Sample size considerations for “omics” data

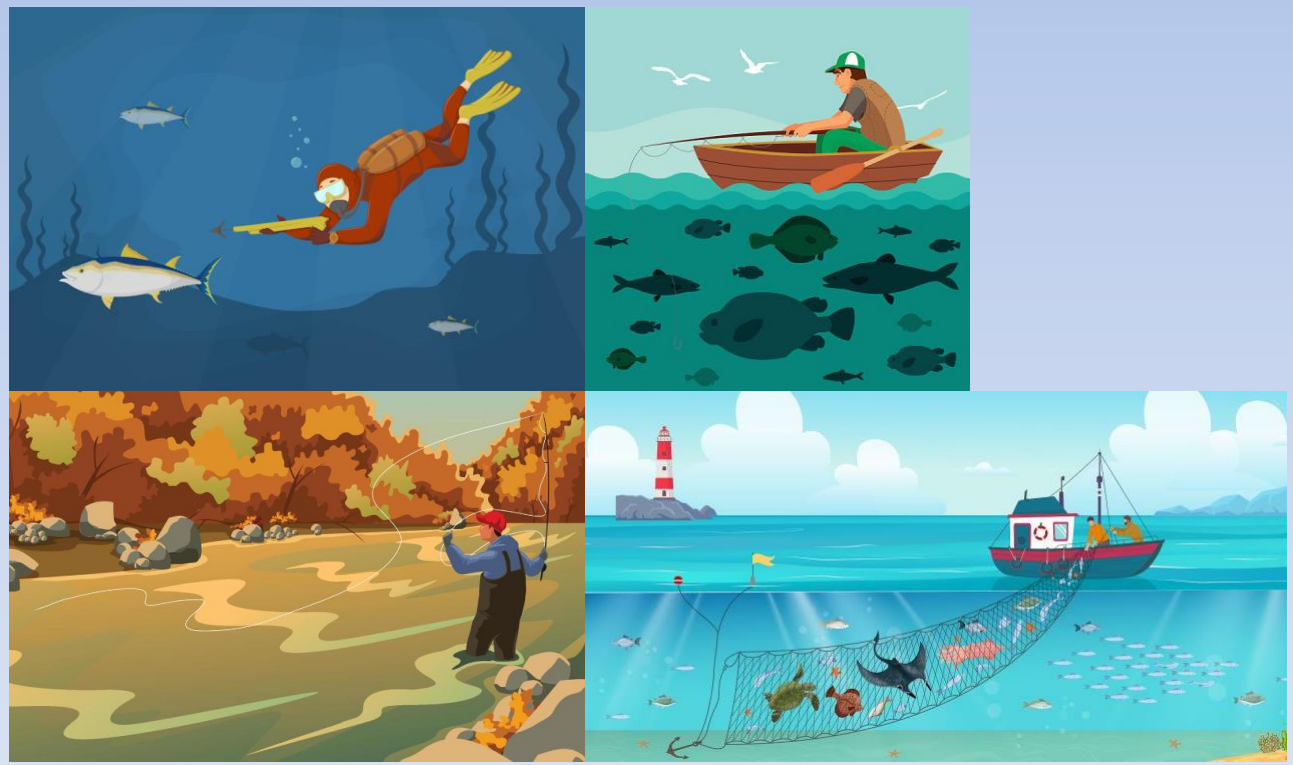
(and other kinds of fishing expeditions)

Jack Wiedrick, MS MA

Biostatistics, Epidemiology, and Research Design (BERD) Program

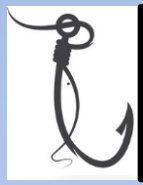
Photo: Belted kingfisher, Eugene. Norman Goo

Some fishing metaphors

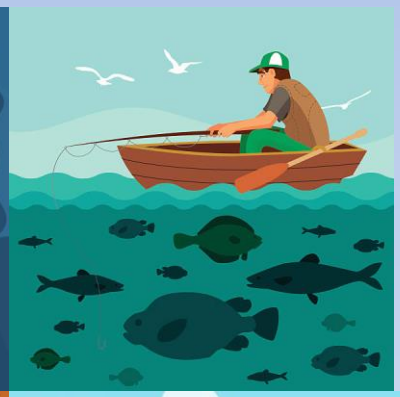


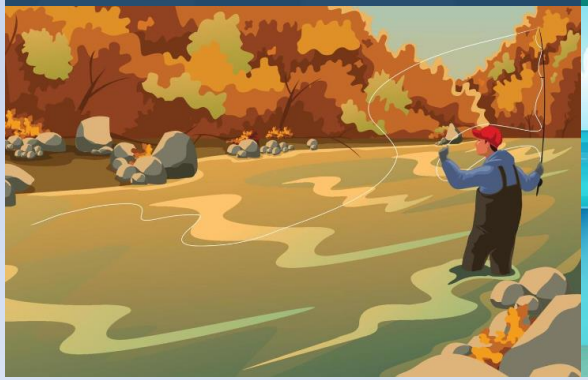
Some fishing metaphors

= measurement



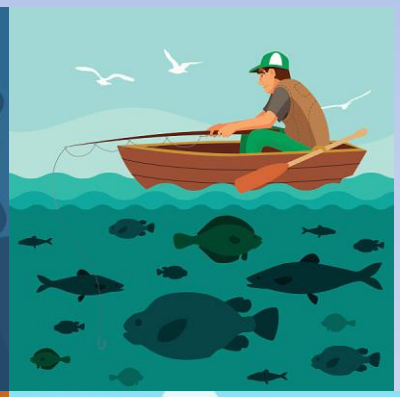
= inference engine

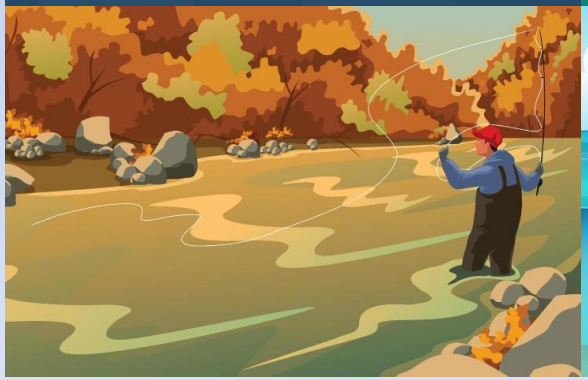




Some fishing metaphors

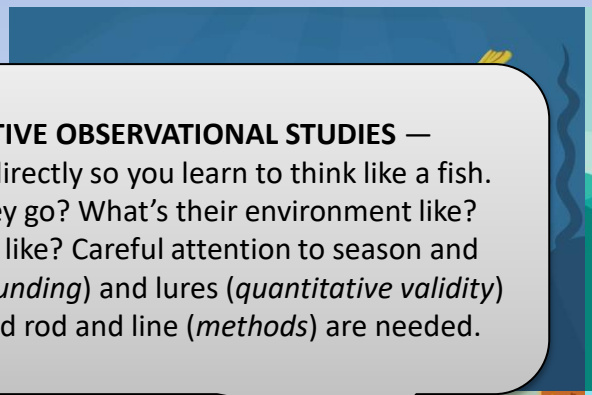
CLINICAL TRIALS and **DESIGNED EXPERIMENTS** — You go exactly where the fish are and watch them as you shoot. Lots of scientific prep and skill required.





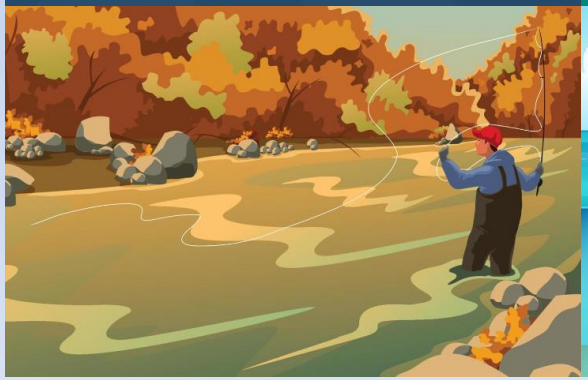
Some fishing metaphors

PROSPECTIVE OBSERVATIONAL STUDIES —
You can't see directly so you learn to think like a fish. Where do they go? What's their environment like? What do they like? Careful attention to season and currents (*confounding*) and lures (*quantitative validity*) as well as good rod and line (*methods*) are needed.

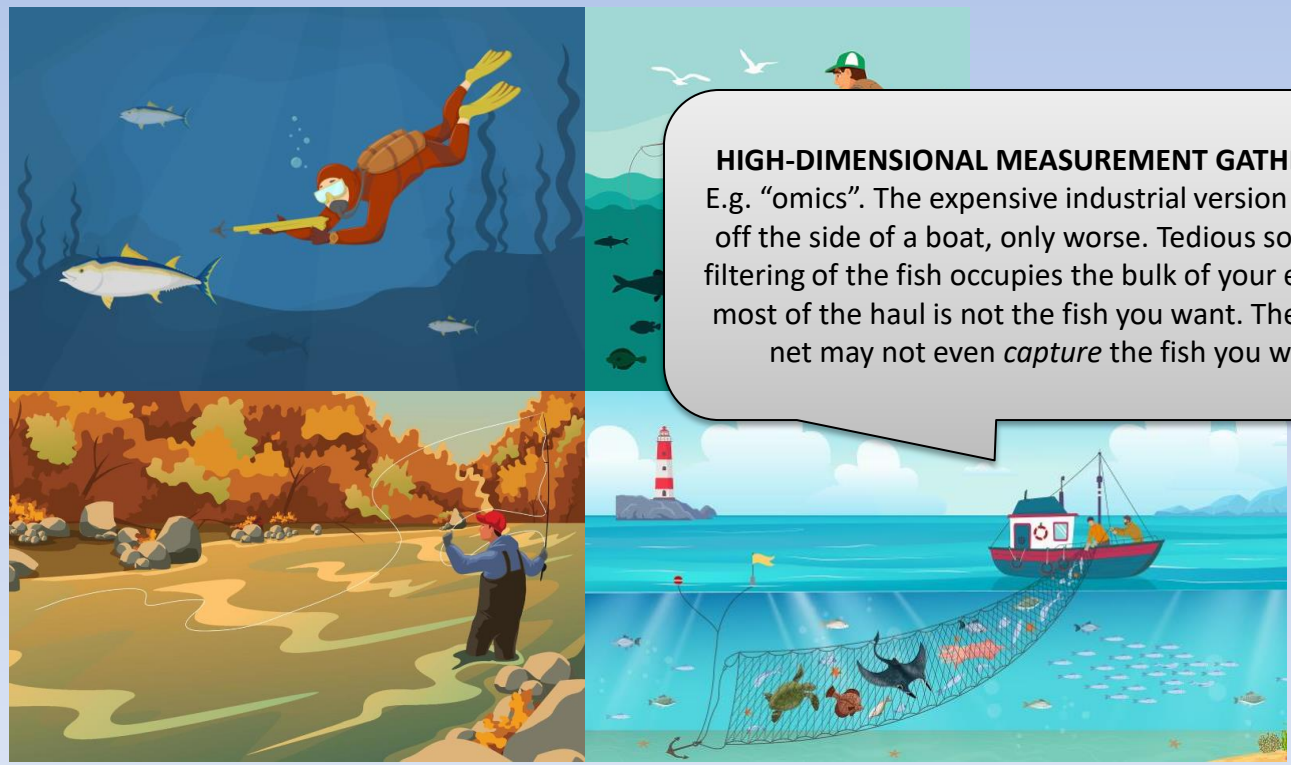


Some fishing metaphors

SURVEYS and **EXPLORATORY STUDIES** —
Row yourself into a fish arena and cast about for whatever fish are there. At best you may know about the *presence* of fish but still can't see them (much). Your equipment is usually generic and subpar.



Some fishing metaphors



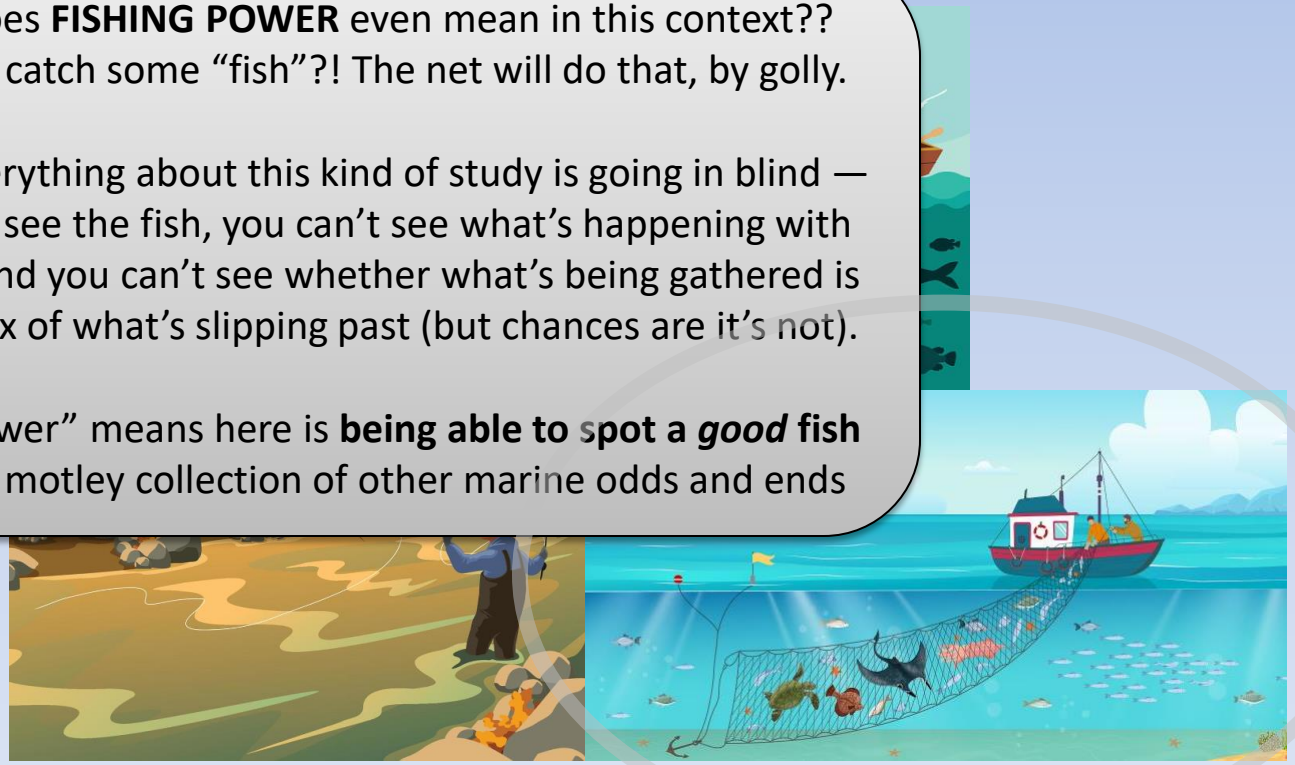
HIGH-DIMENSIONAL MEASUREMENT GATHERING — E.g. “omics”. The expensive industrial version of casting off the side of a boat, only worse. Tedious sorting and filtering of the fish occupies the bulk of your effort, and most of the haul is not the fish you want. The trawling net may not even *capture* the fish you want…

Some fishing metaphors

What does **FISHING POWER** even mean in this context?? Power to catch some “fish”?! The net will do that, by golly.

Nearly everything about this kind of study is going in blind — you can’t see the fish, you can’t see what’s happening with the net, and you can’t see whether what’s being gathered is a good mix of what’s slipping past (but chances are it’s not).

What “power” means here is **being able to spot a *good* fish** amid the motley collection of other marine odds and ends



Data personalities of various omics (and things like omics)

DATA TYPE	TYPICAL # OF FEATURES	BIASES	PROBLEMS
genomics	~1M	variant detection depends on context lower coverage affects detection	sequencing errors higher coverage increases error rate
epigenomics	~10-50M	accuracy depends on read depth	sequencing errors
interactomics	~1-10M	accuracy depends on read depth	sequencing errors
transcriptomics	~1K-10M+	accuracy depends on read depth accuracy depends on abundance	dynamic range depends on context
proteomics	~10K	(targeted) only prespecified proteins (untargeted) only abundant proteins	many missing values peptides shared by protein families
metabolomics	~5K	measurement noise sensitivity/drift	many unknown/unlabeled features
giant questionnaires	~100	only responders provide information	missing/inconsistent responses
administrative data	~10-50K	only system clients represented unknown misclassification errors	many data entry/coding errors broken/compromised linkages
financial pricing/returns data	~100K+ *daily*	not biased	high degree of autocorrelation
ecological sensor data	~50 *minutely*	measurement noise sensitivity/drift	high degree of multicollinearity

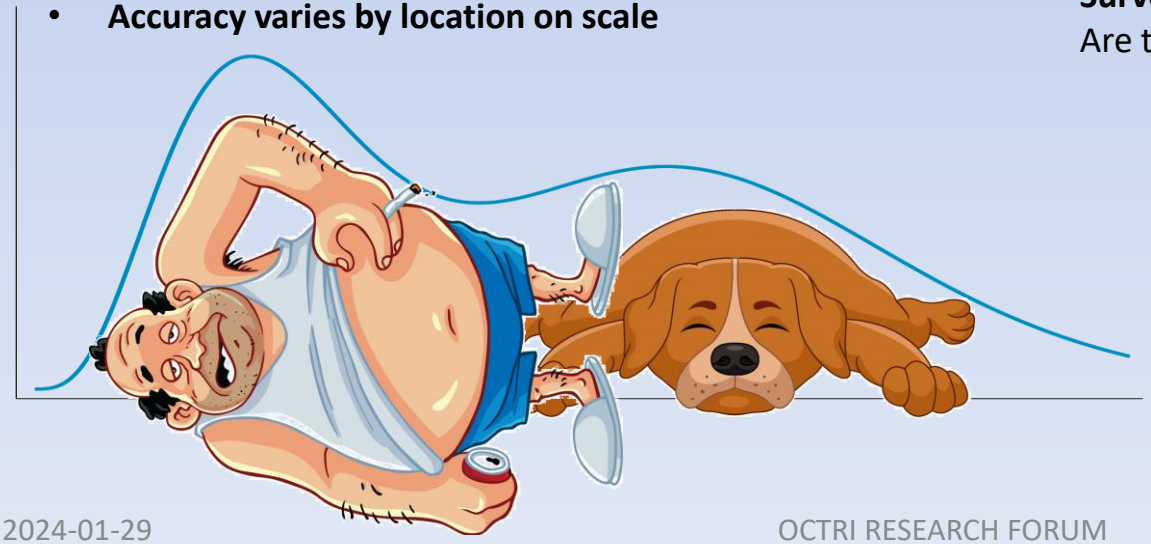
Data personalities of various omics (and things like omics)

MISSING DATA

- **Variable library size…**
  Is it biological?? *Who can say*
- **Missing feature quantifications…**
  Are they below the limit of detection?? *Who can say*
- **Survey nonresponse…**
  Are the responders atypical?? *Who can say*

GRUNGY DATA

- **Extreme skew**
- **Censoring**
- **Clumping of values**
- **Accuracy varies by location on scale**



Data personalities of various omics (and things like omics)

Imputation and normalization can either add or subtract variability, and ***this shifts power curves***.

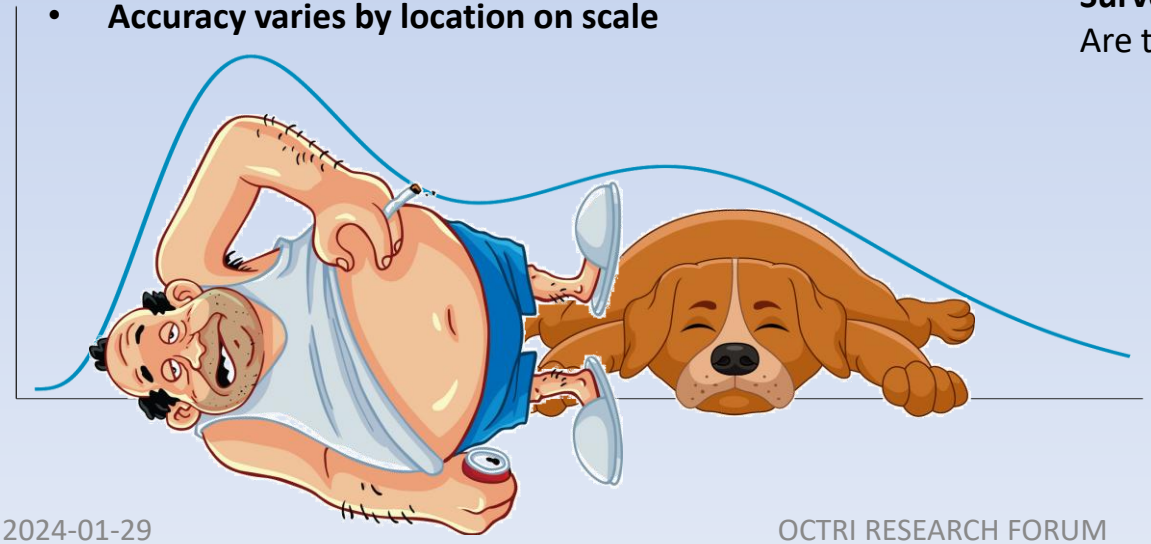
Burden of missing quantifications varies from feature to feature, and ***this makes power curves volatile***.

MISSING DATA

- **Variable library size…**
  Is it biological?? *Who can say*
- **Missing feature quantifications…**
  Are they below the limit of detection?? *Who can say*
- **Survey nonresponse…**
  Are the responders atypical?? *Who can say*

GRUNGY DATA

- **Extreme skew**
- **Censoring**
- **Clumping of values**
- **Accuracy varies by location on scale**



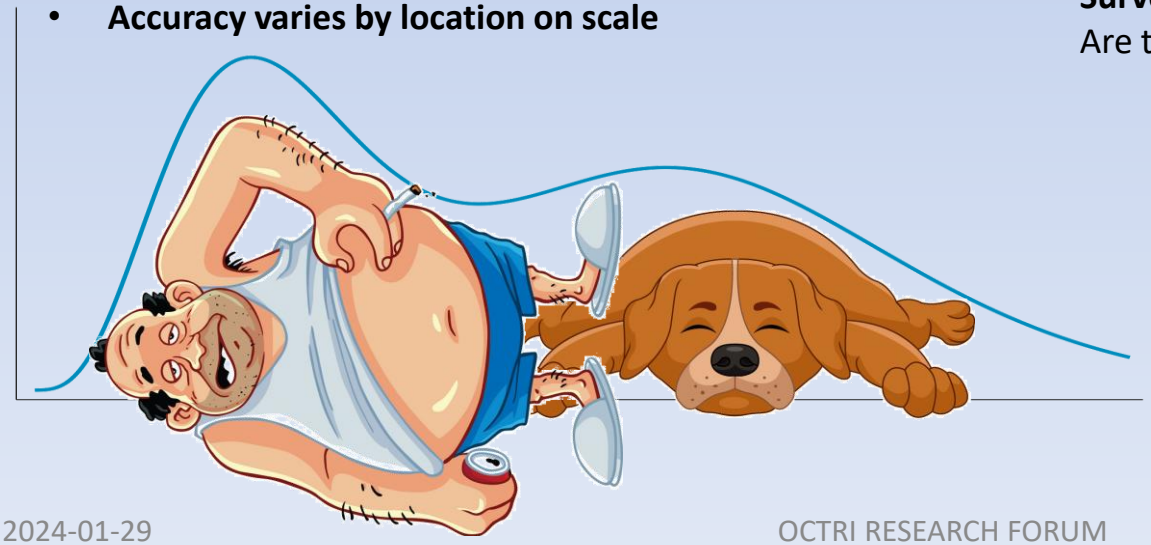
Data personalities of various omics (and things like omics)

Robust methods like median regression or regularized regression may be needed, but performance depends on burden of problems, and ***this shifts power curves***.

Data transformations that help some features will hurt other features, and ***this makes power curves volatile***.

