Design-based conformal prediction for survey sampling

Jerzy Wieczorek
Colby College
jerzy.wieczorek@colby.edu

July 9, 2024
CANSSI-CRT Workshop on
Modern Methods in Survey Sampling
Ottawa, Ontario

Today's talk

- What "conformal prediction" is and how it works
- How it can work for design-based survey data analysis
- Open research questions for survey methodologists

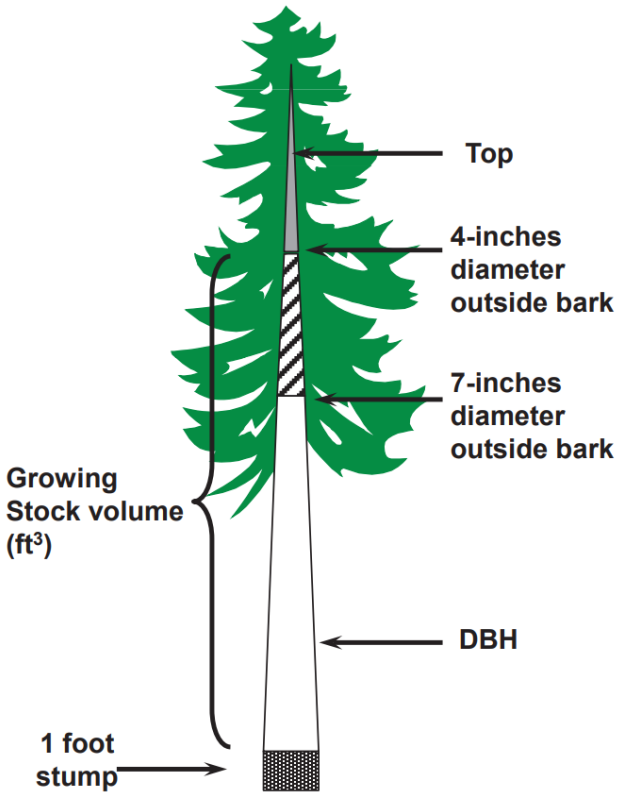
For details, see:
Wieczorek (2023), "Design-based conformal prediction,"
Survey Methodology, 49 (2)

Conformal prediction

Conformal prediction is a way to find prediction intervals, for nearly **arbitrary predictive models** and **with guaranteed finite-sample coverage**, while making **minimal assumptions** about the data distribution.

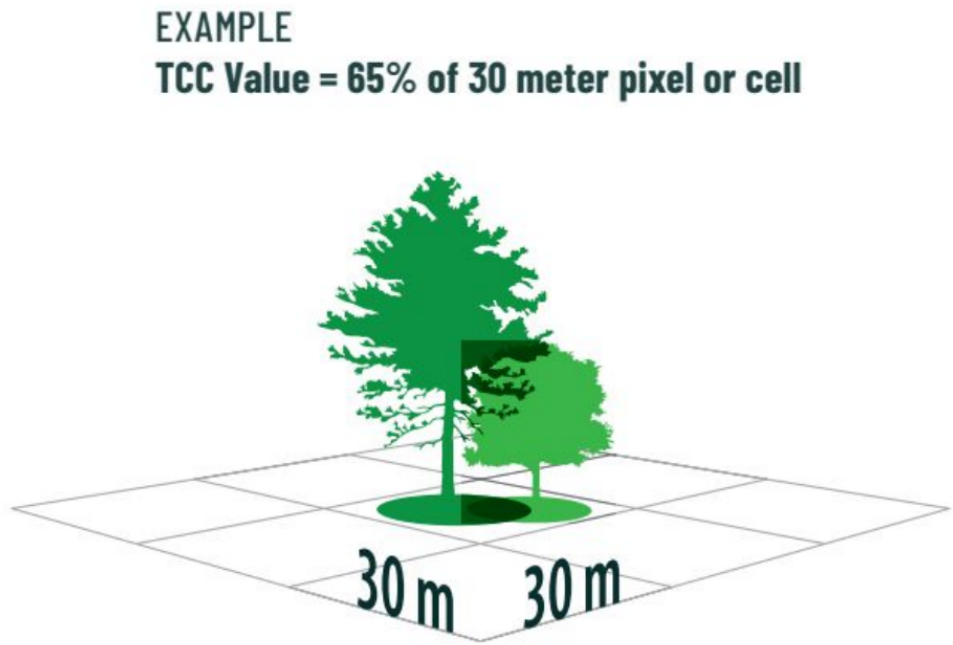
Motivating example: Forest Inventory & Analysis, USFS