GRUNGY DATA

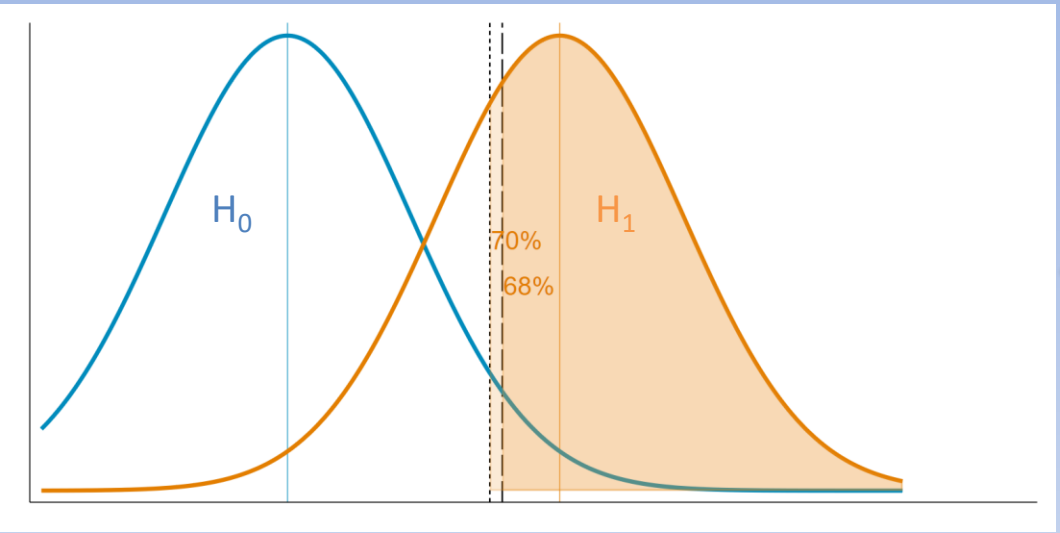
- **Extreme skew**
- **Censoring**
- **Clumping of values**
- **Accuracy varies by location on scale**



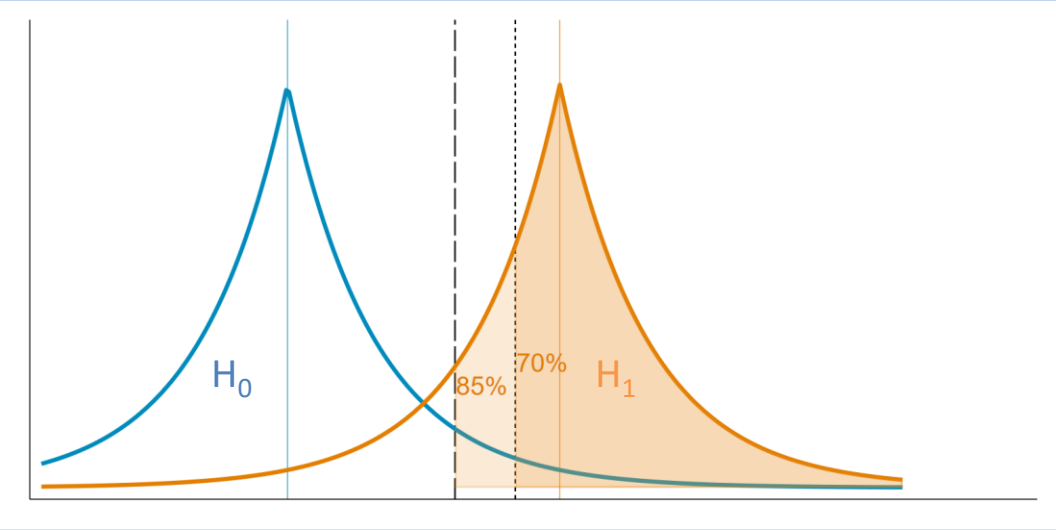
MISSING DATA

- **Variable library size…**
  Is it biological?? *Who can say*
- **Missing feature quantifications…**
  Are they below the limit of detection?? *Who can say*
- **Survey nonresponse…**
  Are the responders atypical?? *Who can say*

Analytical choices do not affect power uniformly

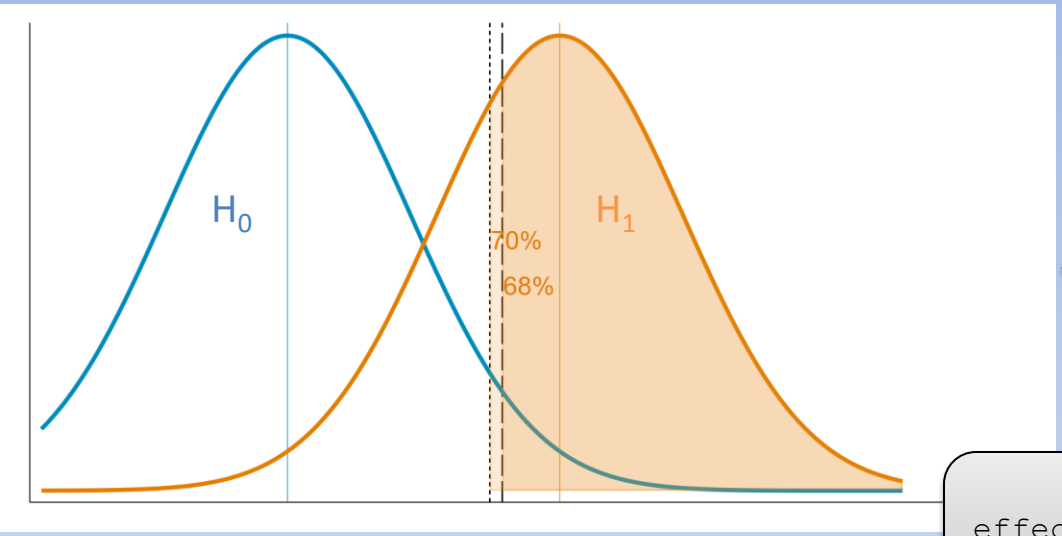


If you subtract the medians of two normal distributions…



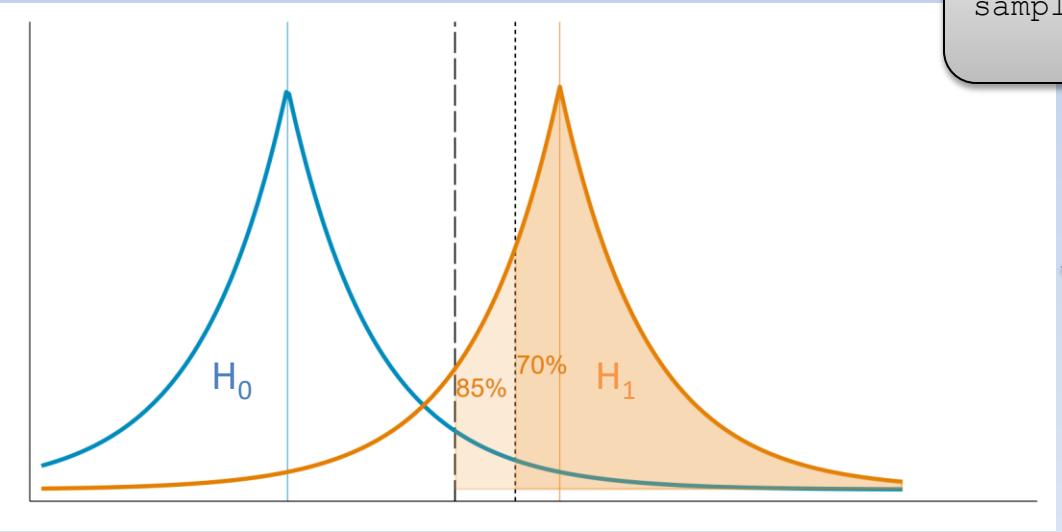
If you subtract the medians of two skewed distributions…

Analytical choices do not affect power uniformly



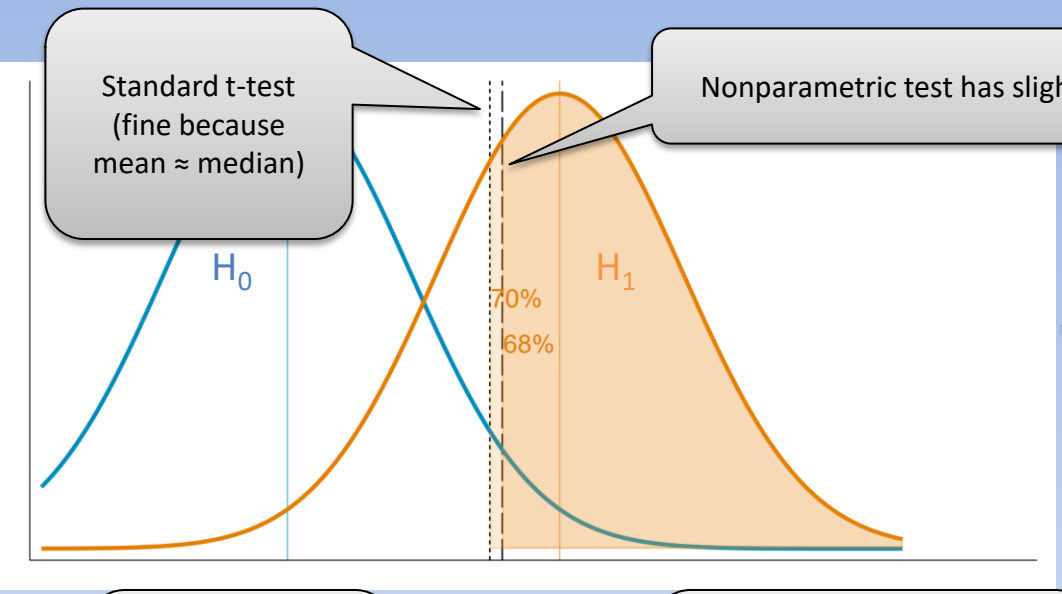
If you subtract the medians of two normal distributions…

```
effect size (d) = 0.35
sample size (n) = 40
```

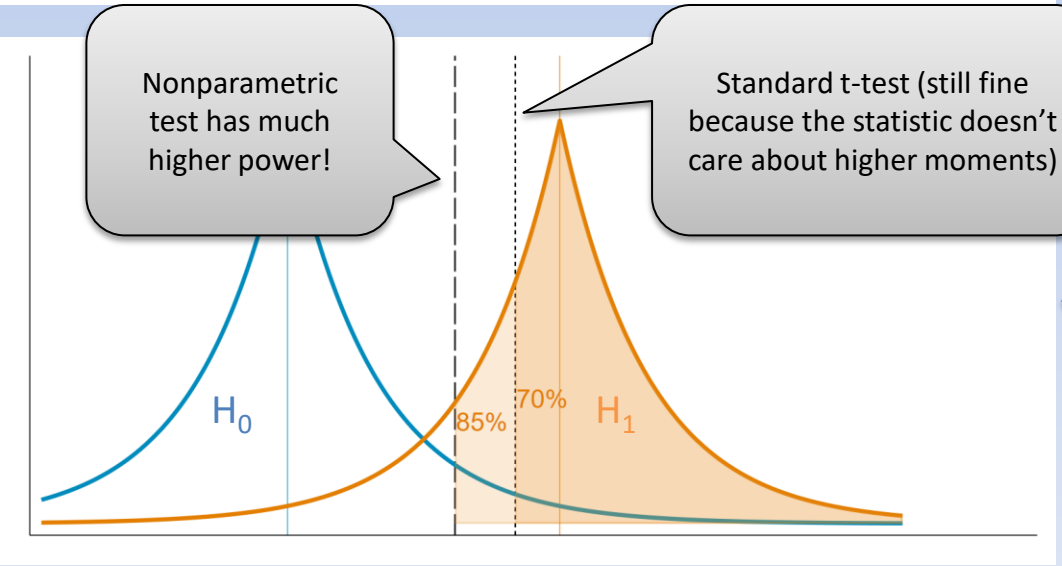


If you subtract the medians of two skewed distributions…

Analytical choices do not affect power uniformly

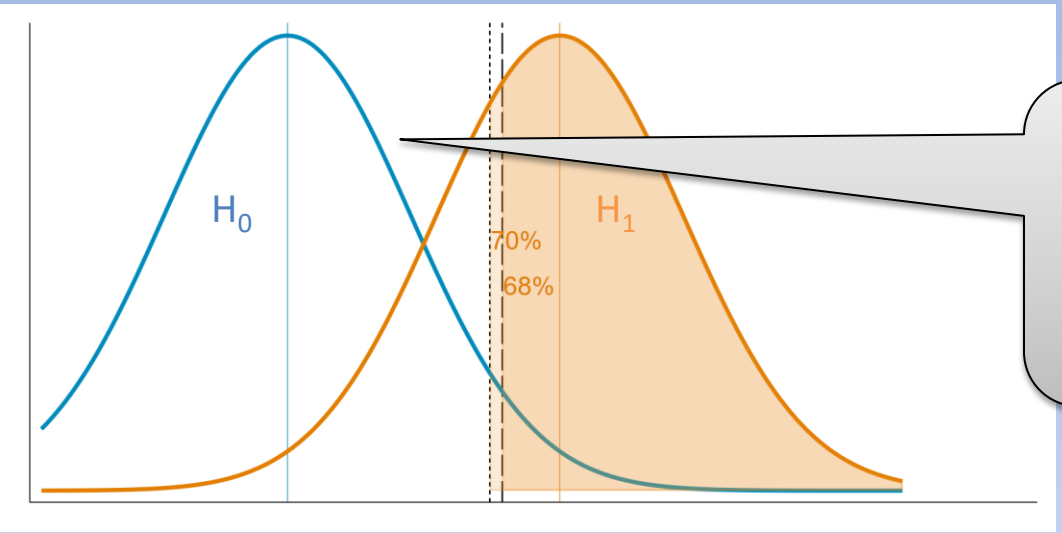


If you subtract the medians of two normal distributions…



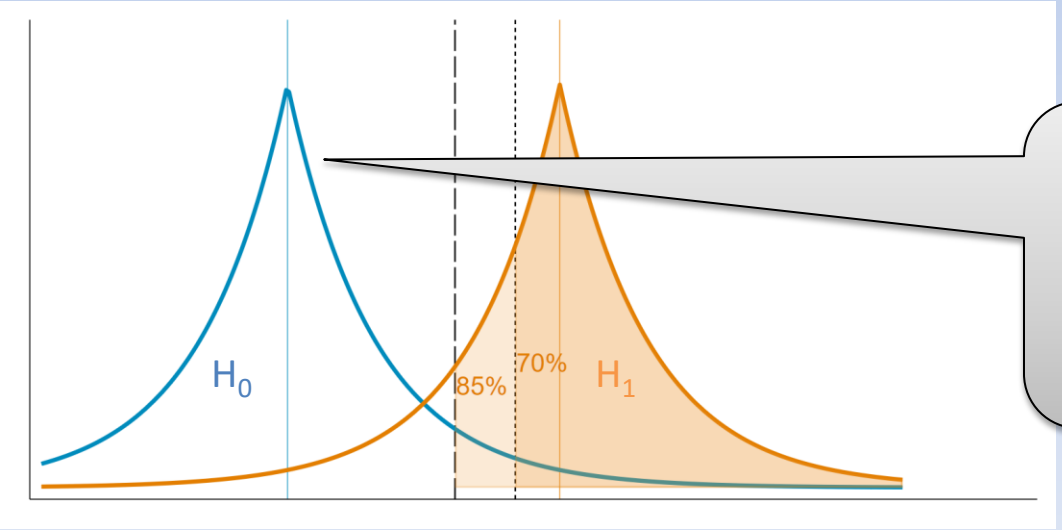
If you subtract the medians of two skewed distributions…

Analytical choices do not affect power uniformly



This is the prior distribution you assume for coefficients when you perform a **RIDGE REGRESSION**!

(interpretable as a Bayesian procedure with independent *normal* priors on all coefficients)



This is the prior distribution you assume for coefficients when you perform a **LASSO REGRESSION**!

(interpretable as a Bayesian procedure with independent *Laplace* priors on all coefficients)

Why multivariate methods fail to be robust in high dimensions

Recall the basic power calculation for a difference of two means:

$n = n_1 + n_2$ is the total sample size
and $r = n_2/n_1$ is the ratio of group sizes

$\frac{\delta}{\sigma}$ is the effect size

$$power = 1 - \beta \approx \Phi\left(\frac{\sqrt{nr}}{1+r}\frac{\delta}{\sigma} - z_\alpha\right)$$

z_α is the upper αth quantile of the standard normal distribution

$\sigma = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$ is a (crudely) pooled standard deviation

Why multivariate methods fail to be robust in high dimensions

Recall the basic power calculation for a difference of two means:

$$power = 1 - \beta \approx \Phi\left(\frac{\sqrt{nr}}{1+r}\frac{\delta}{\sigma} - z_\alpha\right)$$

For a given power tolerance, this implies a total sample size requirement of:

$$n_{\delta/\sigma} \approx \left(\frac{\sigma}{\delta}\right)^2 \left(z_\beta + z_\alpha\right)^2 \left(\frac{(1+r)^2}{r}\right)$$

(scaling to achieve total)

sample size is proportional to the square of the noise-to-signal ratio

proportionality factor depends on your tolerance for making errors

Why multivariate methods fail to be robust in high dimensions

Recall the basic power calculation for a difference of two means:

$$power = 1 - \beta \approx \Phi\left(\frac{\sqrt{nr}}{1+r}\frac{\delta}{\sigma} - z_\alpha\right)$$

For a given power tolerance, this implies a total sample size requirement of:

$$n_{\delta/\sigma} \approx \left(\frac{\sigma}{\delta}\right)^2 \left(z_\beta + z_\alpha\right)^2 \left(\frac{(1+r)^2}{r}\right)$$

BUT DO WE KNOW ANY OF THIS STUFF??!

Why multivariate methods fail to be robust in high dimensions

Recall the basic power calculation for a difference of two means:

$$power = 1 - \beta \approx \Phi\left(\frac{\sqrt{nr}}{1+r}\frac{\delta}{\sigma} - z_\alpha\right)$$

For a given power tolerance, this implies a total sample size requirement of:

Variance depends on the feature…
and we may have millions of them!

$$\approx \left(\frac{\sigma}{\delta}\right)^2 \left(z_\beta + z_\alpha\right)^2 \left(\frac{(1+r)^2}{r}\right)$$

BUT DO WE KNOW ANY OF THIS STUFF??!

Why multivariate methods fail to be robust in high dimensions

Recall the basic power calculation for a difference of two means:

$$power = 1 - \beta \approx \Phi\left(\frac{\sqrt{nr}}{1+r}\frac{\delta}{\sigma} - z_\alpha\right)$$

For a given power tolerance, this implies a total sample size requirement of:

$$n_{\delta/\sigma} \approx \left(\frac{\sigma}{\delta}\right)^2 \left(z_\beta + z_\alpha\right)^2 \left(\frac{(1+r)^2}{r}\right)$$

Group difference also depends…
and could be millions of those too!

KNOW ANY OF THIS STUFF??!

Why multivariate methods fail to be robust in high dimensions

Recall the basic power calculation for a difference of two means:

$$power = 1 - \beta \approx \Phi\left(\frac{\sqrt{nr}}{1+r}\frac{\delta}{\sigma} - z_\alpha\right)$$

For a given power tolerance, this implies a total sample size requirement of:

$$n_{\delta/\sigma} \approx \left(\frac{\sigma}{\delta}\right)^2 \left(z_\beta + z_\alpha\right)^2 \left($$

Apparently we're supposed to "adjust" this somehow too… but for millions of things??

Most of which we don't actually care about…!

BUT DO WE KNOW ANY OF THIS STUFF??!

Why multivariate methods fail to be robust in high dimensions

Recall the basic power calculation for a difference of two means:

$$power = 1 - \beta \approx \Phi\left(\frac{\sqrt{nr}}{1+r}\frac{\delta}{\sigma} - z_\alpha\right)$$

For a given power tolerance, this implies a total sampl

$$n_{\delta/\sigma} \approx \left(\frac{\sigma}{\delta}\right)^2 \left(z_\beta + z_\alpha\right)^2 \left(\frac{\quad}{r}\right)$$

And for a given sample size *n*, adjusting α down implies adjusting β *up*… oh brother…

BUT DO WE KNOW ANY OF THIS STUFF??!

Why multivariate methods fail to be robust in high dimensions

Recall the basic power calculation for a difference of two means:

$$power = 1 - \beta \approx \Phi\left(\frac{\sqrt{nr}}{1+r}\frac{\delta}{\sigma} - z_\alpha\right)$$

For a given power tolerance, this implies a total sample size requirement of:

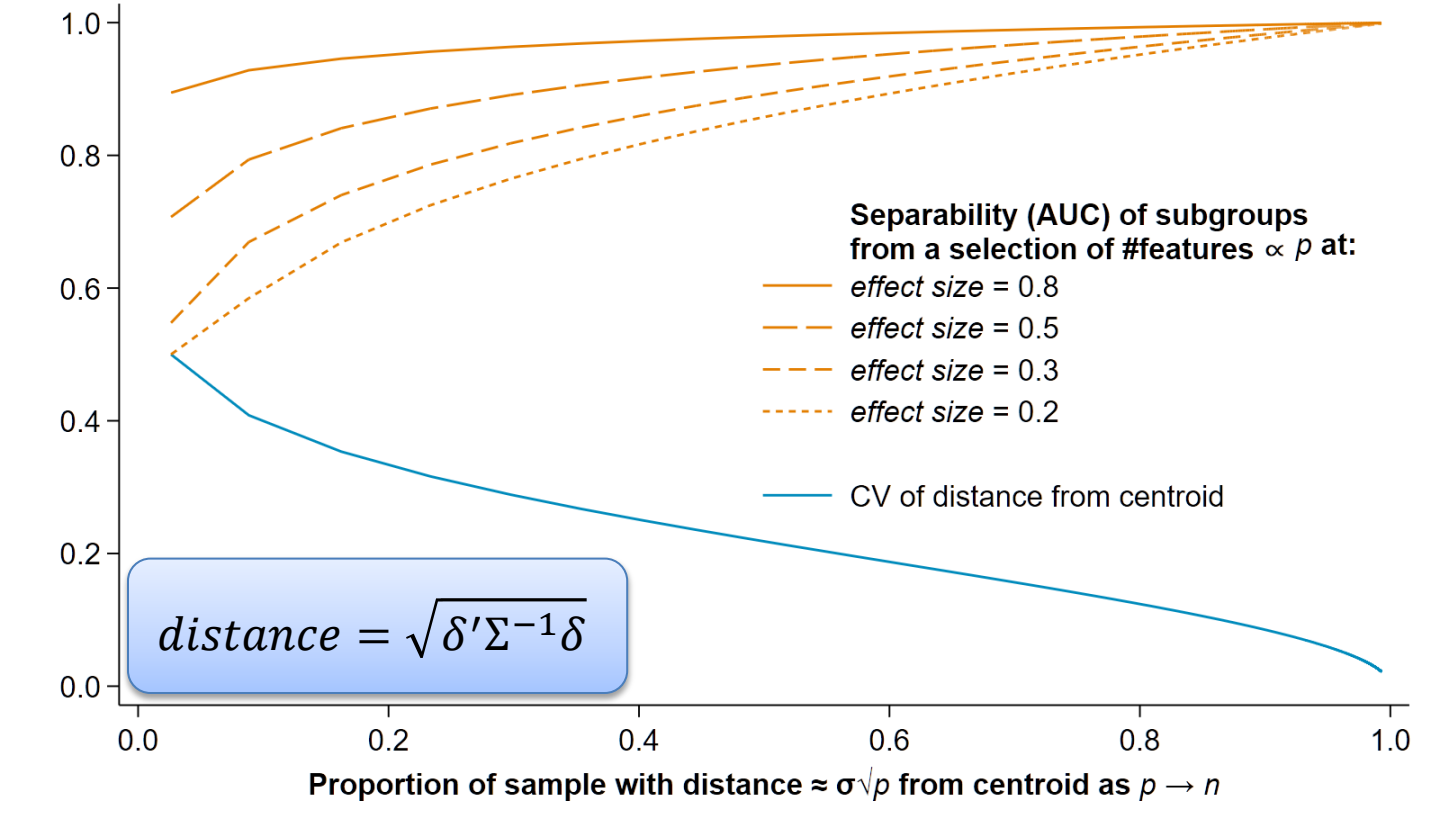
$$n_{\delta/\sigma} \approx \left(\frac{\sigma}{\delta}\right)^2 \left(z_\beta + z_\alpha\right)^2 \left(\frac{(1+r)^2}{r}\right)$$

Our so-called sample size calculation is **UNDERDETERMINED**

It needs to be *sensitive to all the local conditions* applying individually to each feature, but we can't make it that way, not just because we don't know the conditions, but even more crucially because the conditions are *heterogeneous*

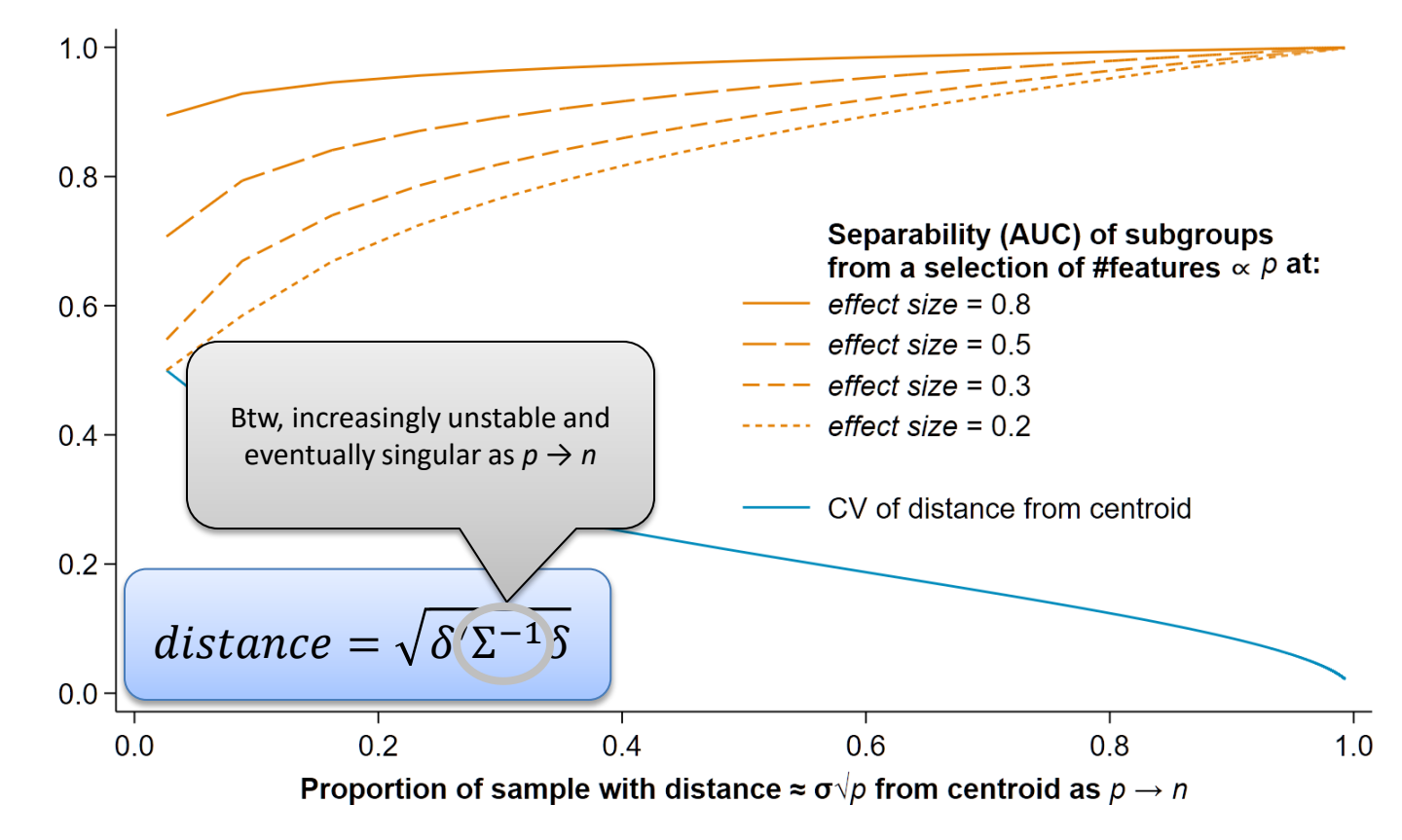
Why multivariate methods fail to be robust in high dimensions

But there's another problem too, also caused by the high dimensionality of the data:



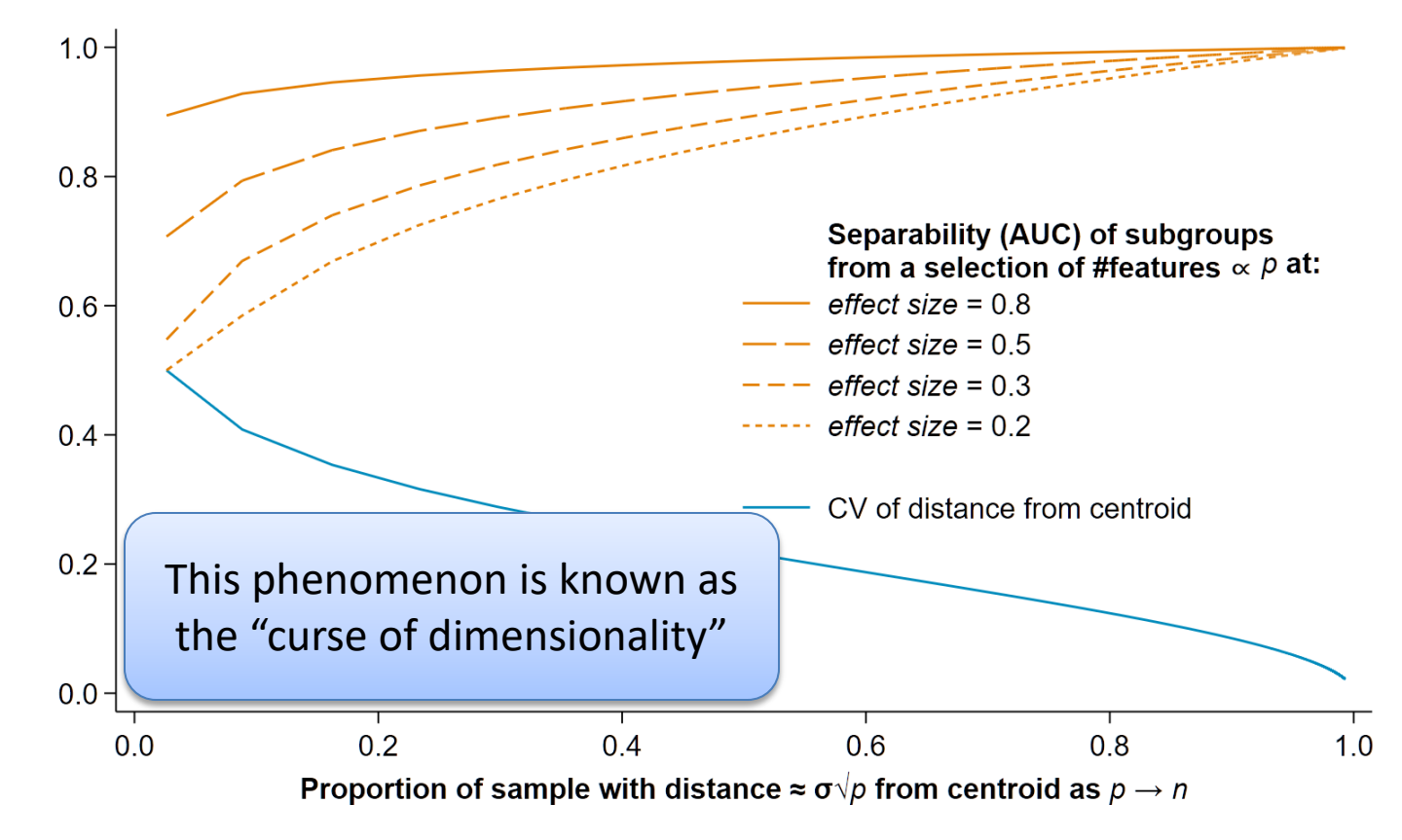
Why multivariate methods fail to be robust in high dimensions

But there's another problem too, also caused by the high dimensionality of the data:



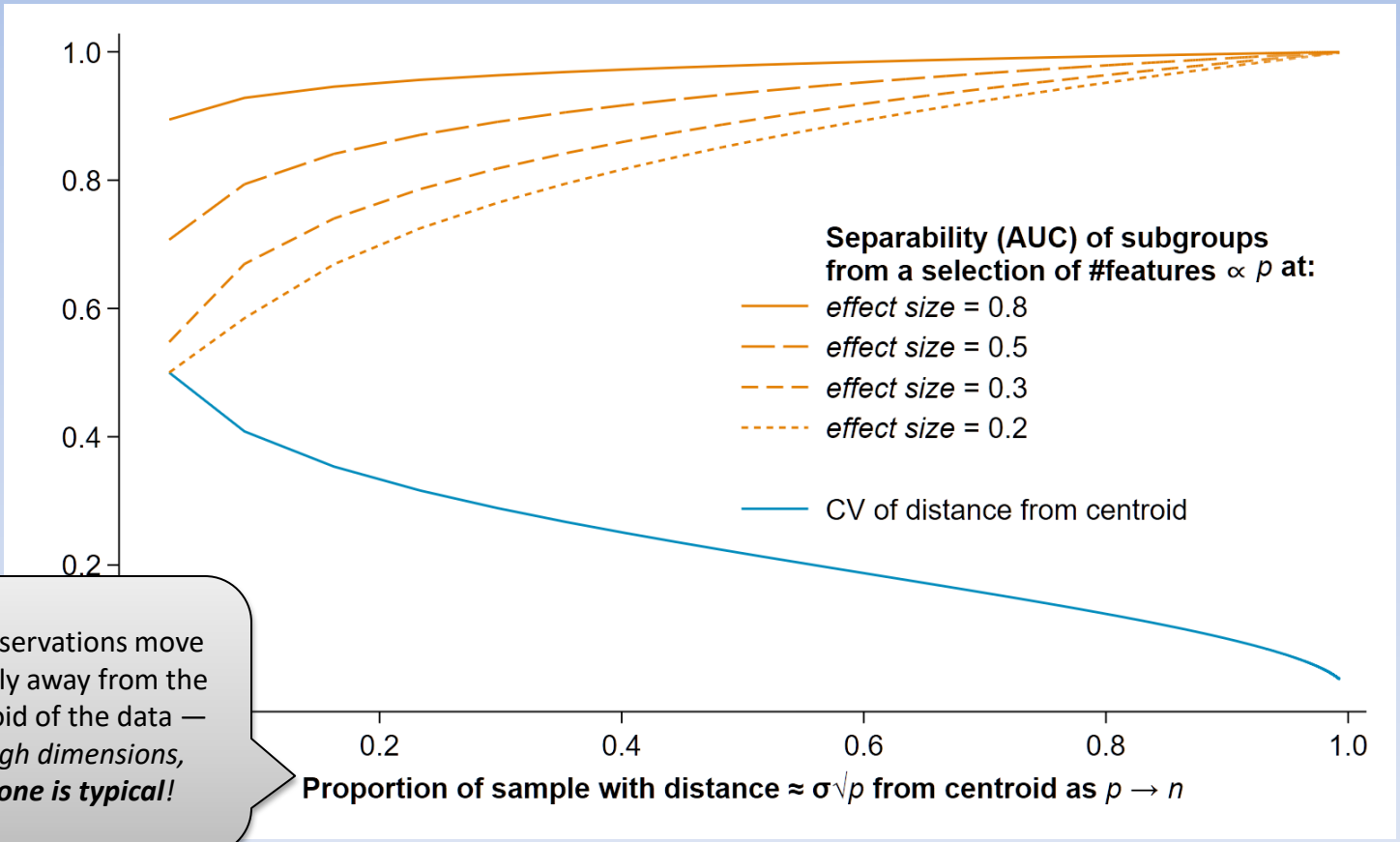
Why multivariate methods fail to be robust in high dimensions

As you add more features — *especially irrelevant ones!* — to data with a fixed sample size, the observations don't cluster together (as you'd expect from "adding more information") but rather *fan apart* and uniformly approach a thin surface inside the data space…



Why multivariate methods fail to be robust in high dimensions

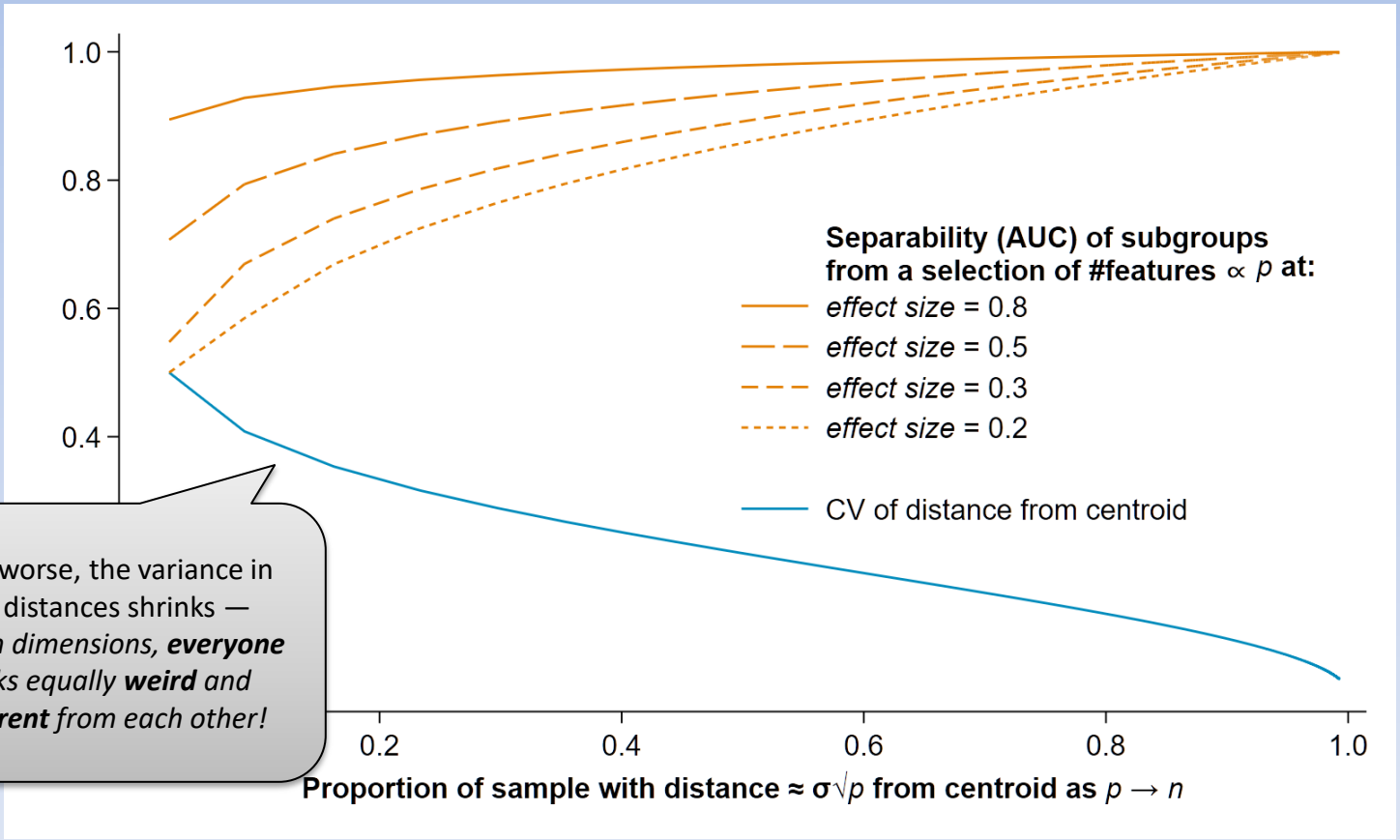
As you add more features — *especially irrelevant ones!* — to data with a fixed sample size, the observations don't cluster together (as you'd expect from "adding more information") but rather *fan apart* and uniformly approach a thin surface inside the data space…



All observations move steadily away from the centroid of the data — *in high dimensions,* ***no one is typical****!*

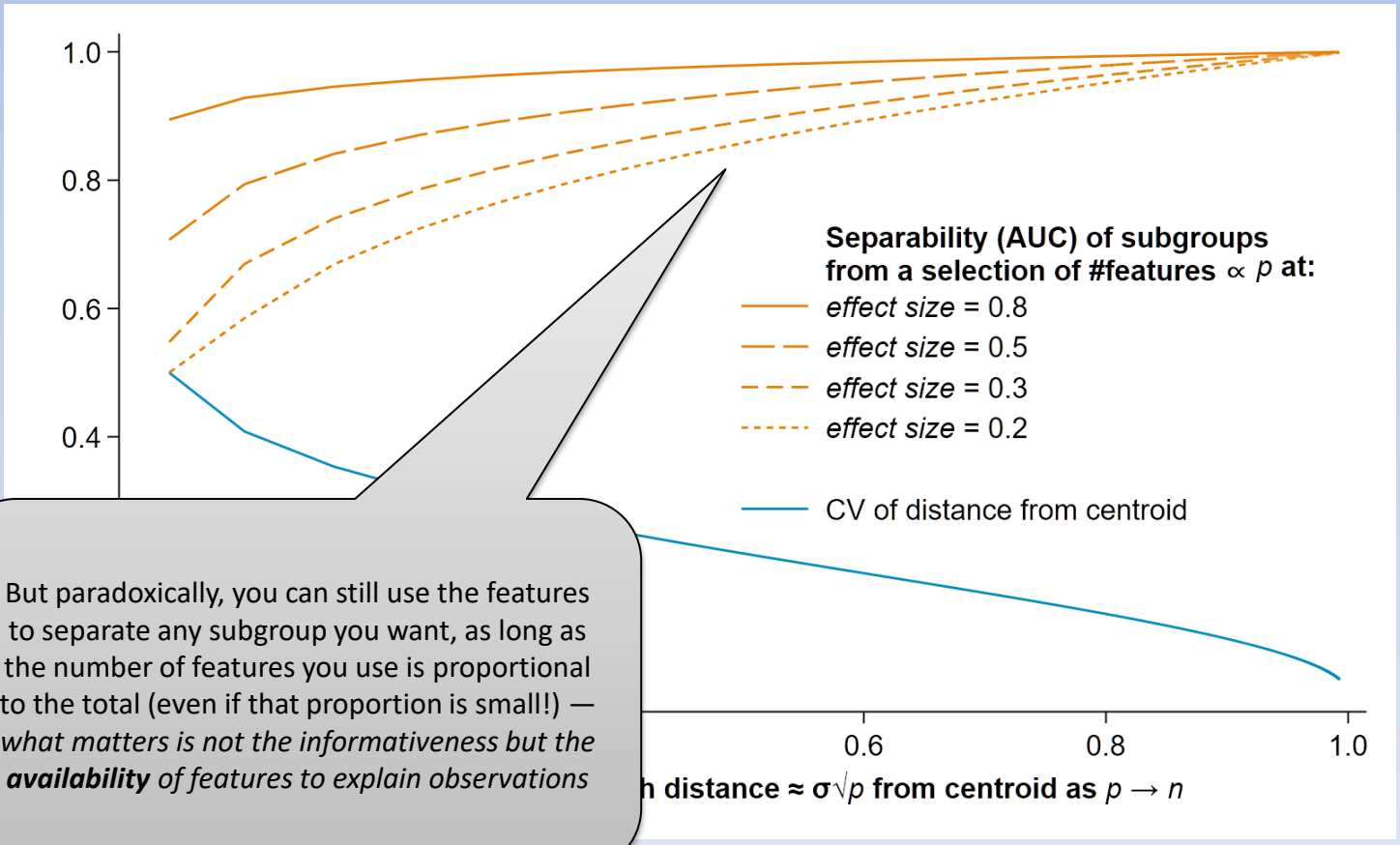
Why multivariate methods fail to be robust in high dimensions

As you add more features — *especially irrelevant ones!* — to data with a fixed sample size, the observations don't cluster together (as you'd expect from "adding more information") but rather *fan apart* and uniformly approach a thin surface inside the data space…



Why multivariate methods fail to be robust in high dimensions

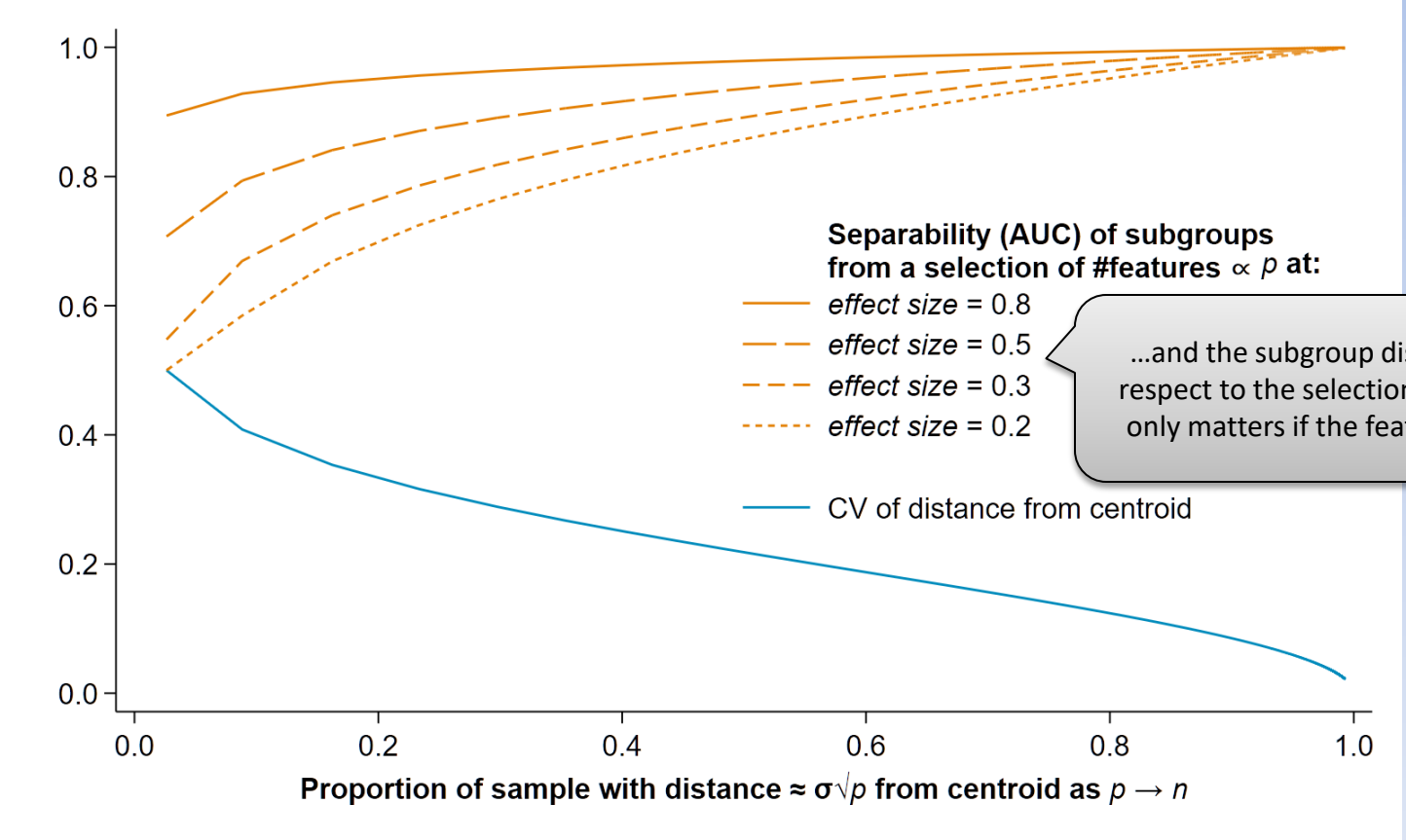
As you add more features — *especially irrelevant ones!* — to data with a fixed sample size, the observations don't cluster together (as you'd expect from "adding more information") but rather *fan apart* and uniformly approach a thin surface inside the data space…



But paradoxically, you can still use the features to separate any subgroup you want, as long as the number of features you use is proportional to the total (even if that proportion is small!) — *what matters is not the informativeness but the* ***availability*** *of features to explain observations*

Why multivariate methods fail to be robust in high dimensions

As you add more features — *especially irrelevant ones!* — to data with a fixed sample size, the observations don't cluster together (as you'd expect from "adding more information") but rather *fan apart* and uniformly approach a thin surface inside the data space…



Why multivariate methods fail to be robust in high dimensions

As you add more features — *especially irrelevant ones!* — to data with a fixed sample size, the observations don't cluster together (as you'd expect from "adding more information") but rather *fan apart* and uniformly approach a thin surface inside the data space…

Now we see our sample size calculation is also **OVERDETERMINED**

We can choose any sample size we like as long as we're willing to use enough features to explain them. Even if we feel we're being parsimonious (e.g. using regularized regression), very high availability of features will always lead to high predictability of subgroups (aka *overfitting*) unless we apply stringent filters

Why multivariate methods fail to be robust in high dimensions

Multivariate thinking in very high dimensions thus leads to a **PARADOX**

- ✓ We can't choose any globally appropriate sample size due to feature heterogeneity
- ✓ At the same time, any sample size we choose is adequate due to feature availability