Survey: sites **sampled** by US Forest Service
Example: timber volume (board ft) in a forest stand



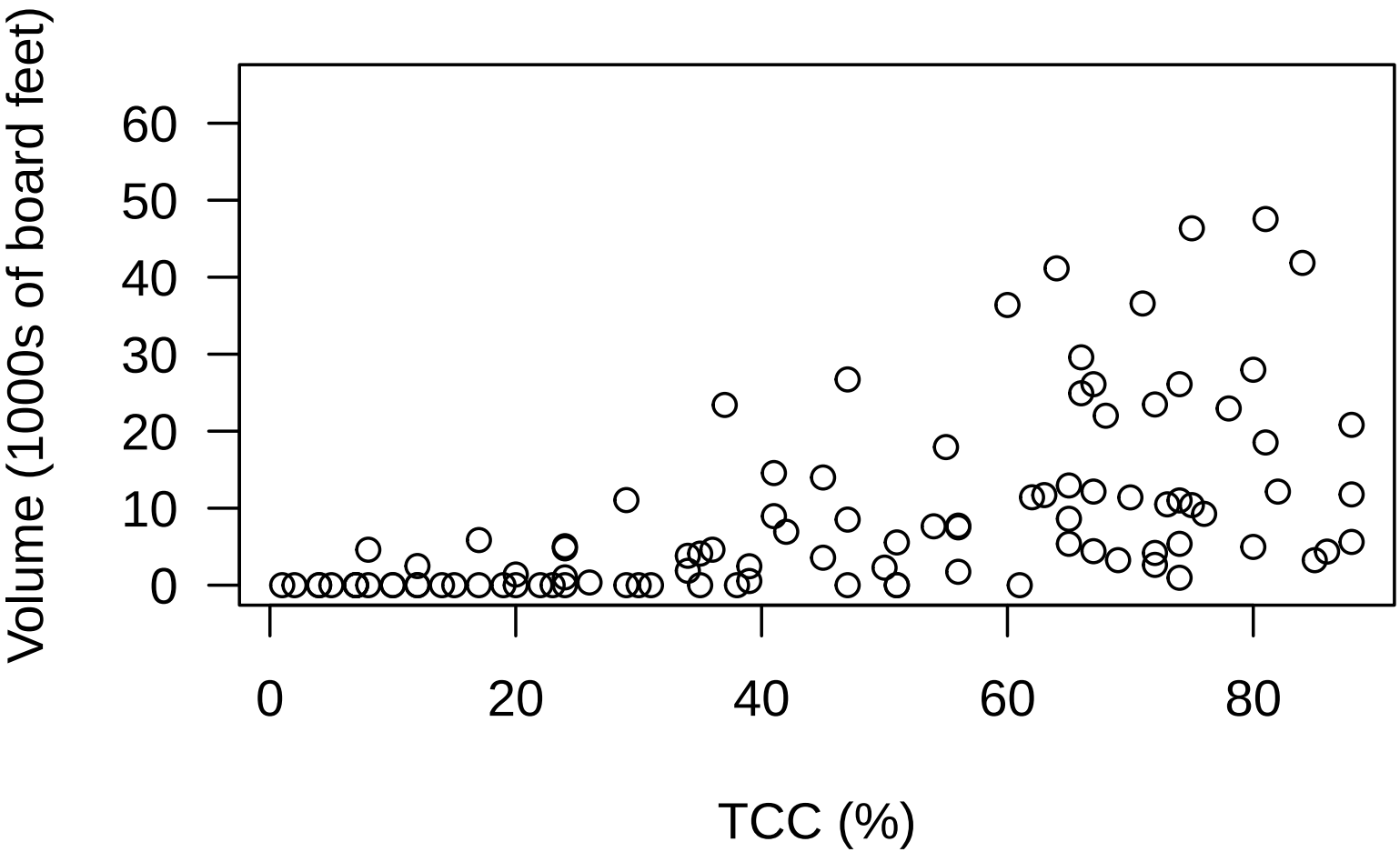
Berg et al. (2015)

Auxiliary data: Satellite, precipitation, elevation, etc. Known for **entire** US.
Example: TCC = Tree Canopy Cover



USFS (2020)

Simple data example: predict Volume using TCC



USFS goals include **predictive modeling**

- Using sampled sites, build models to predict survey variables (eg timber volume) from auxiliary data (eg TCC).
- Estimate unit-level predictions for out-of-sample pixels.
- **Today's talk:**
  Find **prediction intervals (PIs)** for these pixels.