Why multivariate methods fail to be robust in high dimensions

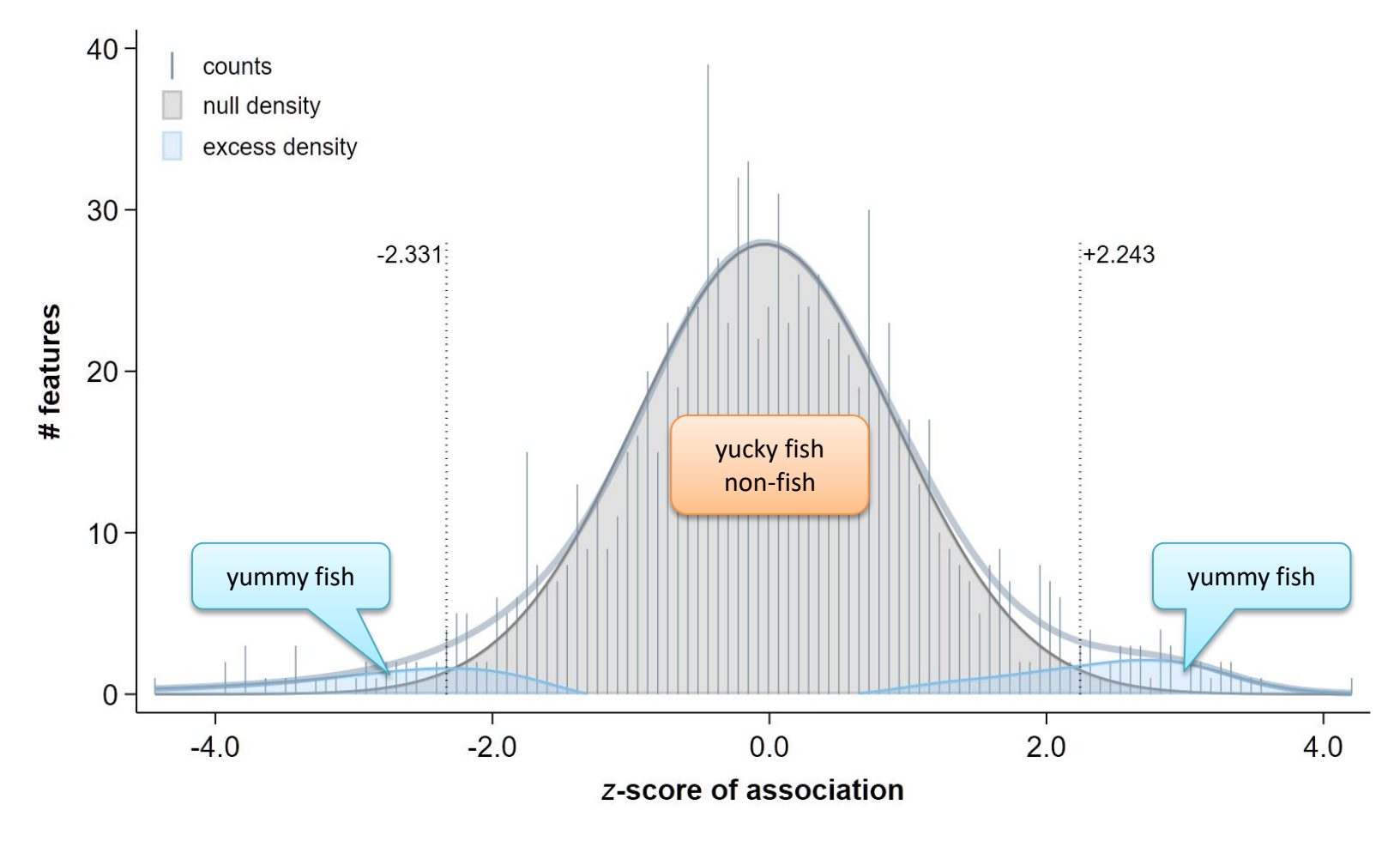
Multivariate thinking in very high dimensions thus leads to a **PARADOX**

- ✓ We can't choose any globally appropriate sample size due to feature heterogeneity
- ✓ At the same time, any sample size we choose is adequate due to feature availability

We need another approach…

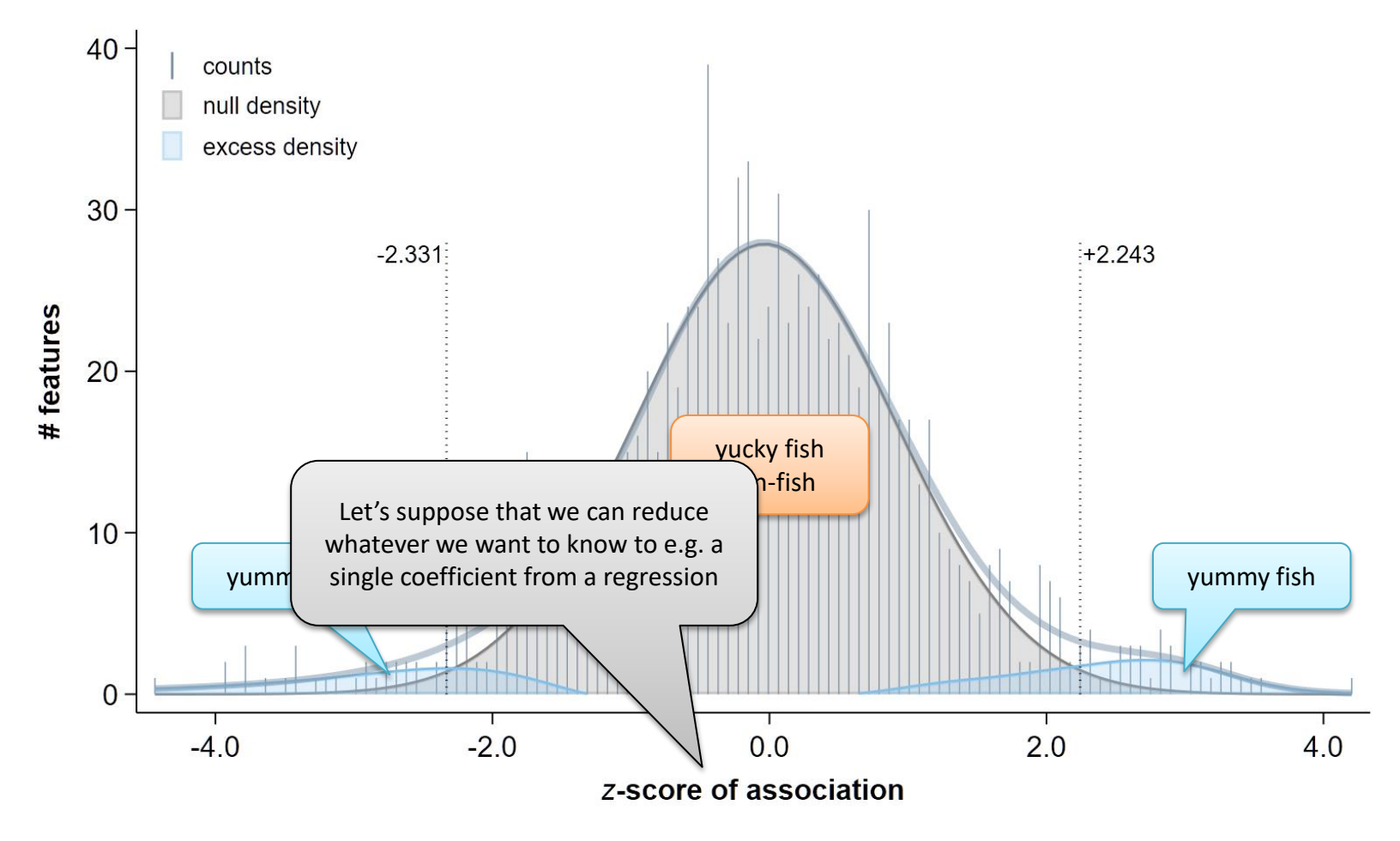
So much complexity, so little time

Multivariate thinking leads to a dead end. We should start thinking *distributionally*…



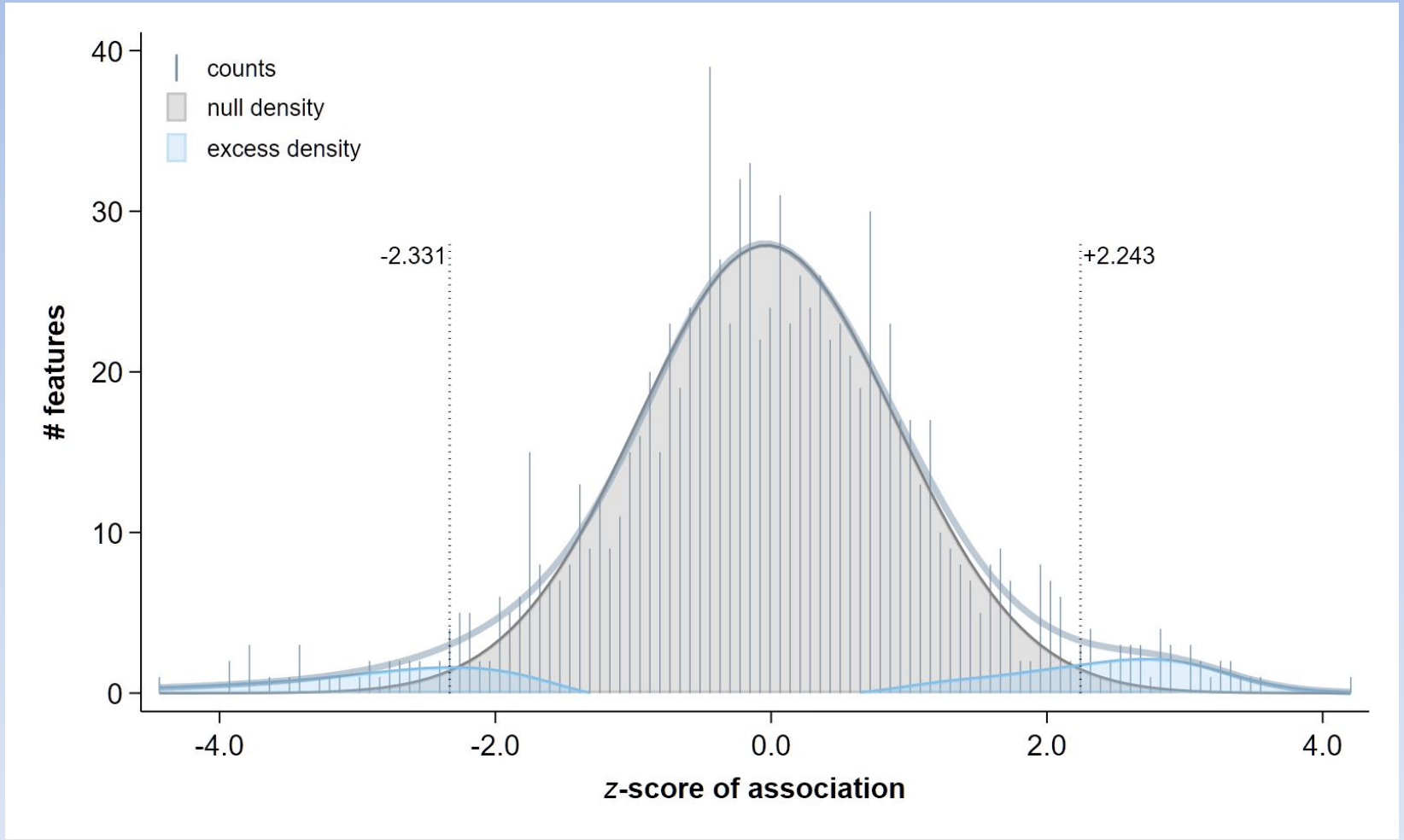
So much complexity, so little time

Multivariate thinking leads to a dead end. We should start thinking *distributionally*…



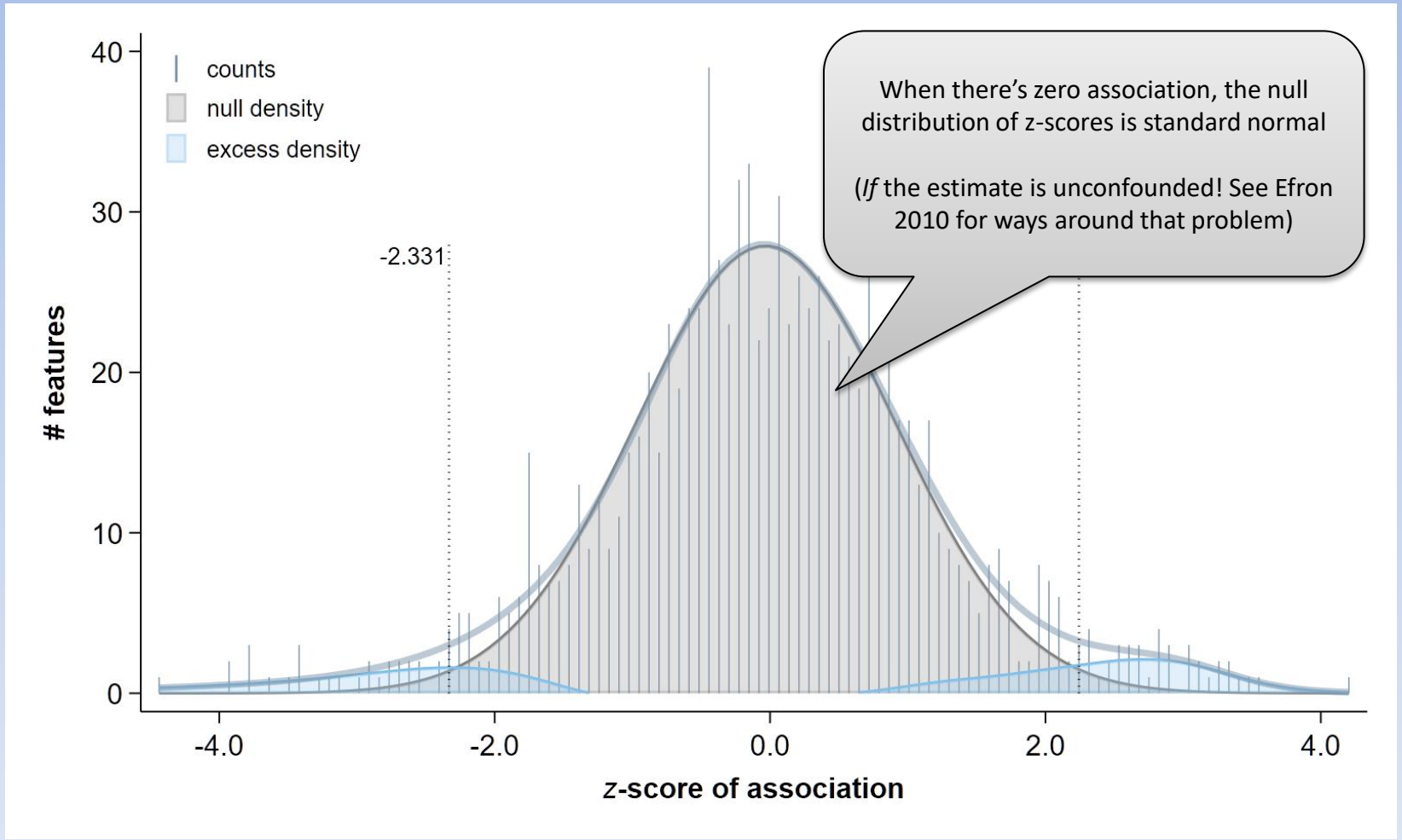
So much complexity, so little time

The idea: Fit a finite-mixture model to the association scores. Excess density is exciting!



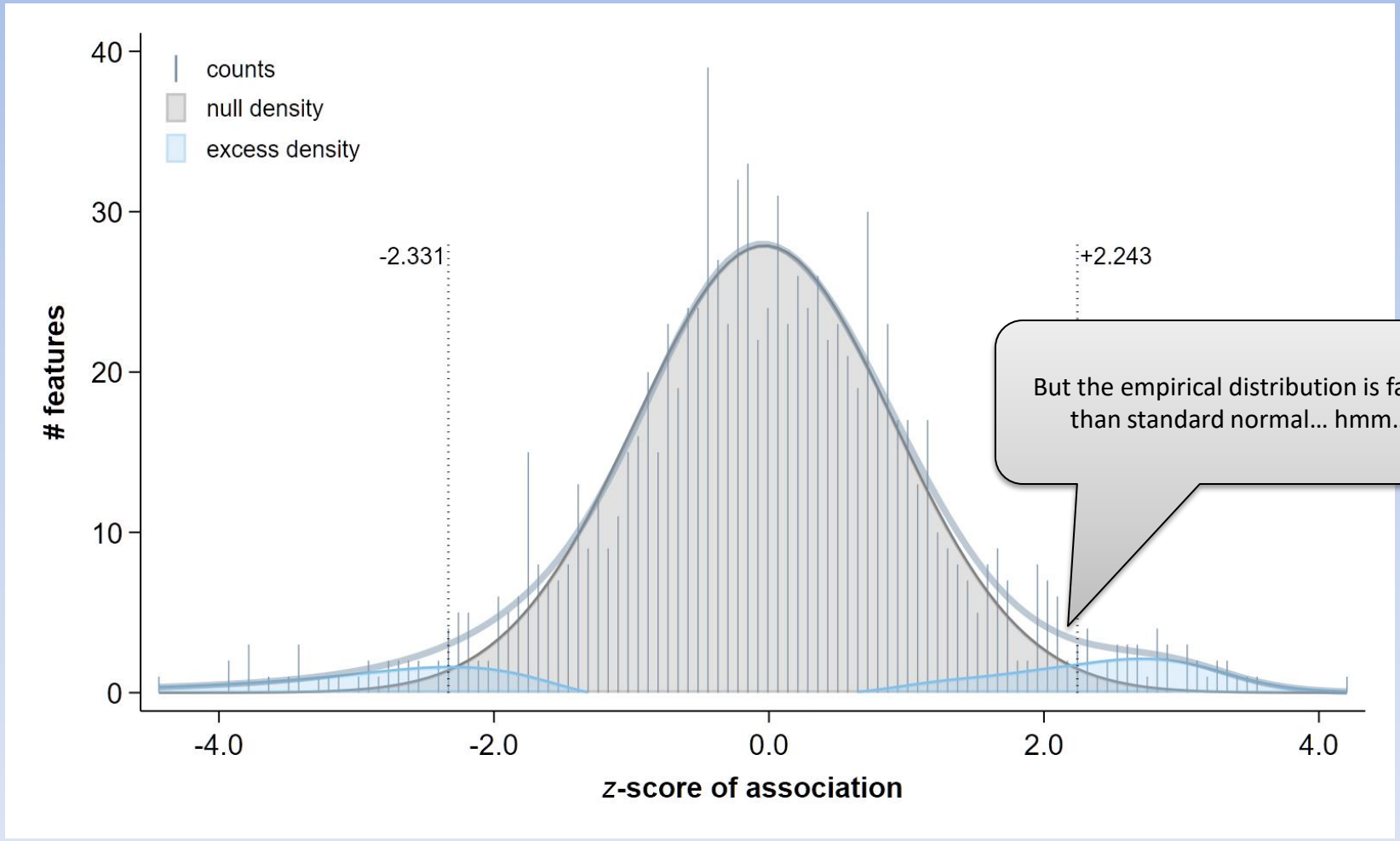
So much complexity, so little time

The idea: Fit a finite-mixture model to the association scores. Excess density is exciting!



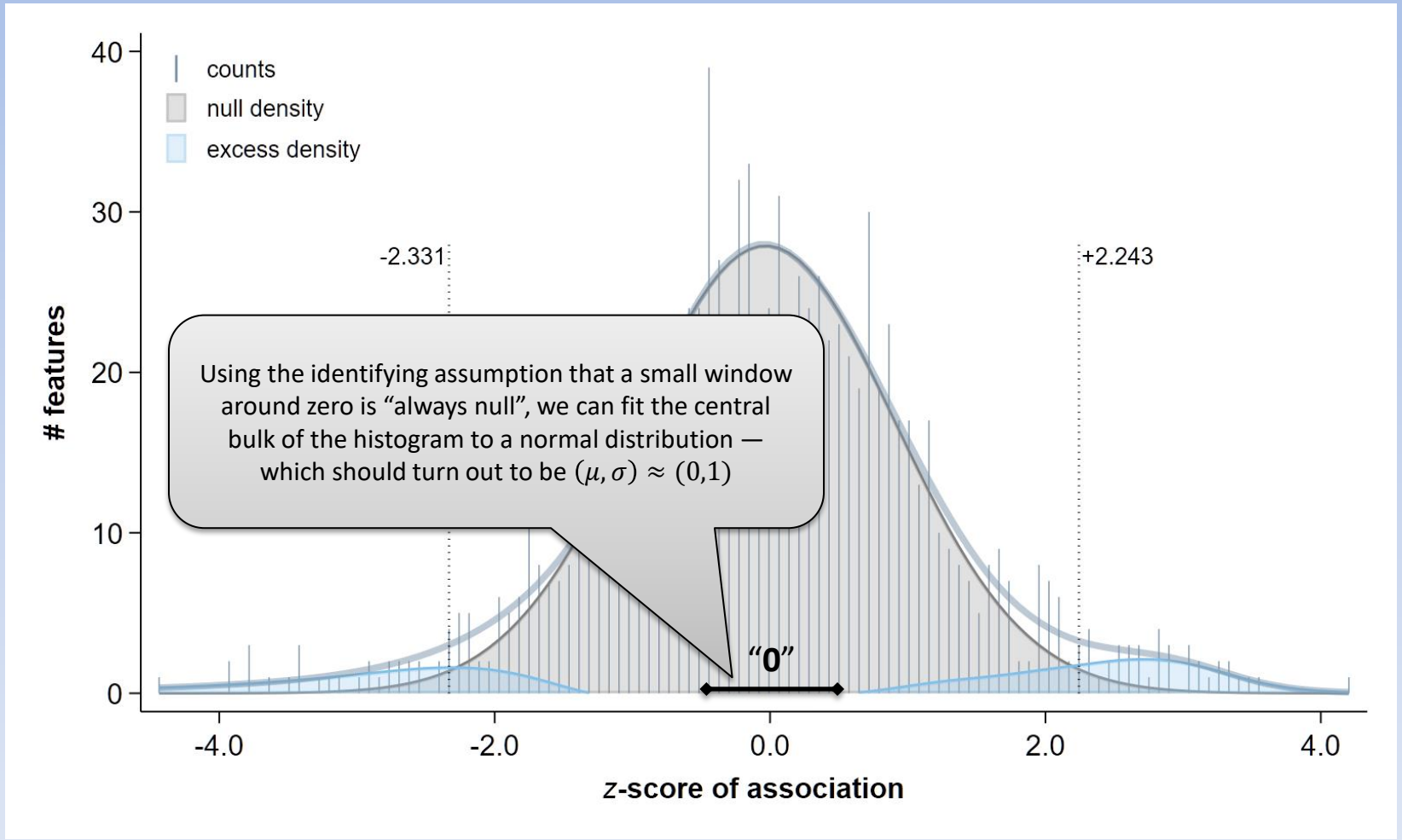
So much complexity, so little time

The idea: Fit a finite-mixture model to the association scores. Excess density is exciting!



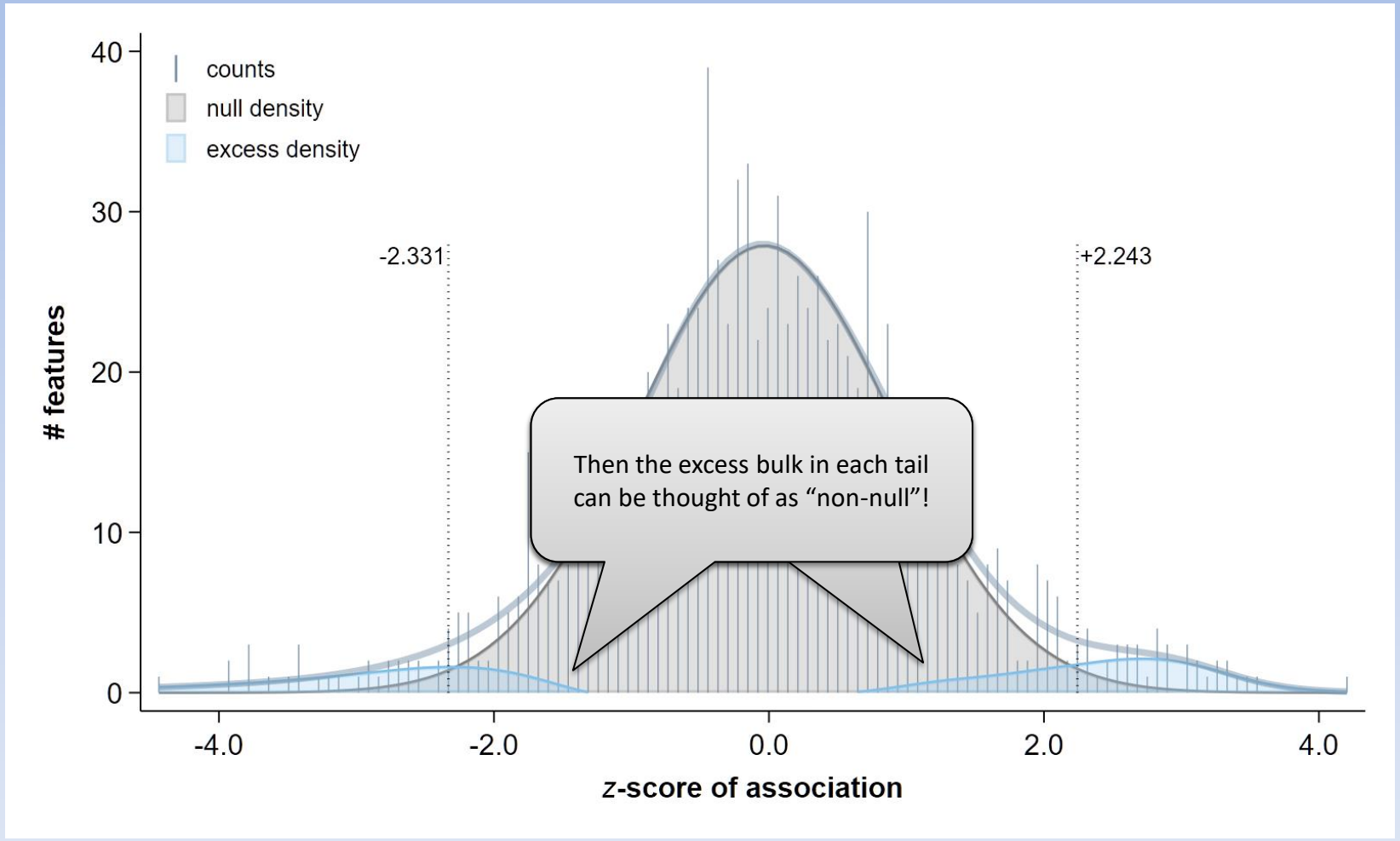
So much complexity, so little time

The idea: Fit a finite-mixture model to the association scores. Excess density is exciting!



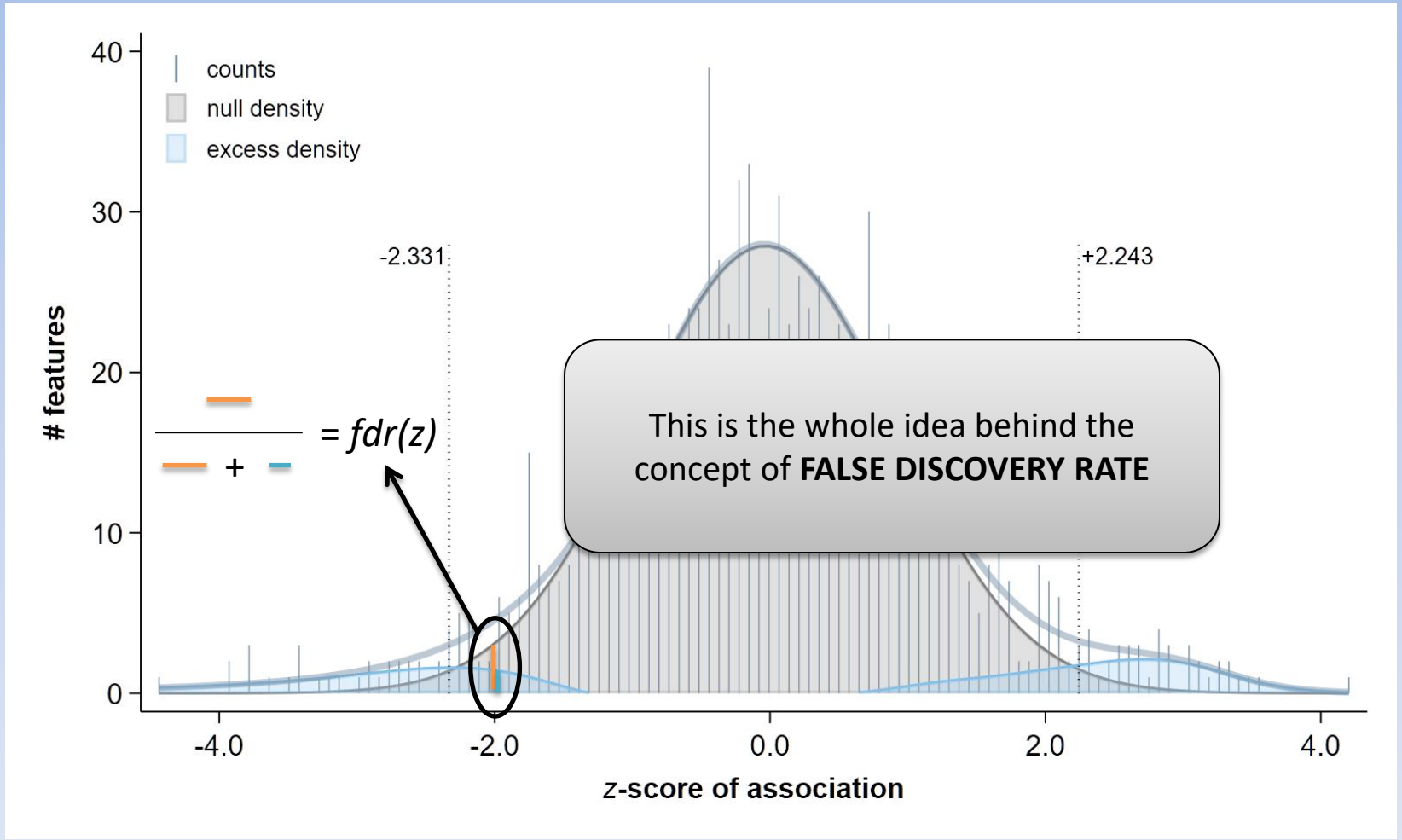
So much complexity, so little time

The idea: Fit a finite-mixture model to the association scores. Excess density is exciting!



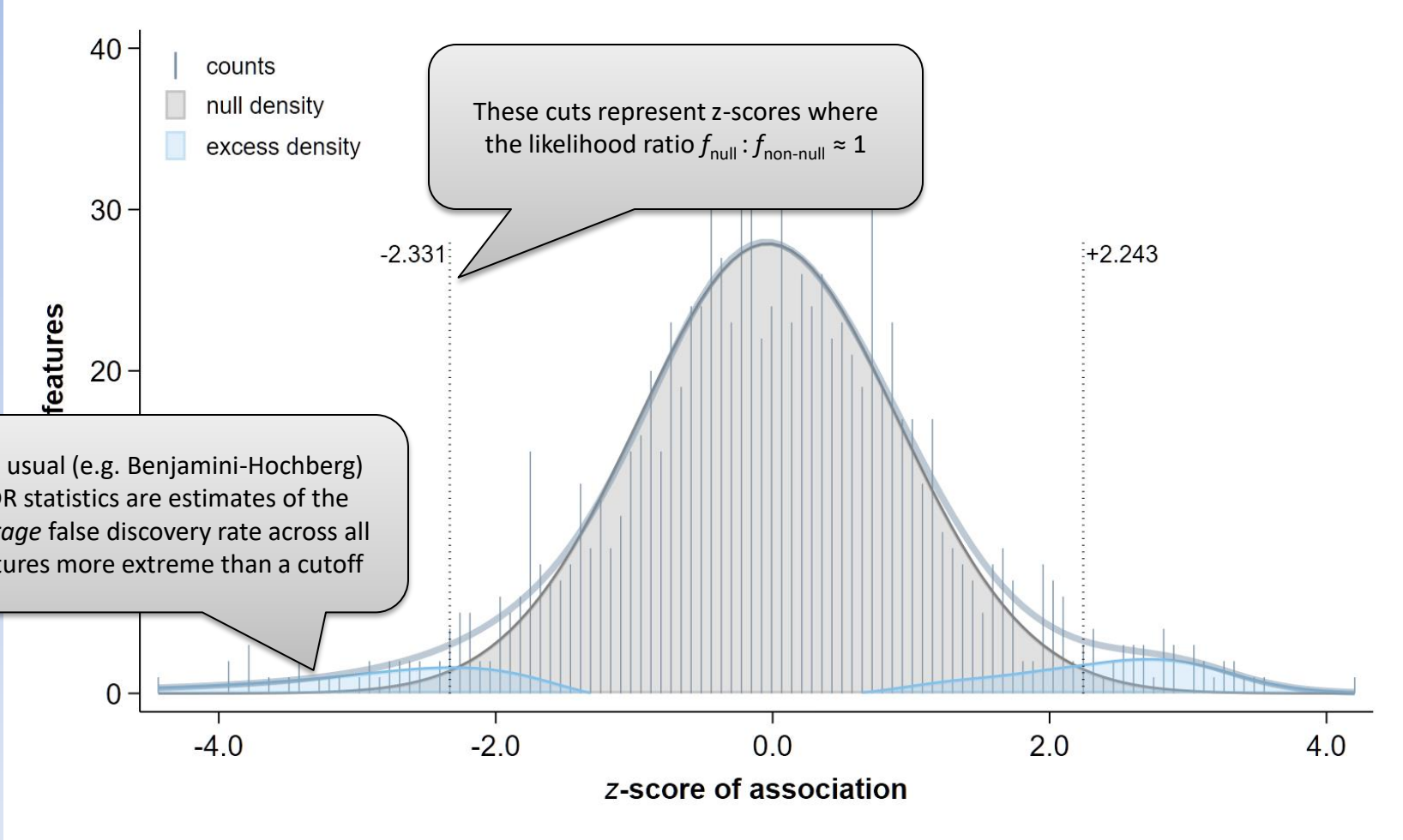
So much complexity, so little time

The idea: Fit a finite-mixture model to the association scores. Excess density is exciting!



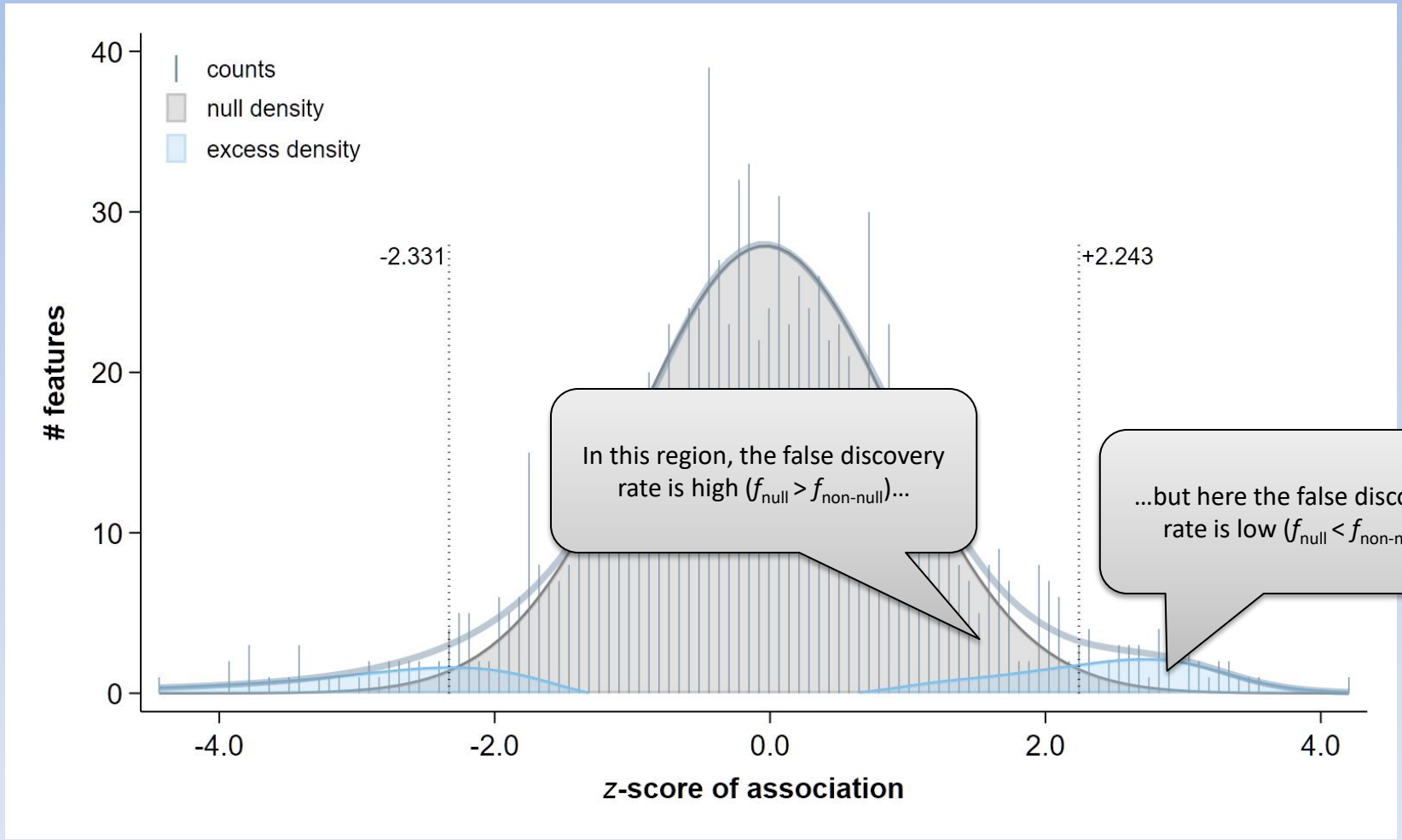
So much complexity, so little time

The idea: Fit a finite-mixture model to the association scores. Excess density is exciting!



So much complexity, so little time

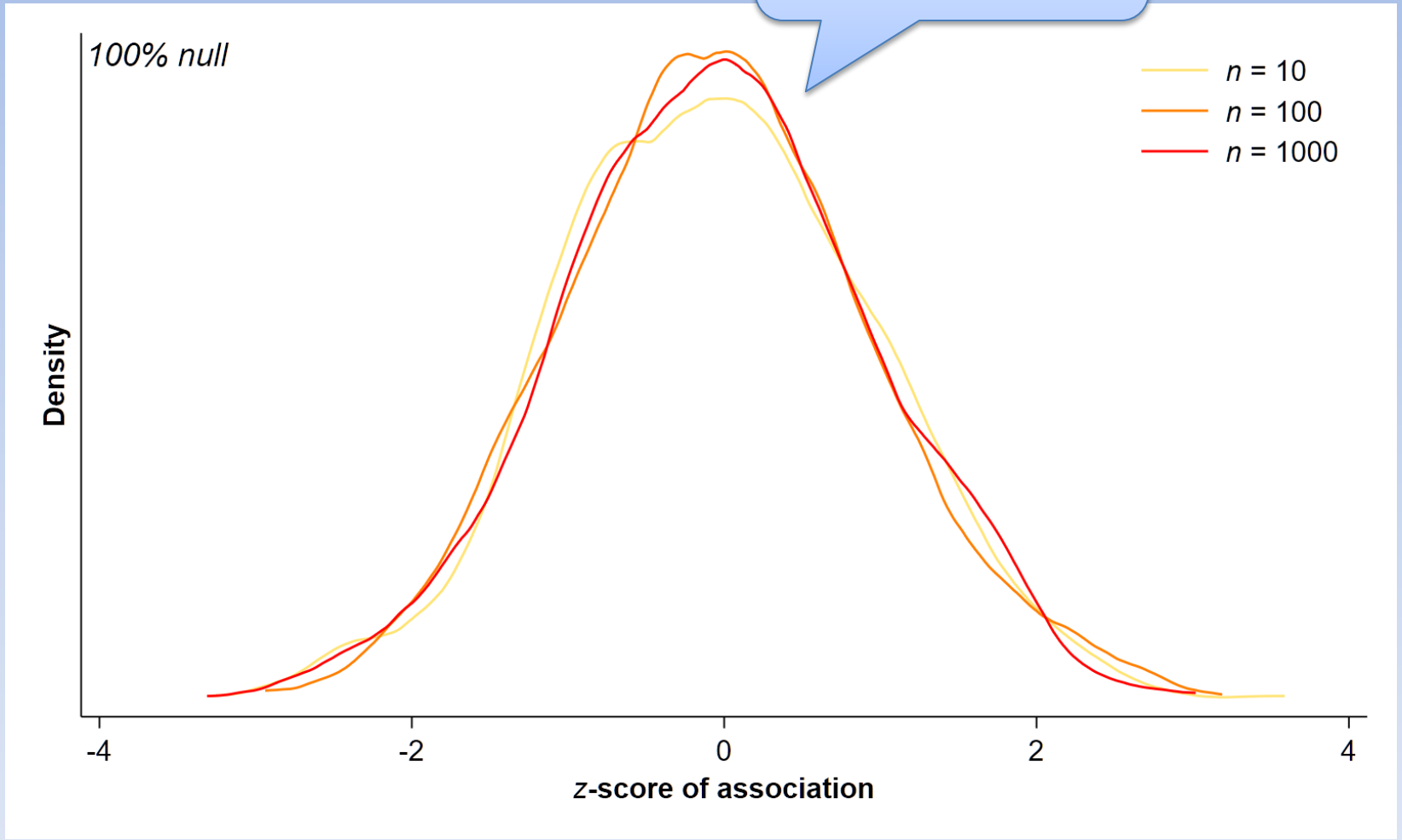
The idea: Fit a finite-mixture model to the association scores. Excess density is exciting!



So much complexity, so little time

How does sample size impact *fdr*?

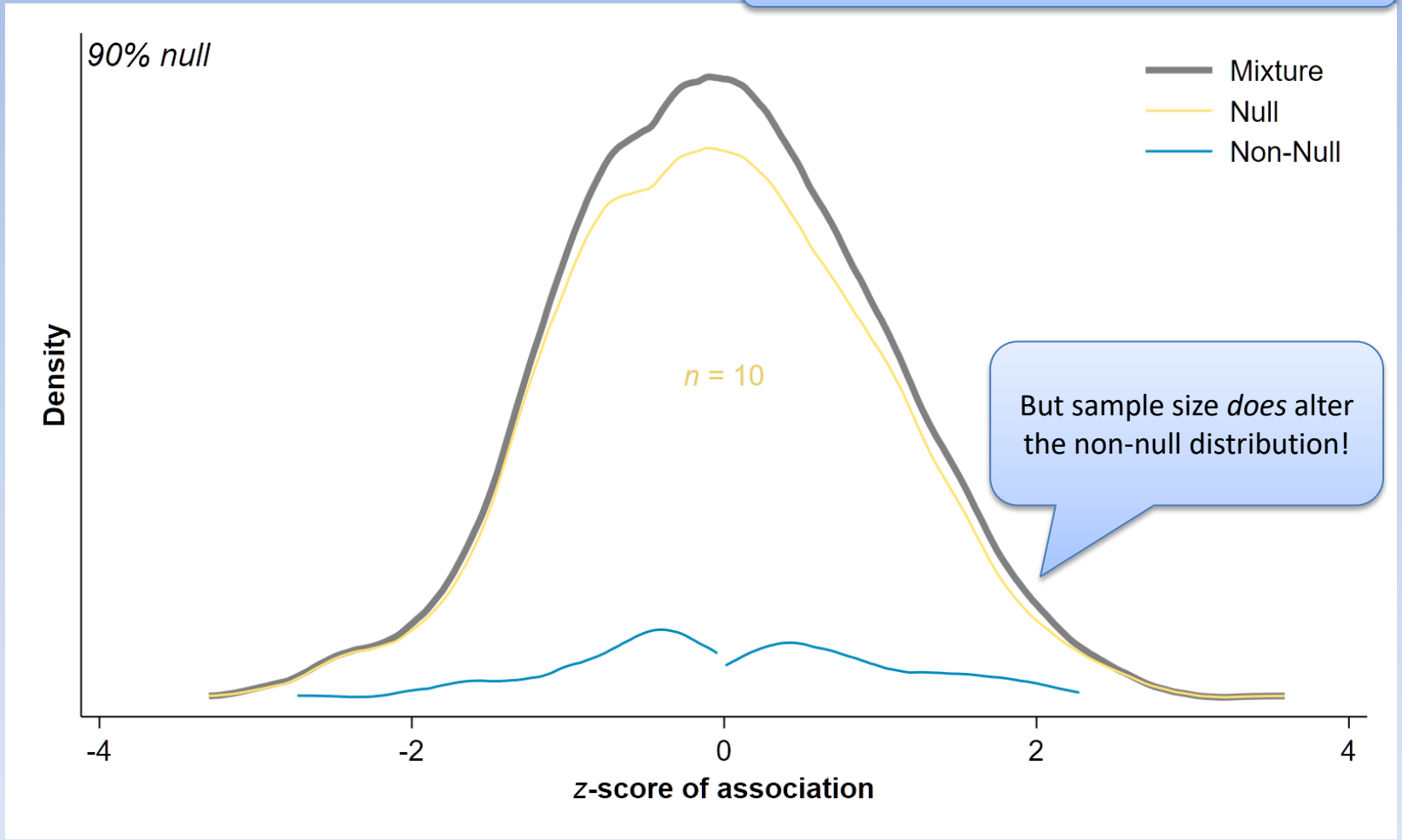
Sample size doesn't alter the null distribution at all



So much complexity, so little time

How does sample size impact *fdr*?

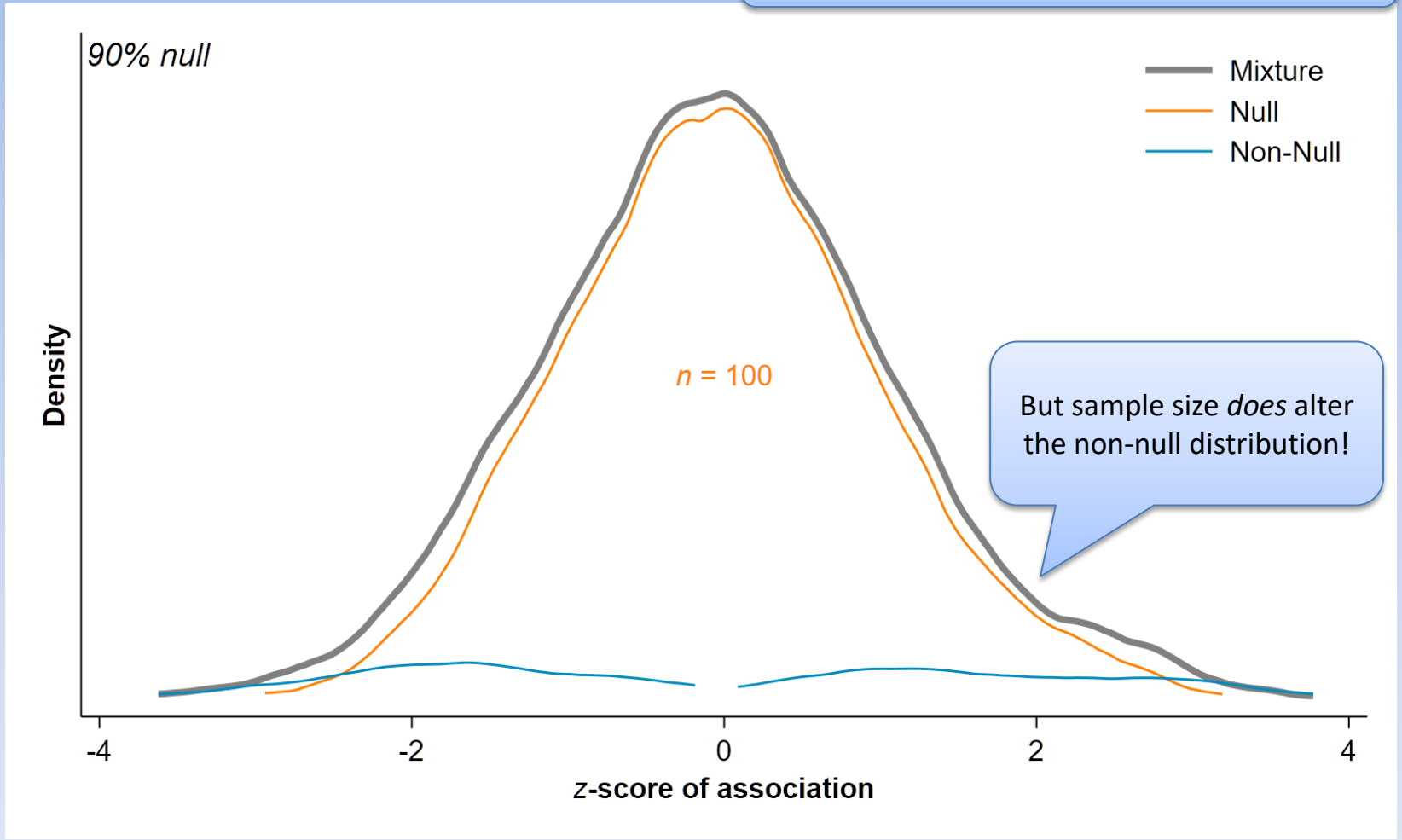
|effect size| = Lognormal(μ=0.3,σ=2)



So much complexity, so little time

How does sample size impact *fdr*?