Prediction intervals vs Confidence intervals

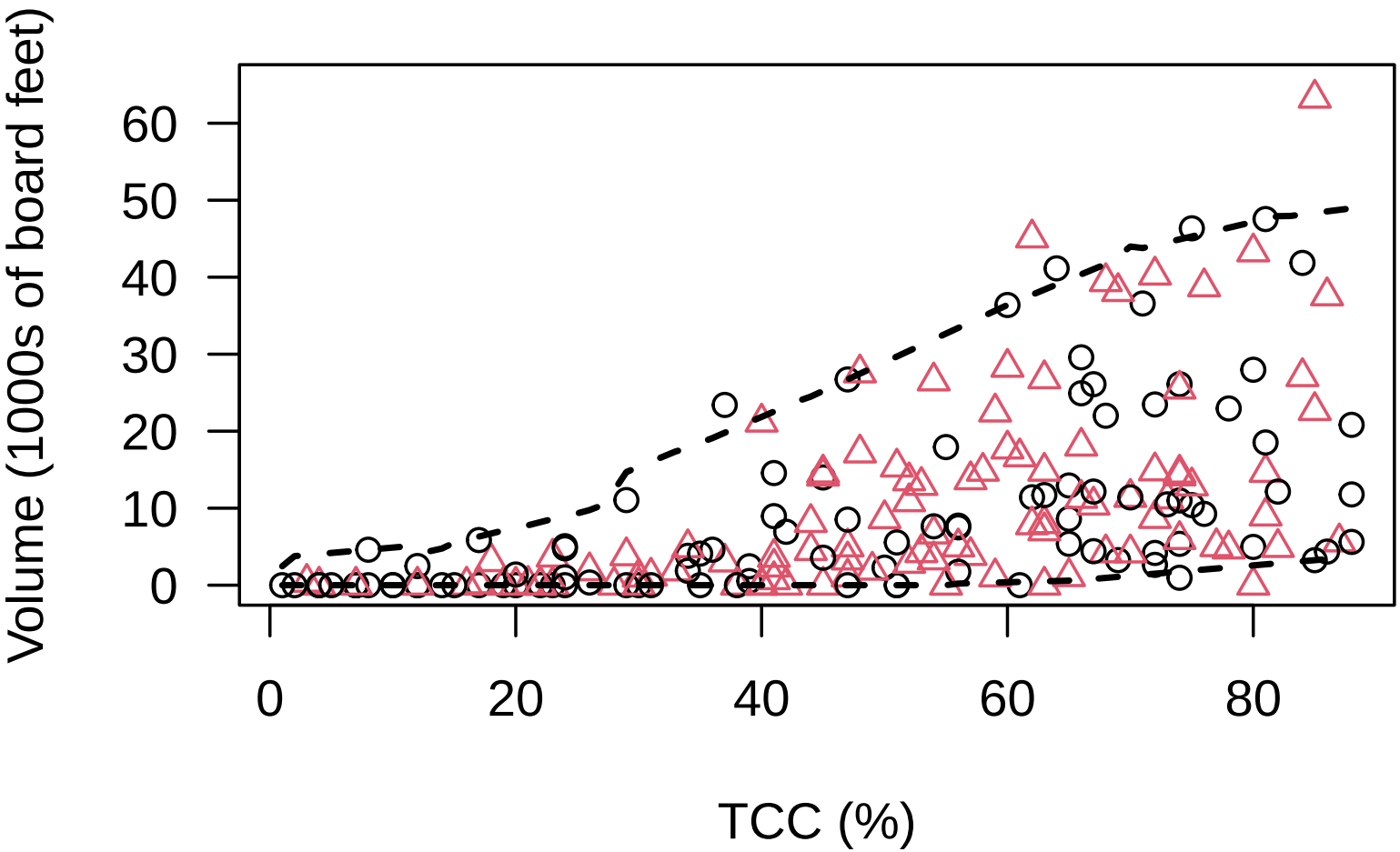
- 90% **confidence interval** for $\mathbb{E}(Y|X = x)$:
  Among all pixels with these same auxiliary data values, we're 90% confident that their **mean** timber volume is between 15,000 and 19,000 board feet.

Prediction intervals vs Confidence intervals

- 90% **confidence interval** for $\mathbb{E}(Y|X = x)$:
  Among all pixels with these same auxiliary data values,
  we're 90% confident that their **mean** timber volume is
  between 15,000 and 19,000 board feet.

- 90% **prediction interval** for Y at $X = x$:
  For a pixel with these auxiliary data values,
  we're 90% confident that its **individual** timber volume is
  between 4,000 and 30,000 board feet.