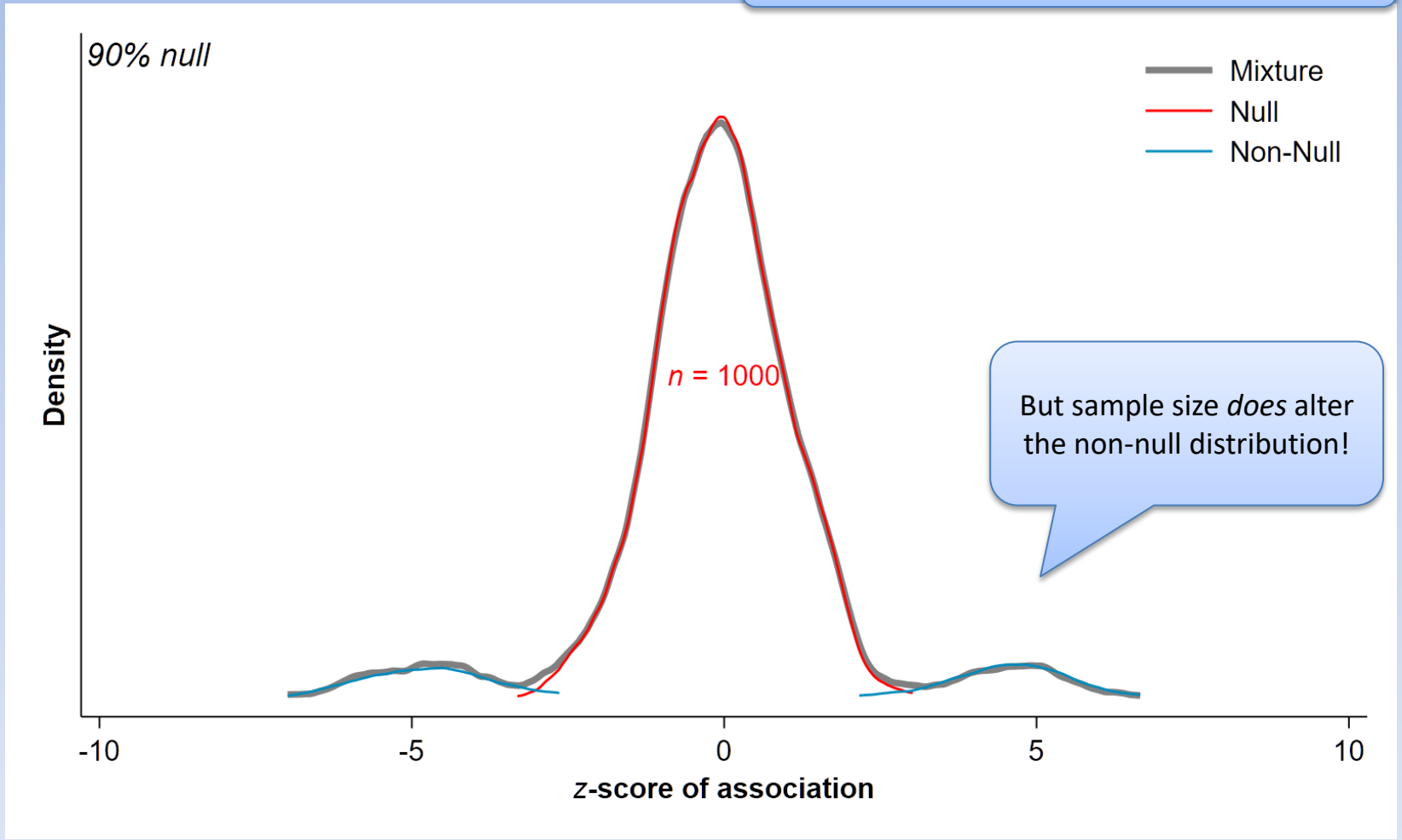
|effect size| = Lognormal(μ=0.3,σ=2)



So much complexity, so little time

How does sample size impact *fdr*?

|effect size| = Lognormal(μ=0.3,σ=2)



How do we "control" false discoveries? (We actually can't)

Concern about false discovery rates is fundamentally a Bayesian idea:

	+	−	
T	TP	FN	FN/T = β
F	FP	TN	FP/F = α
	FP/+ = *FDR*	FN/− = *FOR*	

Bayesian decision theory says to look at the **COLUMNS** — *Let's try not to regret too many of the decisions we make!*

Classical statistical theory says to look only at the **ROWS** — *Let's try to limit the rate of false assertions we make!*

How do we "control" false discoveries? (We actually can't)

Concern about false discovery rates is fundamentally a Bayesian idea:

	+	−	
T	TP	FN	FN/T = β
F	FP	TN	FP/F = α
	FP/+ = *FDR*	FN/− = *FOR*	

Bayesian decision theory says to look at the **COLUMNS** — *Let's try not to regret too many of the decisions we make!*

"FOR" = false omission rate

Classical statistical theory says to look only at the **ROWS** — *Let's try to limit the rate of false assertions we make!*

How do we "control" false discoveries? (We actually can't)

Concern about false discovery rates is fundamentally a Bayesian idea:

	+	−	
T	TP	FN	FN/T = β
F	FP	TN	FP/F = α
	FP/+ = *FDR*	FN/− = *FOR*	

It means we give up the concept of *control by design* (e.g. if T = 0 and we declare anything at all a "discovery" then our *FDR* = 100%), but in exchange we assert *local control* to base decisions on the data after we actually observe it. We can be **sensitive to the distribution** of **our findings** rather than fuss about adding a dirty dish to the Cupboard of Scientific Truths.

How do we "control" false discoveries? (We actually can't)

Concern about false discovery rates is fundamentally a Bayesian idea:

	+	−	
T	TP	FN	FN/T = β
F	FP	TN	FP/F = α
	FP/+ = *FDR*	FN/− = *FOR*	

It means we give up the concept of *control by design* (e.g. if T = 0 and we declare anything at all a "discovery" then our *FDR* = 100%), but in exchange we assert *local control* to base decisions on the data after we actually observe it. We can be **sensitive to the distribution** of **our findings** rather than fuss about adding a dirty dish to the Cupboard of Scientific Truths.

I would argue no such Cupboard exists.
Or if it does, it's all dirty dishes in there.

How do we "control" false discoveries? (We actually can't)

Concern about false discovery rates is fundamentally a Bayesian idea:

	+	−	
T	TP	FN	FN/T = β
F	FP	TN	FP/F = α
	FP/+ = *FDR*	FN/− = *FOR*	

It means we give up the concept of *control by design* (e.g. if T = 0 and we declare anything at all a "discovery" then our *FDR* = 100%), but in exchange we assert *local control* to base decisions on the data after we actually observe it. We can be **sensitive to the distribution** of **our findings** rather than fuss about adding a dirty dish to the Cupboard of Scientific Truths.

Bonferroni-style FWER control means refusing to decide on a set ("fail to reject") unless *all* rejections for the set are correct (at < α error rate), whereas Tukey-style ("higher criticism") control means accepting a set if *at least one* rejection is correct. **FDR is neither of those.**

How do we "control" false discoveries? (We actually can't)

Concern about false discovery rates is fundamentally a Bayesian idea:

	+	−	
T	TP	FN	FN/T = β
F	FP	TN	FP/F = α
	FP/+ = *FDR*	FN/− = *FOR*	

It means we give up the concept of *control by design* (e.g. if T = 0 and we declare anything at all a "discovery" then our *FDR* = 100%), but in exchange we assert *local control* to base decisions on the data after we actually observe it. We can be **sensitive to the distribution** of **our findings** rather than fuss about adding a dirty dish to the Cupboard of Scientific Truths.

Bonferroni-style FWER control means refusing to decide on a set ("fail to reject") unless *all* rejections for the set are correct (at < α error rate), whereas Tukey-style ("higher criticism") control means accepting a set if *at least one* rejection is correct. **FDR is neither of those.**

FDR is simply a *statistic* — a local estimate of the proportion of false discoveries among some set of discoveries. Not hypothesis-testing but *calibration* of the z-score distribution.

Using false discovery rate tolerance to estimate sample size

IDEA #1

- Assume a prior probability θ (proportion of the features you think are truly associated)
- Set β to shoot for a target power (1 − β)
- Given β, figure out how to adjust α to arrive at a particular (version of) *FDR*
- Use β and the adjusted α to estimate sample size for a single feature in the usual way

Works about like Bonferroni correction, just less!

Using false discovery rate tolerance to estimate sample size

IDEA #1

- Assume a prior probability θ (proportion of the features you think are truly associated)
- Set β to shoot for a target power (1 − β)
- Given β, figure out how to adjust α to arrive at a particular (version of) *FDR*
- Use β and the adjusted α to estimate sample size for a single feature in the usual way

$$adjusted\ \alpha = \frac{FDR}{1 - FDR} \cdot \frac{\theta}{1 - \theta} \cdot (1 - \beta)$$

Odds of false discovery: Tune to some value that you're comfortable with

Odds of a real association: Penalty for having a sparse population of good features

Probability of detecting a proposed effect size at this α

Using false discovery rate tolerance to estimate sample size

IDEA #1

- Assume a prior probability θ (proportion of the features you think are truly associated)
- Set β to shoot for a target power (1 − β)
- Given β, figure out how to adjust α to arrive at a particular (version of) *FDR*
- Use β and the adjusted α to estimate sample size for a single feature in the usual way

$$adjusted\ \alpha = \frac{FDR}{1 - FDR} \cdot \frac{\theta}{1 - \theta} \cdot (1 - \beta)$$

This really only makes sense for sparse scenarios (θ << 50%) and *FDR* < 50%. Perversely set both odds terms to 1 and get back an "adjusted α" of 0.8 (!)

Using false discovery rate tolerance to estimate sample size

IDEA #1

- Assume a prior probability θ (proportion of the features you think are truly associated)
- Set β to shoot for a target power (1 − β)
- Given β, figure out how to adjust α to arrive at a particular (version of) *FDR*
- Use β and the adjusted α to estimate sample size for a single feature in the usual way

$$adjusted\ \alpha = \frac{FDR}{1 - FDR} \cdot \frac{\theta}{1 - \theta} \cdot (1 - \beta)$$

EXAMPLE
Assume θ=10% (we doubt that too much of the panel is truly relevant) and 1 − β = 80% and set *FDR*=10% with the goal of being moderately but not overly stringent. The formula gives adjusted α=0.0099, which is nearly equivalent to the very weak Bonferroni adjustment of α' = 0.05/5. So powering in the usual way for a single association is equivalent to making the assumption that if 10% of our features have that magnitude of association, then our decision rule accepting 10% false discoveries will work for us as long as we do a FWER correction for 5 tests, regardless of the actual size of the panel.

Using false discovery rate tolerance to estimate sample size

IDEA #1

- Assume a prior probability θ (proportion of the features you think are truly associated)
- Set β to shoot for a target power (1 − β)
- Given β, figure out how to adjust α to arrive at a particular (version of) *FDR*
- Use β and the adjusted α to estimate sample size for a single feature in the usual way

$$adjusted\ \alpha = \frac{FDR}{1 - FDR} \cdot \frac{\theta}{1 - \theta} \cdot (1 - \beta)$$

The downside is that this is pretty gimmicky. It's quite difficult to justify the assumption that we can accurately guess what fraction of our data feature panel will have true associations of significant magnitude. For dense scenarios (e.g. θ ≈ 50%), stringent FDR rules are needed to arrive at reasonable values, and the formula ignores the curse of dimensionality by claiming the penalty for assuming 100K associated genes in a panel of 1M is qualitatively no different than assuming 1 associated outcome among a total of 10 that you'll check.

Using false discovery rate tolerance to estimate sample size

IDEA #1

- Assume a prior probability θ (proportion of the features you think are truly associated)
- Set β to shoot for a target power (1 − β)
- Given β, figure out how to adjust α to arrive at a particular (version of) *FDR*
- Use β and the adjust[illegible] in the usual way

EXAMPLE (continued)
Having obtained adjusted α=0.0099, now we need to assume that 10% of our features will at minimum have some appreciable magnitude of association. Since we're assuming some sparsity going in, choosing something modest like Cohen's *d*=0.3 may be acceptable (albeit unlikely to actually pan out). Calculating sample size for that yields n≈525 needed.

The downside is that this is pretty gimmicky. It's quite difficult to justify the assumption that we can accurately guess what fraction of our data feature panel will have true associations of significant magnitude. For dense scenarios (e.g. θ ≈ 50%), stringent FDR rules are needed to arrive at reasonable values, and the formula ignores the curse of dimensionality by claiming the penalty for assuming 100K associated genes in a panel of 1M is qualitatively no different than assuming 1 associated outcome among a total of 10 that you'll check.

Using false discovery rate tolerance to estimate sample size

IDEA #2

- Assume a prior probability θ (proportion of the features you think are truly associated)
- Use θ to convert *FDR* and *FOR* into α and β
- Use α and β to estimate sample size for a single feature in the usual way

Hey, it's the same table!

Using false discovery rate tolerance to estimate sample size

IDEA #2

- Assume a prior probability θ (proportion of the features you think are truly associated)
- Use θ to convert *FDR* and *FOR* into α and β
- Use α and β to estimate sample size for a single feature in the usual way

$$\alpha = \frac{FDR}{1 - FOR - FDR} \cdot \frac{\theta - FOR}{1 - \theta}$$

$$\beta = \frac{FOR}{1 - FOR - FDR} \cdot \frac{1 - FDR - \theta}{\theta}$$

Using false discovery rate tolerance to estimate sample size

IDEA #2

- Assume a prior probability θ (proportion of the features you think are truly associated)
- Use θ to convert *FDR* and *FOR* into α and β
- Use α and β to estimate sample size for a single feature in the usual way

$$\alpha = \frac{FDR}{1 - FOR - FDR} \cdot \frac{\theta - FOR}{1 - \theta}$$

$$\beta = \frac{FOR}{1 - FOR - FDR} \cdot \frac{1 - FDR - \theta}{\theta}$$

EXAMPLE
Assume θ=20% (we're somewhat hopeful our omics will offer something) and we don't want to miss too much of what's there, so we set *FOR*=5% (note that FOR < θ and 1 − *FDR* > θ are required for the conversion to work sensibly), but in compensation we're willing to accept *FDR*=20%. The formulas magically give α=0.05 and β=0.20. So powering in the usual way for a single association is equivalent to making the assumption that 20% of our features will have that magnitude of association, and that our decision rule accepting 20% false discoveries will lead to only 5% of our omissions being true associations.

Using false discovery rate tolerance to estimate sample size

IDEA #2

- Assume a prior probability θ (proportion of the features you think are truly associated)
- Use θ to convert *FDR* and *FOR* into α and β
- Use α and β to estimate sample size for a single feature in the usual way

$$\alpha = \frac{FDR}{1 - FOR - FDR} \cdot \frac{\theta - FOR}{1 - \theta}$$

$$\beta = \frac{FOR}{1 - FOR - FDR} \cdot \frac{1 - FDR - \theta}{\theta}$$

This is pretty gimmicky too. It's difficult to arrive at combinations of α and β that will fly with grant reviewers, and (as with IDEA #1) even more difficult to justify the assumption that we can accurately guess what fraction of our data feature panel will show associations of significant magnitude. For realistically sparse scenarios (e.g. θ < 10%), stringent FDR rules are needed to arrive at reasonable values.

Using false discovery rate tolerance to estimate sample size

IDEA #2

- Assume a prior probability θ (proportion of the features you think are truly associated)
- Use θ to convert *FDR* and *FOR* into α and β
- Use α and β to estimate sample size for a single feature in the usual way

EXAMPLE (continued)
Having obtained α=0.05 and β=0.20, now we need to assume that 20% of our features will at minimum have some appreciable magnitude of association. However, something like Cohen's *d*=0.5 is not going to fly! There's basically no way 20% of any omics panel will be that predictive; maybe *d*=0.2 is more realistic (altho probably still optimistic). Calculating sample size for that yields n≈800 needed.

This is pretty gimmicky too. It's difficult to arrive at combinations of α and β that will fly with grant reviewers, and (as with IDEA #1) even more difficult to justify the assumption that we can accurately guess what fraction of our data feature panel will show associations of significant magnitude. For realistically sparse scenarios (e.g. θ < 10%), stringent FDR rules are needed to arrive at reasonable values.

Using false discovery rate tolerance to estimate sample size

Surely we can do better!? Those first ideas are easy but arguably dorky.
Here's IDEA #3

- Simulate some "data" that looks like what we expect
- Parse the simulated finite mixture into components and calculate *fdr* values
- Calculate the *expected* false discovery rate $E_{non-null}[fdr]$ across non-null components
  - If this value is large, the features we want are mostly hiding amongst the nulls
  - If this value is small, the features we want are mostly separated from the nulls
- $E_{non-null}[fdr] < fdr_{cutoff}$ reflects a scenario where power is good
  (i.e. the "excess density" according to our cutoff is mostly outside the null density)
  - Iteratively adjust the sample size in the simulations until you obtain this!

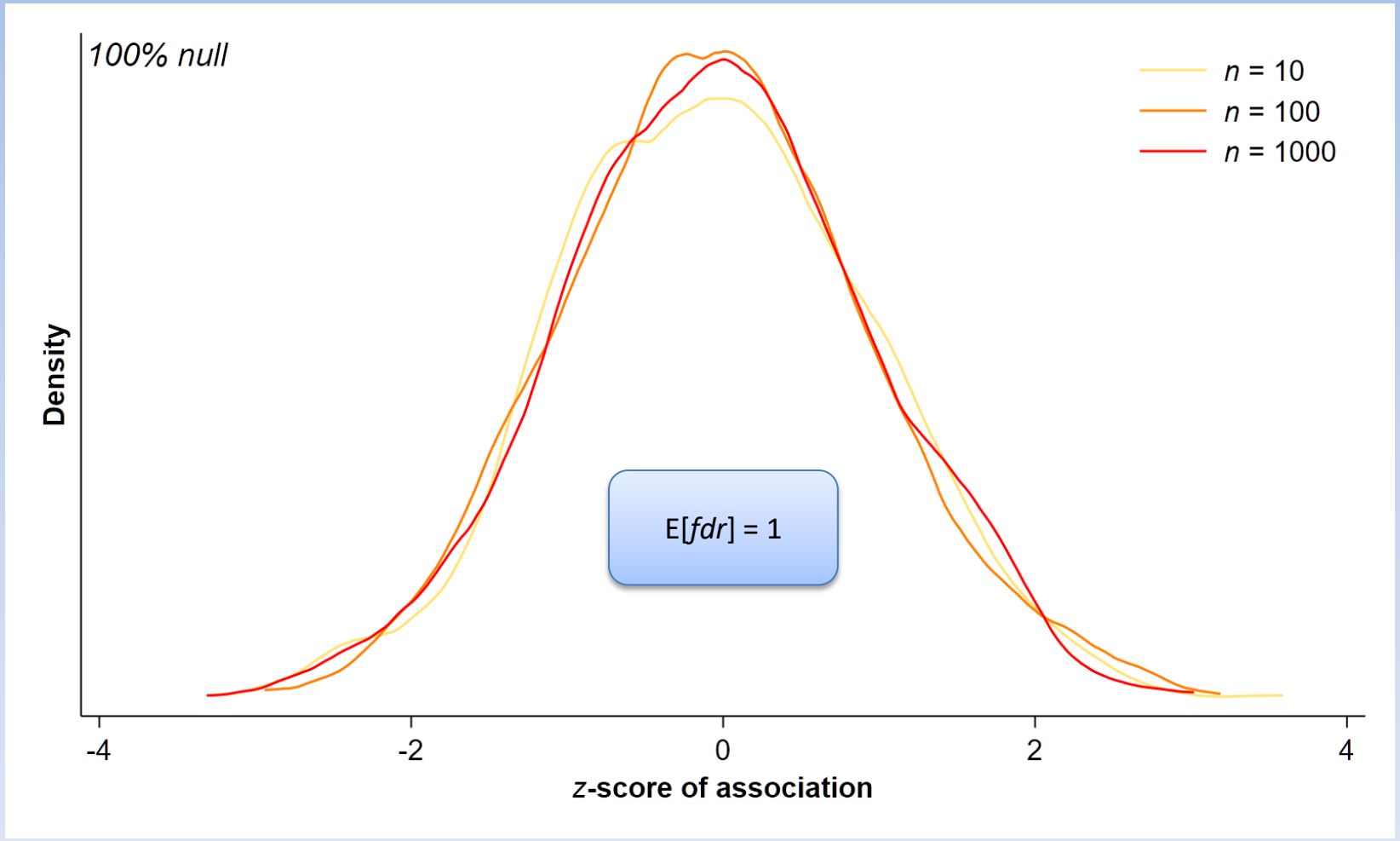
Using false discovery rate tolerance to estimate sample size

Surely we can do better!? Those first ideas are easy but arguably dorky.
Here's IDEA #3

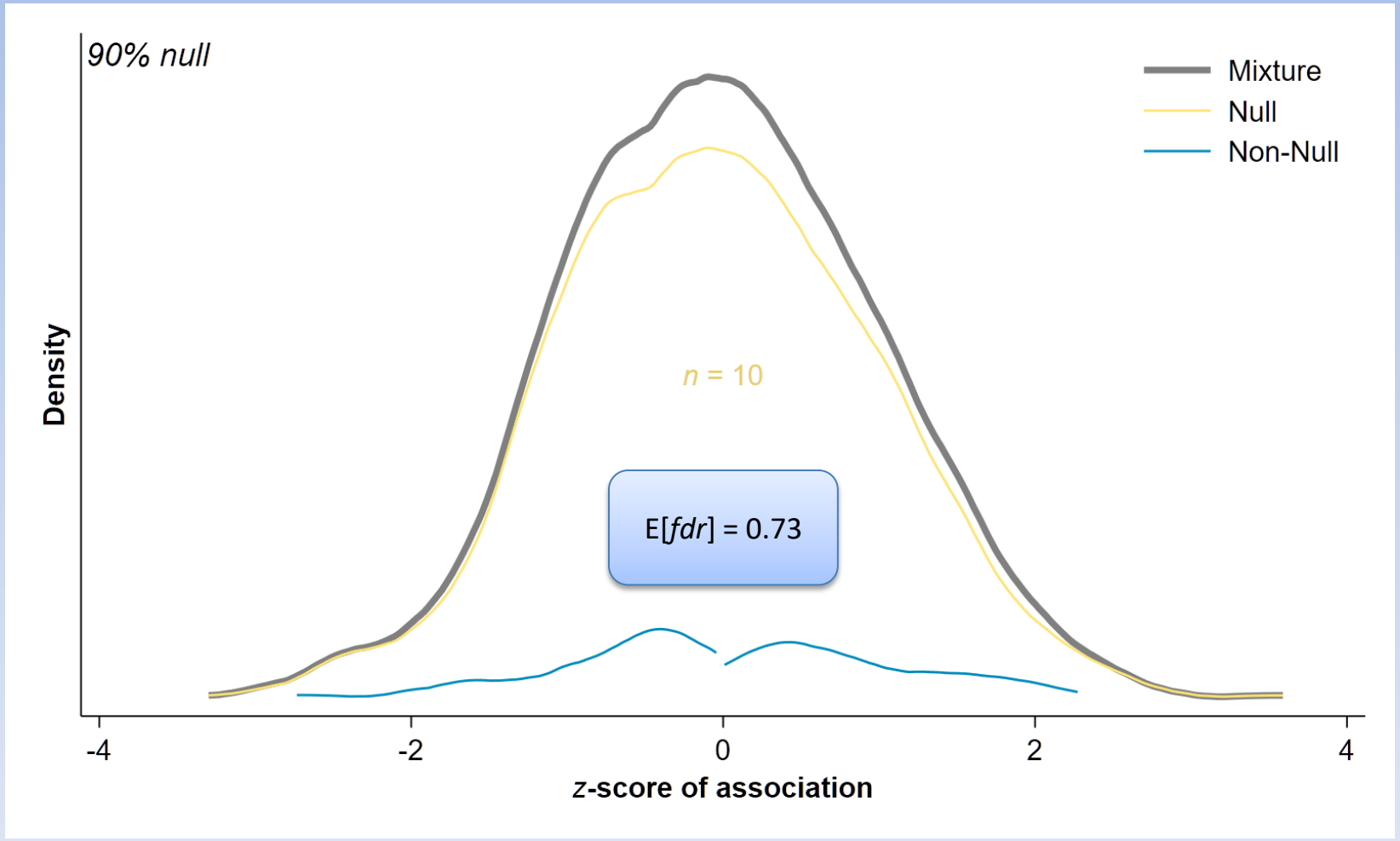
- Simulate some "data" that looks like what we expect
- Parse the simulated finite mixture into components and calculate *fdr* values
- Calculate the *expected* false discovery rate $E_{non-null}[fdr]$ across non-null components
  - If this value is large, the features we want are mostly hiding amongst the nulls
  - If this value is small, the features we want are mostly separated from the nulls
- $E_{non-null}[fdr] < fdr_{cutoff}$ reflects a scenario where power is good
  (i.e. the "excess density" according to our cutoff is mostly outside the null density)
  - Iteratively adjust the sample size in the simulations until you obtain this!

EXAMPLE
For the simulations we did above, let's try this! In those, we simulated 1000 features with a 9:1 split of null:non-null features, and in the latter the median effect size was 0.3 (with large variance). As a sanity check, we also tried it with the 100% null simulation and got expected *fdr* of 1 as we should. Values were > 10% for sample sizes $n < 1000$ (including one at $n=250$ not shown, with expected *fdr* of 0.37), suggesting that $n \approx 1000$ is needed if we want to use *fdr* = 10% as a cutoff for discoveries there.

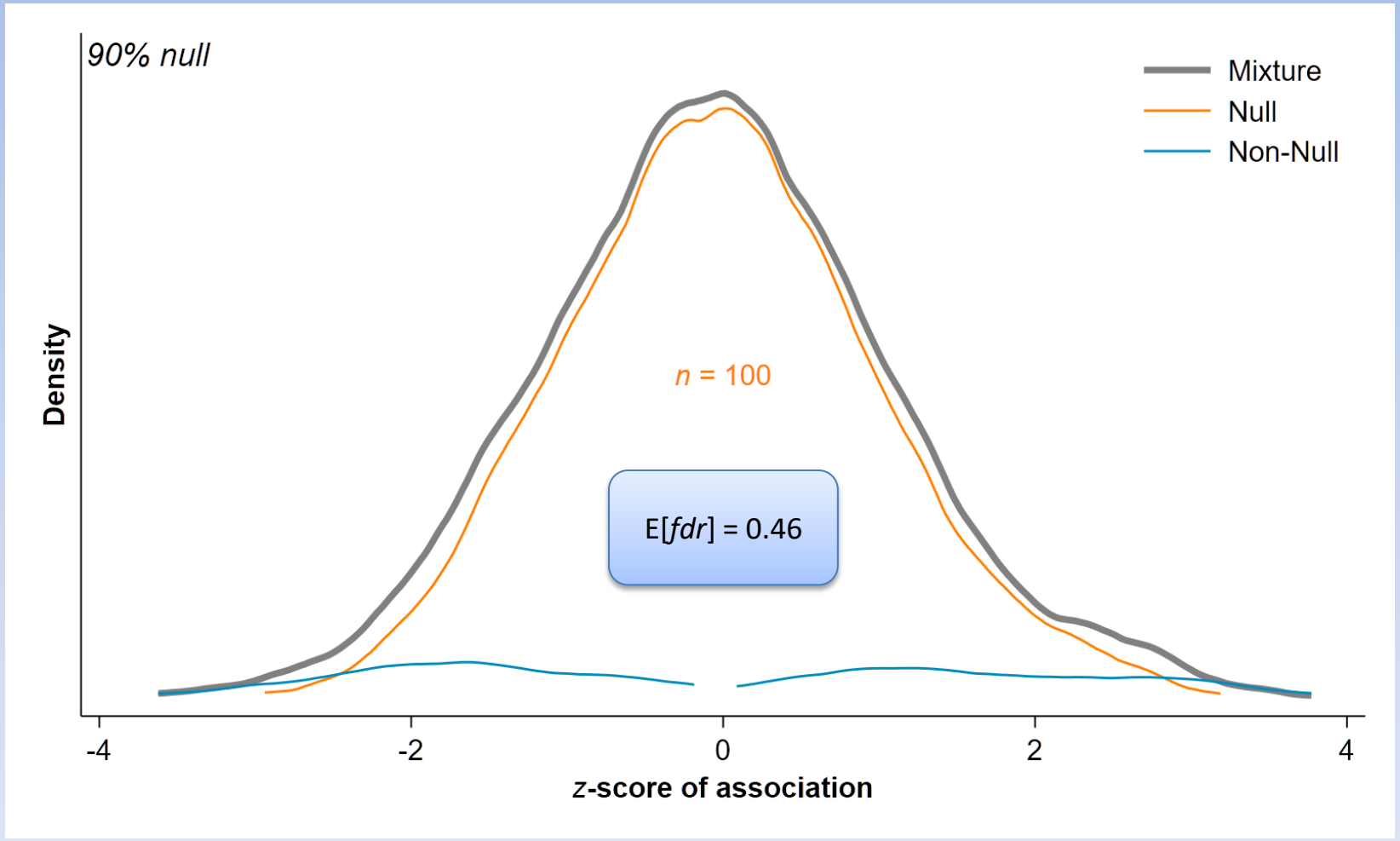
Using false discovery rate tolerance to estimate sample size



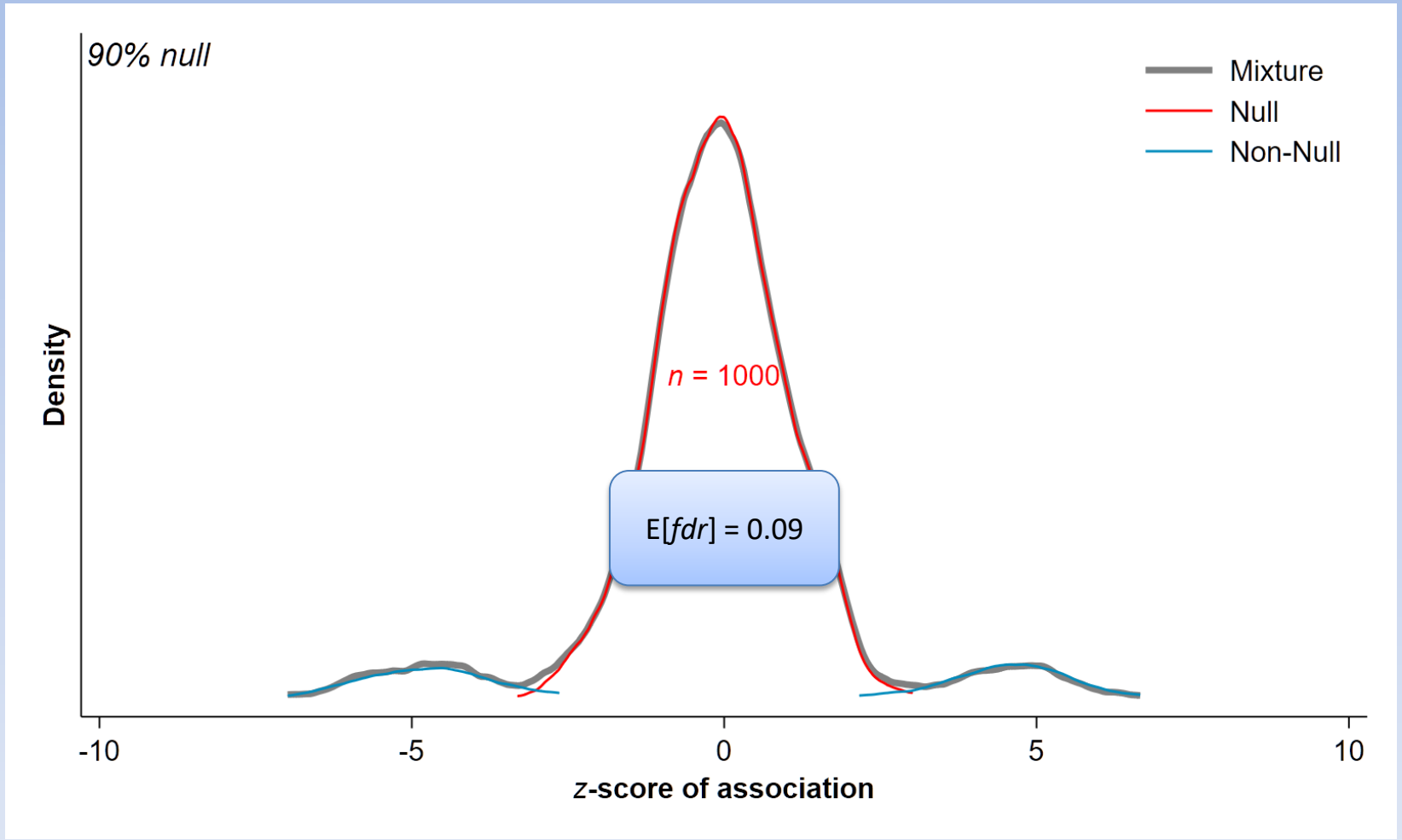
Using false discovery rate tolerance to estimate sample size



Using false discovery rate tolerance to estimate sample size



Using false discovery rate tolerance to estimate sample size



What about a fully Bayesian approach?

What *thinking distributionally* means is that we don't have to view the high-dimensional dataset as a bunch of different things. We can view it as *one* thing — an average feature — that has a distribution attached to it.

What about a fully Bayesian approach?

What *thinking distributionally* means is that we don't have to view the high-dimensional dataset as a bunch of different things. We can view it as *one* thing — an average feature — that has a distribution attached to it.

For the purpose of estimating sample size, it's the average feature that's most relevant to us. We also don't want to pretend that every associated feature has the same effect size; that the *average* one has some particular effect size is a more natural assumption.

What about a fully Bayesian approach?

What *thinking distributionally* means is that we don't have to view the high-dimensional dataset as a bunch of different things. We can view it as *one* thing — an average feature — that has a distribution attached to it.

For the purpose of estimating sample size, it's the average feature that's most relevant to us. We also don't want to pretend that every associated feature has the same effect size; that the *average* one has some particular effect size is a more natural assumption.

We can also define "average" however we like via choice of prior distribution. For example, use the same reasoning as for the identifying assumption in the finite-mixture analysis of a small "null window" near zero where only null associations live, we can say that any z-score too close to zero is automatically discounted.

What about a fully Bayesian approach?

What *thinking distributionally* means is that we don't have to view the high-dimensional dataset as a bunch of different things. We can view it as *one* thing — an average feature — that has a distribution attached to it.

For the purpose of estimating sample size, it's the average feature that's most relevant to us. We also don't want to pretend that every associated feature has the same effect size; that the *average* one has some particular effect size is a more natural assumption.

We can also define "average" however we like via choice of prior distribution. For example, use the same reasoning as for the identifying assumption in the finite-mixture analysis of a small "null window" near zero where only null associations live, we can say that any z-score too close to zero is automatically discounted.

Also, from a Bayesian point of view a two-sided alternative hypothesis makes little sense. Placing more prior weight near zero is a more natural way to express directional uncertainty. Once we have observed a direction, the power we care about is the power to assign high probability to the *observed* direction relative to the opposite direction.

What about a fully Bayesian approach?

The most natural Bayesian analogue to study power is *(pre)posterior risk* — your (expected) uncertainty (under e.g. squared-error loss) after having seen and analyzed the data. Adopting the posterior variance of the association as the posterior risk is IDEA #4

- Assume a normal-normal model: $p(\delta|y) \sim N(\delta|\theta, (1+r)^2\sigma^2/nr) \cdot N(\theta|\mu_\delta, \tau^2)$
- Calculate the posterior risk given assumptions for (σ, τ)
- Figure out how the posterior risk varies as a function of *n*
- Justify a particular value of posterior risk (lower is better!) and target *n* to achieve that

What about a fully Bayesian approach?

The most natural Bayesian analogue to study power is *(pre)posterior risk* — your (expected) uncertainty (under e.g. squared-error loss) after having seen and analyzed the data. Adopting the posterior variance of the association as the posterior risk is IDEA #4

- Assume a normal-normal model: $p(\delta|y) \sim N(\delta|\theta, (1+r)^2\sigma^2/nr) \cdot N(\theta|\mu_\delta, \tau^2)$
- Calculate the posterior risk given assumptions for (σ, τ)
- Figure out how the posterior risk varies as a function of *n*
- Justify a particular value of posterior risk (lower is better!) and target *n* to achieve that

$$posterior\ risk \approx \sigma_\delta^2|y = \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}\frac{r}{(1+r)^2}}$$

(same scaling to total as before)

$$n \approx \left(\frac{\sigma}{\sigma_\delta|y}\right)^2 \left(1 - \left(\frac{\sigma_\delta|y}{\tau}\right)^2\right)^+ \left(\frac{(1+r)^2}{r}\right)$$

sample size is proportional to the square of the ratio of sampling noise to posterior parameter uncertainty (no mention of "error rates" — irrelevant! the uncertainty captures everything we need to know)

proportionality factor depends on how large our prior uncertainty was — if we were already pretty close so that $\tau \approx \sigma_\delta|y$ then we don't need more research, but if we were far off as $\tau >> \sigma_\delta|y$ then we need a lot more

What about a fully Bayesian approach?

The most natural Bayesian analogue to study power is *(pre)posterior risk* — your (expected) uncertainty (under e.g. squared-error loss) after having seen and analyzed the data. Adopting the posterior variance of the association as the posterior risk is IDEA #4

- Assume a normal-normal model: $p(\delta|y) \sim N(\delta|\theta, (1+r)^2\sigma^2/nr) \cdot N(\theta|\mu_\delta, \tau^2)$
- Calculate the posterior risk given assumptions for (σ, τ)
- Figure out how the posterior risk varies as a function of *n*
- Justify a particular value of posterior risk (lower is better!) and target *n* to achieve that

$$posterior\ risk \approx \sigma_\delta^2|y = \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}\frac{r}{(1+r)^2}}$$

Doesn't depend on the data (*y*)!
Normal-normal model is nice this way

$$n \approx \left(\frac{\sigma}{\sigma_\delta|y}\right)^2 \left(1 - \left(\frac{\sigma_\delta|y}{\tau}\right)^2\right)^+ \left(\frac{(1+r)^2}{r}\right)$$

As τ grows large, the calculation approaches the usual frequentist one (where the size of $\sigma_\delta|y$ we demand was specified as a function of α and β)

What about a fully Bayesian approach?

The most natural Bayesian analogue to study power is *(pre)posterior risk* — your (expected) uncertainty (under e.g. squared-error loss) after having seen and analyzed the data. Adopting the posterior variance of the association as the posterior risk is IDEA #4

- Assume a normal-normal model: $p(\delta|y) \sim N(\delta|\theta, (1+r)^2\sigma^2/nr) \cdot N(\theta|\mu_\delta, \tau^2)$
- Calculate the posterior risk given assumptions for (σ, τ)
- Figure out how the posterior risk varies as a function of *n*
- Justify a particular value of posterior risk (lower is better!) and target *n* to achieve that

$$posterior\ risk \approx \sigma_\delta^2|y = \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}\frac{r}{(1+r)^2}}$$

EXAMPLE
In the simulations we did above, we assumed a typical effect size of 0.3, so if the standard error is about 15% of the standard deviation then a 95% confidence interval would just touch zero, but the Bayes factor for such a tight decision is at most 4:1, which feels risky. We would prefer our posterior risk to feel something more like the confidence level, e.g. Bayes factor of 20:1 or more. That means we want our standard error at least 30-50% smaller, such as 5-10% of the standard deviation at most; let's pick 8%. We'll assume the variance of effect sizes across all effects is much larger than the standard error of any single effect. If we have equal group sample sizes this gives n≈625 as the recommended total.

$$\left(\left(\frac{\sigma_\delta|y}{\tau}\right)^2\right)^+ \left(\frac{(1+r)^2}{r}\right)$$

What about a fully Bayesian approach?

The most natural Bayesian analogue to study power is *(pre)posterior risk* — your (expected) uncertainty (under e.g. squared-error loss) after having seen and analyzed the data. Adopting the posterior variance of the association as the posterior risk is IDEA #4

- Assume a normal-normal model: $p(\delta|y) \sim N(\delta|\theta, (1+r)^2\sigma^2/nr) \cdot N(\theta|\mu_\delta, \tau^2)$
- Calculate the posterior risk given assumptions for (σ, τ)
- Figure out how the posterior risk varies as a function of *n*
- Justify a particular value of posterior risk (lower is better!) and target *n* to achieve that

$$posterior\ risk \approx \sigma_\delta^2|y = \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}\frac{r}{(1+r)^2}}$$

$$n \approx \left(\frac{\sigma}{\sigma_\delta|y}\right)^2 \left(1 - \left(\frac{\sigma_\delta|y}{\tau}\right)^2\right)^+ \left(\frac{(1+r)^2}{r}\right)$$

This is sorta Bayesian but feels pretty simplistic. All we're really saying is that we want all of the standard errors to be quite small relative to the standard deviations. Nothing about effect sizes or estimated discovery probability that reviewers will want to hear about.

What about a fully Bayesian approach?

A more principled Bayesian approach is to find an expression for *expected power* at a given sample size and then vary the sample size to see where expected power gets good enough. The expectation can be with respect to our prior belief that an observed effect of given size will have the correct sign. (This may vary with the size! More on that later.)
That's IDEA #5

- Assume a normal prior distribution for δ given a location: $p(\delta|\theta) \sim N(\delta|\theta, \tau^2)$
- Calculate the expected power at δ for a given α and *n*, assuming δ=θ with variance τ^2
- Vary *n* (you could also vary α for fixed *n*, as with FDR approaches) to find good power

What about a fully Bayesian approach?

A more principled Bayesian approach is to find an expression for *expected power* at a given sample size and then vary the sample size to see where expected power gets good enough. The expectation can be with respect to our prior belief that an observed effect of given size will have the correct sign. (This may vary with the size! More on that later.)
That's <u>IDEA #5</u>

- Assume a normal prior distribution for δ given a location: $p(\delta|\theta) \sim N(\delta|\theta, \tau^2)$
- Calculate the expected power at δ for a given α and *n*, assuming δ=θ with variance τ^2
- Vary *n* (you could also vary α for fixed *n*, a[illegible]th FDR approaches) to find good po[illegible]er

Wait, we're still caring about α??! Yes — that's because we're still describing frequentist power, just going about it in a Bayesian way, factoring in uncertainty

A thorn in our side is we don't know how to pick this. But we can use normal distribution theory! If δ is normal with mean θ >0 and standard deviation τ, then the z-score for a realized value δ_0 is $z_0=(\delta_0-\theta)/\tau$; choosing $\delta_0=0$ gives $z_0=-\theta/\tau$ or equivalently $\tau=-\theta/z_0$ where $z_0=\Phi^{-1}(P(\delta<\delta_0=0))$, so we can estimate τ by guessing how often δ would be negative

What about a fully Bayesian approach?

A more principled Bayesian approach is to find an expression for *expected power* at a given sample size and then vary the sample size to see where expected power gets good enough. The expectation can be with respect to our prior belief that an observed effect of given size will have the correct sign. (This may vary with the size! More on that later.)
That's IDEA #5

- Assume a normal prior distribution for δ given a location: $p(\delta|\theta) \sim N(\delta|\theta, \tau^2)$
- Calculate the expected power at δ for a given α and *n*, assuming δ=θ with variance τ^2
- Vary *n* (you could also vary α for fixed *n*, as with FDR approaches) to find good power

$$E_P[power](\alpha, n, r) \approx \int \Phi\left(\frac{\sqrt{nr}}{1+r}\frac{\delta}{\sigma} - z_\alpha\right) dP(\delta|\theta)$$

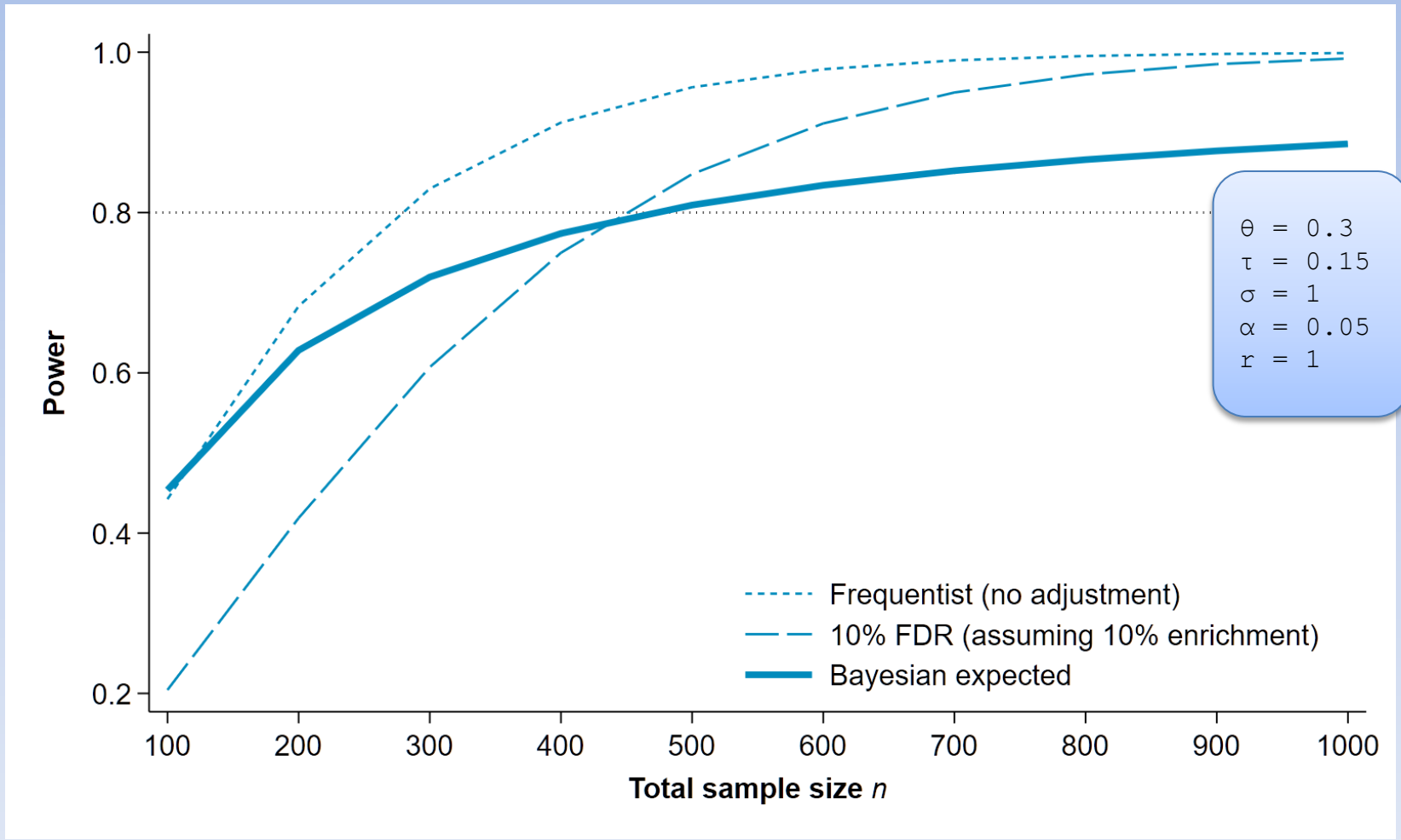
$$= \Phi\left(\frac{\delta - z_\alpha \sigma'}{\sqrt{\sigma'^2 + \tau^2}}\right)$$

After some algebra… closed-form!

$$\sigma'^2 = \frac{\sigma^2}{n}\frac{(1+r)^2}{r}$$

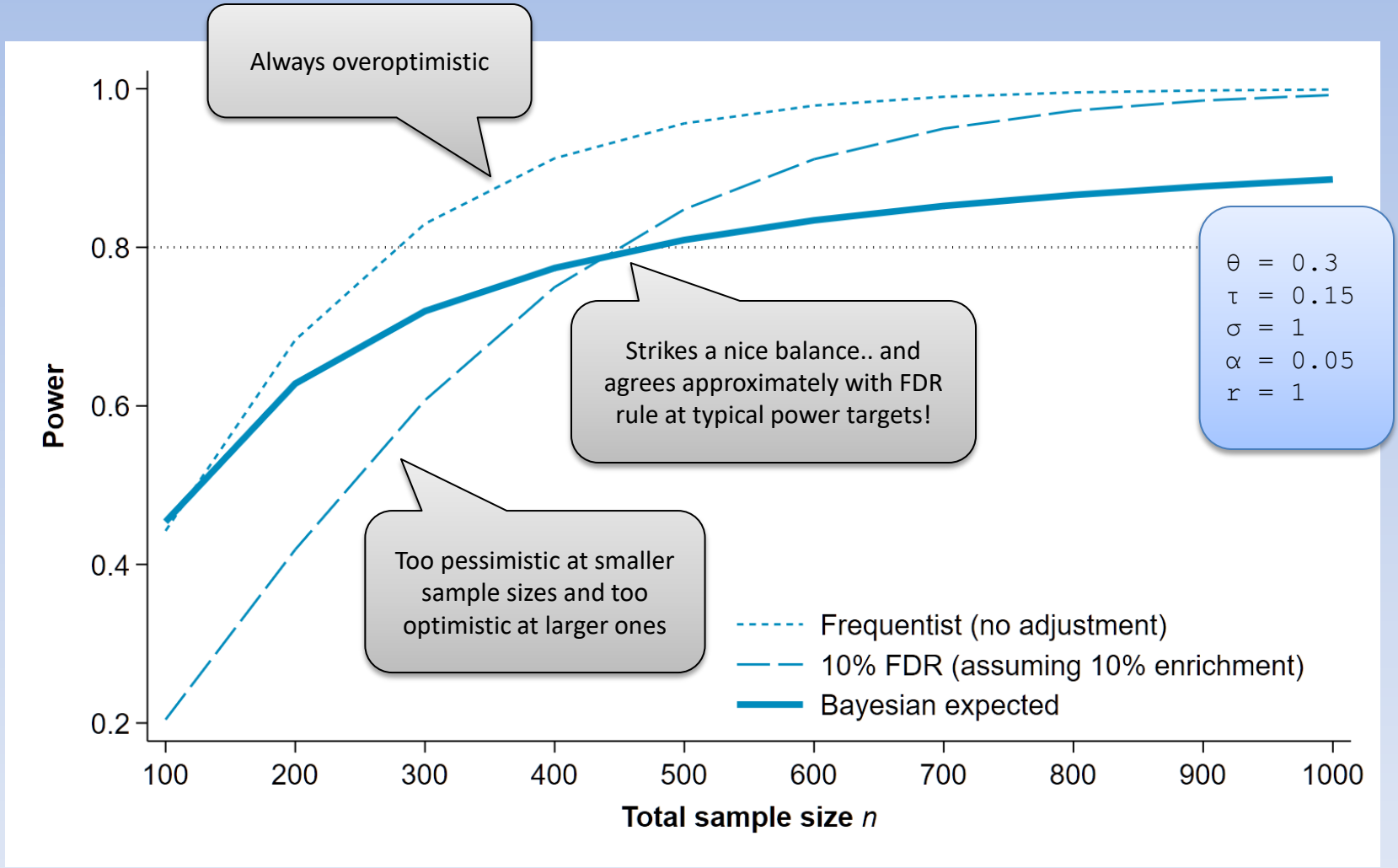
What about a fully Bayesian approach?

Trying out IDEA #5…



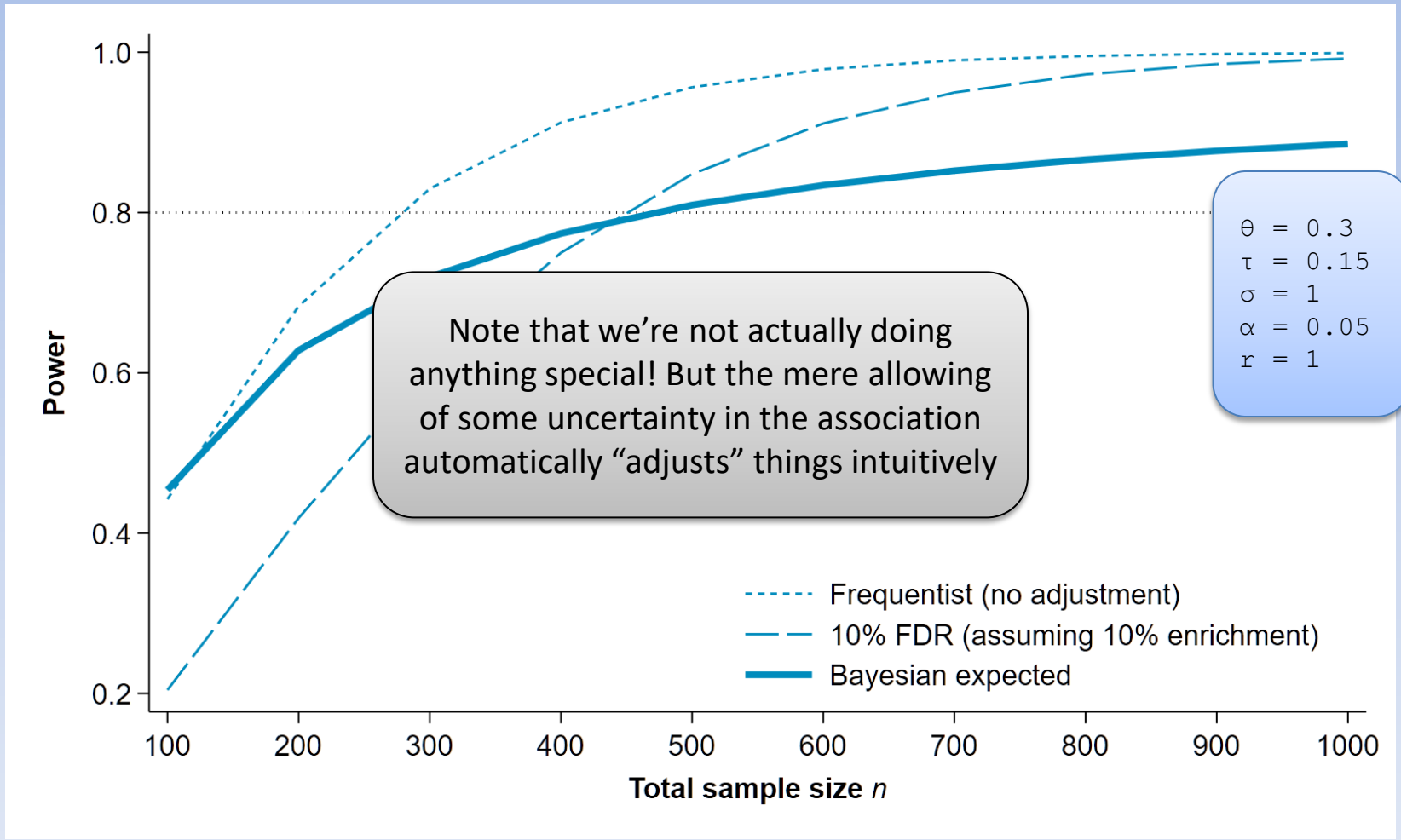
What about a fully Bayesian approach?

Trying out IDEA #5…



What about a fully Bayesian approach?

Trying out IDEA #5…



What about a fully Bayesian approach?

The last idea didn't actually take things far enough. What we really should do is calculate an expectation of expected power, where on top of averaging over uncertainty in effect size estimates, we also average over *uncertainty in the range of effect sizes* across the features. This is our (final) IDEA #6

- Assume a distribution for the effect sizes across features
- Assume a distribution for the within-feature uncertainty in the effect (e.g. confounding)
- Assume a distribution for the per-feature standard deviation in each group
- Assume a distribution for the per-feature sample size loss (missing data) in each group
- Integrate over everything and report expected power *for the entire experiment*

What about a fully Bayesian approach?

The last idea didn't actually take things far enough. What we really should do is calculate an expectation of expected power, where on top of averaging over uncertainty in effect size estimates, we also average over *uncertainty in the range of effect sizes* across the features. This is our (final) IDEA #6

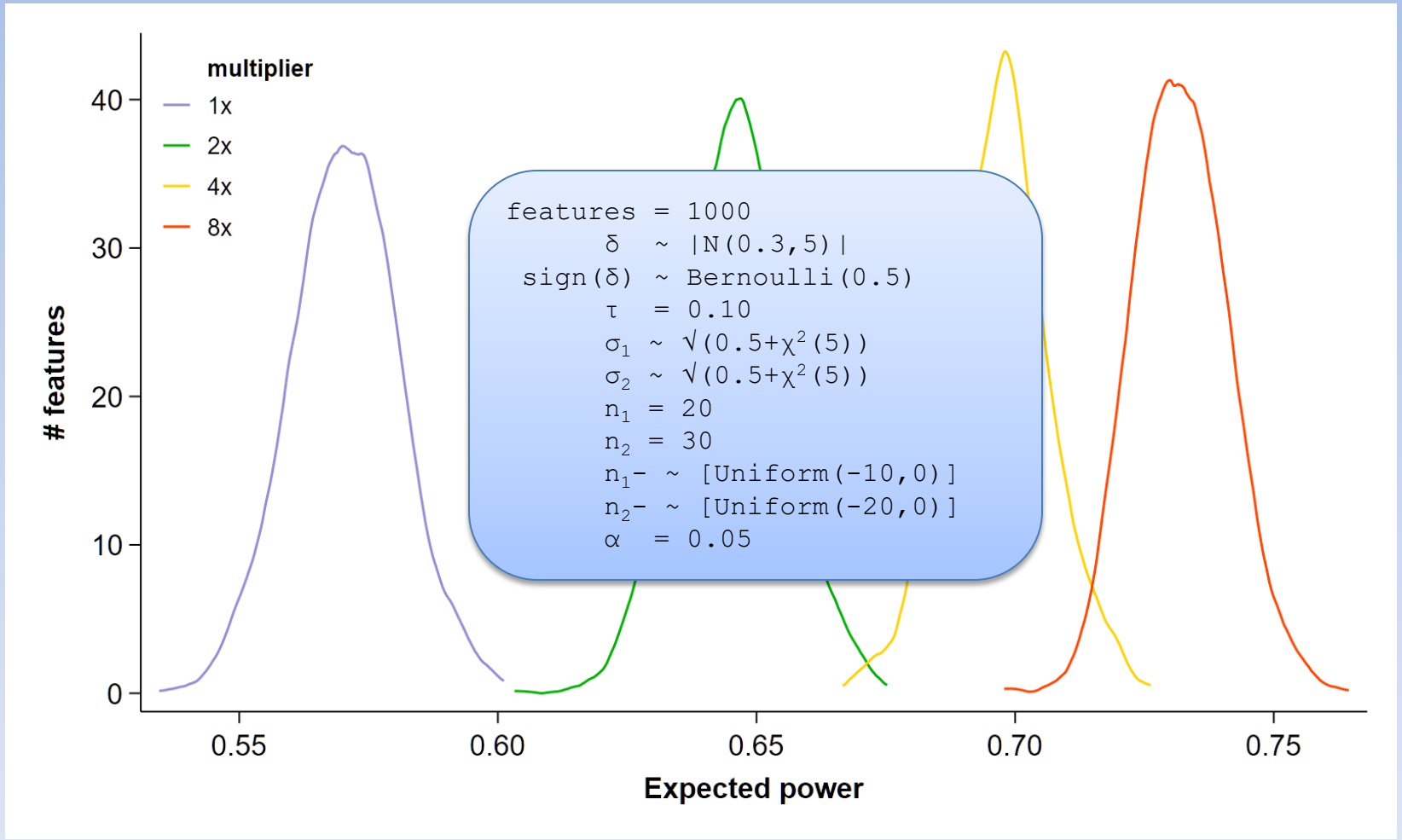
- Assume a distribution for the effect sizes across features
- Assume a distribution for the within-feature uncertainty in the effect (e.g. confounding)
- Assume a distribution for the per-feature standard deviation in each group
- Assume a distribution for the per-feature sample size loss (missing data) in each group
- Integrate over everything and report expected power *for the entire experiment*

$$\int\int\int\int\int\int Power(H|\delta,\tau,\sigma_1,\sigma_2,n_1^-,n_2^-)dP(\delta,\tau,\sigma_1,\sigma_2,n_1^-,n_2^-)$$

Essentially this (ugh)

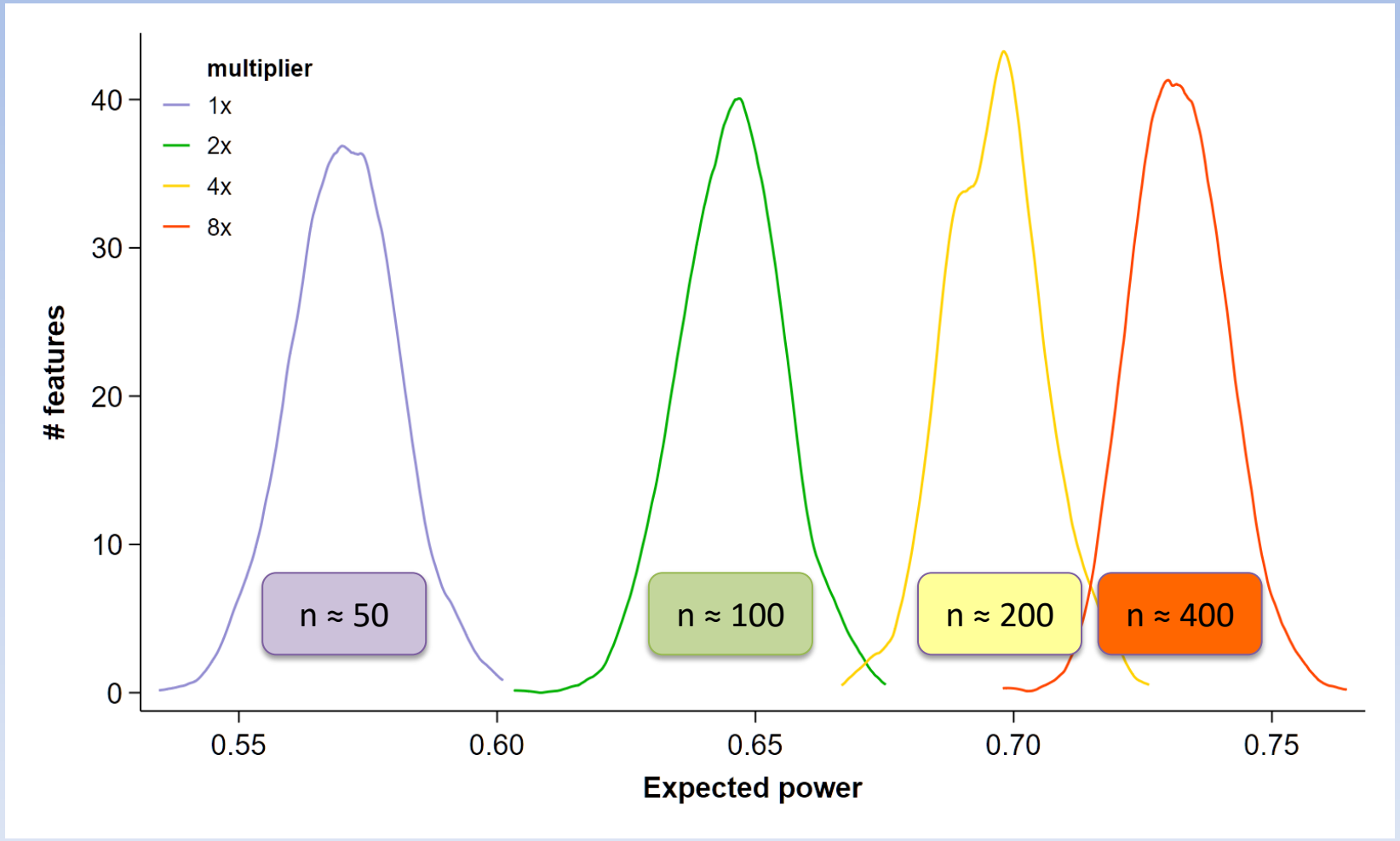
What about a fully Bayesian approach?

IDEA #6 is very difficult to implement analytically, but easy to simulate…



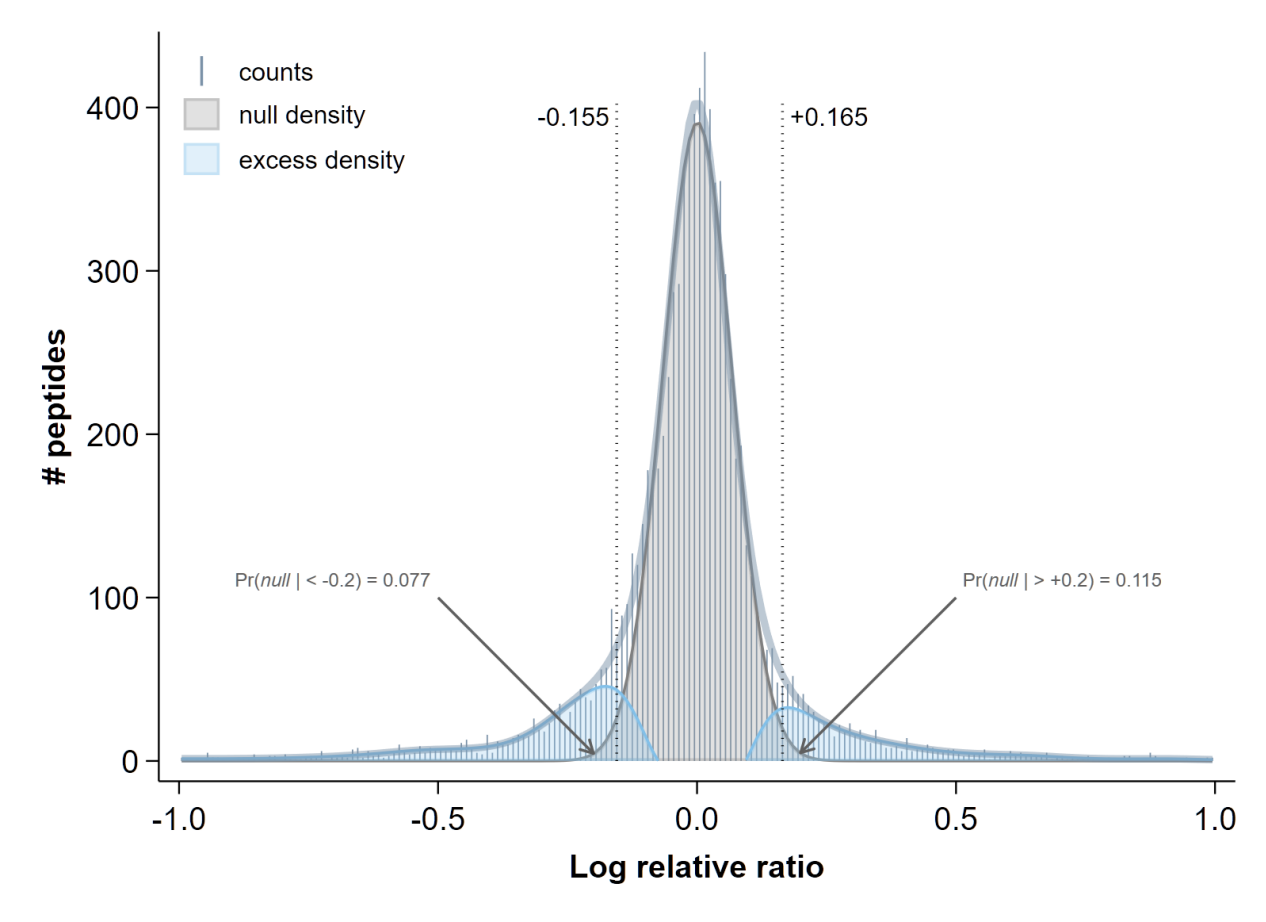
What about a fully Bayesian approach?

IDEA #6 is very difficult to implement analytically, but easy to simulate…



What about a fully Bayesian approach?

In fact, PASS has a set of modules ("Assurance") that will do this! (In fewer dimensions…)
As a test of principle, I applied it to some real proteomics data (8952 peptides) that I have:



What about a fully Bayesian approach?

In fact, PASS has a set of modules ("Assurance") that will do this! (In fewer dimensions…)
As a test of principle, I applied it to some real proteomics data (8952 peptides) that I have:

PASS 2023, v23.0.2 2024-01-25 19:09:05 1

Assurance for Two-Sample T-Tests Allowing Unequal Variance

Numeric Results

Solve For: Assurance
Hypotheses: H0: $\delta \leq 0$ vs. H1: $\delta > 0$
Prior Type: Independent Univariate Distributions

Prior Distributions
δ: Lognormal (Mean = -2.85, SD = 1.35).
$\sigma 1$: Lognormal (Mean = 0.4, SD = 0.65), Left Boundary = -2.5, Right Boundary = 5.
$\sigma 2$: Lognormal (Mean = -0.2, SD = 0.65), Left Boundary = -2.5, Right Boundary = 5.

Assurance*	Power‡	N1	N2	N	Expected Mean Diff E(δ)	Expected Group 1 Std Dev E(σ1)	Expected Group 2 Std Dev E(σ2)	Alpha
0.44172	0.70233	1035	335	1370	0.1648	1.69127	0.99787	0.05

* The number of points used for computation of the prior(s) was 50.
‡ Power was calculated using $\delta = E(\delta) = 0.1648$, $\sigma 1 = E(\sigma 1) = 1.69127$, and $\sigma 2 = E(\sigma 2) = 0.99787$.

With the (median) sample sizes we actually used

What about a fully Bayesian approach?

In fact, PASS has a set of modules ("Assurance") that will do this! (In fewer dimensions…)
As a test of principle, I applied it to some real proteomics data (8952 peptides) that I have:

PASS 2023, v23.0.2 2024-01-25 19:18:56 1

Assurance for Two-Sample T-Tests Allowing Unequal Variance

Numeric Results

Solve For: Assurance
Hypotheses: H0: $\delta \leq 0$ vs. H1: $\delta > 0$
Prior Type: Independent Univariate Distributions

Doubling the sample sizes

Prior Distributions
δ: Lognormal (Mean = -2.85, SD = 1.35).
$\sigma 1$: Lognormal (Mean = 0.4, SD = 0.65), Left Boundary = -2.5, Right Boundary = 5.
$\sigma 2$: Lognormal (Mean = -0.2, SD = 0.65), Left Boundary = -2.5, Right Boundary = 5.

Assurance*	Power‡	N1	N2	N	Expected Mean Diff E(δ)	Expected Group 1 Std Dev E(σ1)	Expected Group 2 Std Dev E(σ2)	Alpha
0.53019	0.92394	2070	670	2740	0.1648	1.69127	0.99787	0.05

* The number of points used for computation of the prior(s) was 50.
‡ Power was calculated using $\delta = E(\delta) = 0.1648$, $\sigma 1 = E(\sigma 1) = 1.69127$, and $\sigma 2 = E(\sigma 2) = 0.99787$.

A few interesting approaches that we didn't cover

- **Harmonic-mean p-value averaging**
  https://www.pnas.org/doi/10.1073/pnas.1814092116
- **Sequential approaches (run a few at a time, then run more, etc)**
  Stuart & Ord Ch24
- **Bayesian LASSO**
  Hastie et al Ch3
- **Bayesian expected utility**
  Carlin & Louis p117f

Some final thoughts and tips…

- **Collect pilot data** if at all possible! You need some understanding of the measurement characteristics (overall variance, average reliability, distributional features like skew, etc)
  - People who perform omics measurements professionally will often have past datasets lying around that they can share with you to help you get a sense of this
- It rarely makes sense to run omics with sample sizes smaller than ***n* ≈ hundreds**. Reliability tends to be very poor, normalization is usually critical, and a lot of missing data can quickly ruin the integrity of smaller studies, even if the missingness is random
  - Even in pilot settings, n ≈ 50-100 should be considered a bare minimum, and you shouldn't expect to learn much about any scientific hypothesis
- Be open and honest about your **uncertainty**! These kinds of studies are often exploratory, and should be approached with that mindset. You're trying to justify the *informativeness* of the data you'll collect, not to assert that it will lead to any firm conclusions
  - "Informative" means you can learn *something* from it. That something doesn't have to be the answer to your scientific question. Usually it will just be a nudge or a jolt
- In the strictest sense, "power" is not well-defined in high-dimensional settings. The best you can do is **provide a range of power possibilities** over some domain of uncertainty

FURTHER READING…

- ***Measurement: Theory and Practice*** (2004) by Hand
  Careful but readable theoretic treatment of the philosophy of measurement and quantitative validity — HIGHLY RECOMMENDED
- ***Alternative Methods of Regression*** (1993) by Birkes, Dodge
  Robust approaches to regression analysis that carry less risk in noisy or high-dimensional settings
- ***Introduction to Modern Nonparametric Statistics*** (2004) by Higgins
  Useful practical survey of rank-based methods in various (including multivariate; Chapter 6) settings and bootstrapping (Chapter 8)

- ***Statistical Power Analysis for the Behavioral Sciences***, 2e (1988) by Cohen
  Classical discussion of power and sample-size estimation with much good practical advice
- ***Kendall's Advanced Theory of Statistics, Volume 2: Classical Inference and Relationship***, 5e (1991) by Stuart, Ord
  Chapters 21-25 give detailed treatment of power for inference, including discussion of sequential and nonparametric procedures

- **PASS (Power Analysis & Sample Size) Software** (2024)
  Implements a broad plethora of sample-size estimation procedures (https://www.ncss.com/software/pass/)
- **Harmonization of quality metrics and power calculation in multi-omic studies**, *Nature Communications* 11:3092 (2020)
  Describes their MultiPower software and considerations for omics power analysis (https://doi.org/10.1038/s41467-020-16937-8)

- ***The Elements of Statistical Learning: Data Mining, Inference, and Prediction***, 2e (2009) by Hastie, Tibshirani, Friedman
  Excellent theoretical survey of statistical methods applicable in high-dimensional settings
- ***Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*** (2010) by Efron
  Theoretical treatment of false discovery rate estimation and its application to inference in high-dimensional settings
- ***Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*** (2016) by Efron, Hastie
  Applications-oriented synthesis of the above two books

- ***Bayesian Data Analysis***, 3e (2013) by Gelman, Carlin, Stern, Dunson, Vehtari, Rubin
  The go-to textbook for modern applied Bayesian methods
- ***Bayes and Empirical Bayes Methods for Data Analysis***, 2e (2000) by Carlin, Louis
  Practical treatment of Bayesian procedures from the perspective of relating frequentist performance to decision-theoretic risk