Example of 90% PI band for Volume using TCC



How to find prediction intervals (PIs)?

For linear regression, there are known equations for PIs if certain conditions are met (errors are Normally distributed, etc.).

But not guaranteed to work when conditions aren't met;
and not developed for many other statistical models or machine learning algorithms.

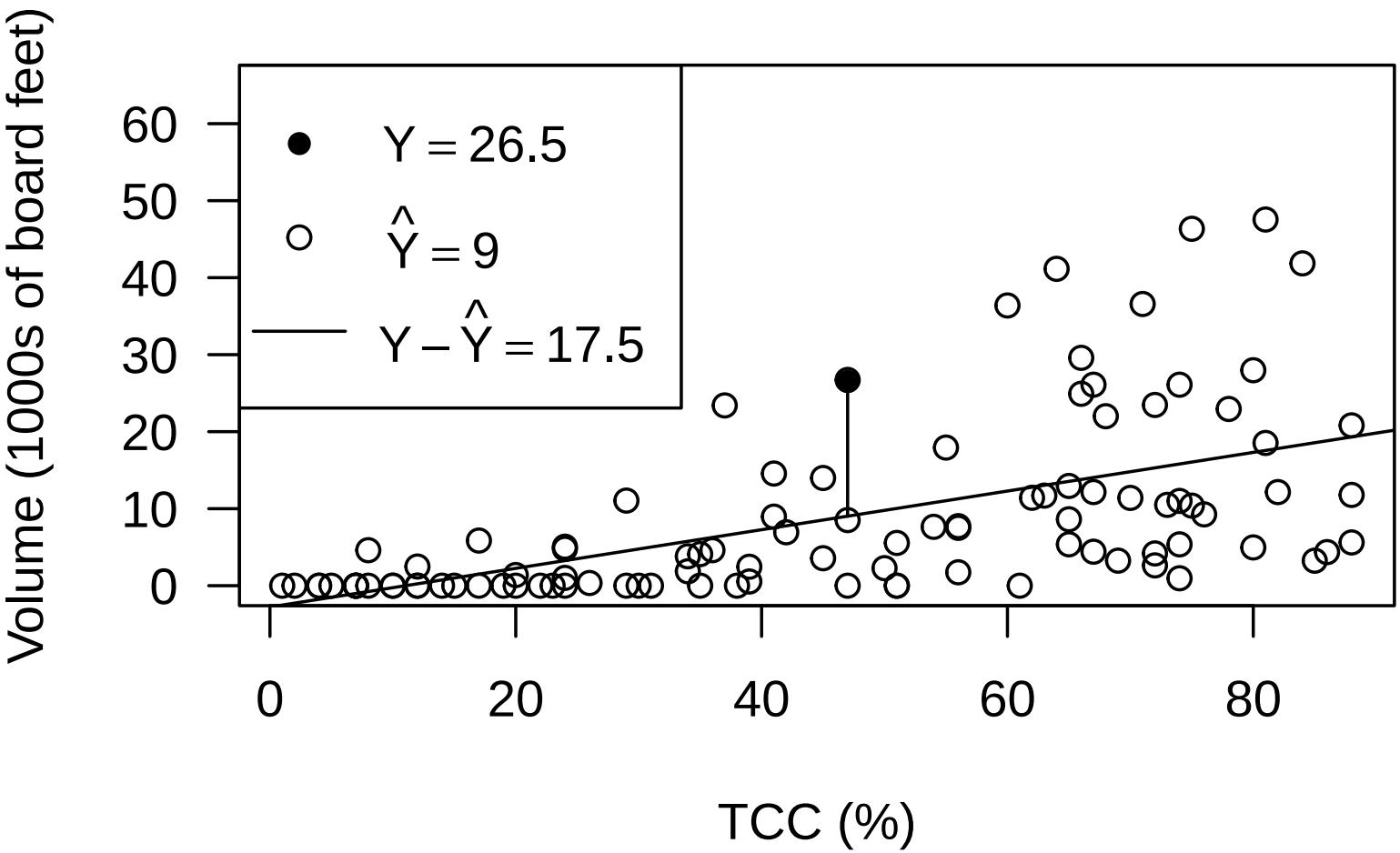
Conformal prediction

Conformal prediction is a way to find prediction intervals, for nearly **arbitrary predictive models** and **with guaranteed finite-sample coverage**, while making **minimal assumptions** about the data distribution.

Conformal prediction: intuition

Let data be iid or SRS.

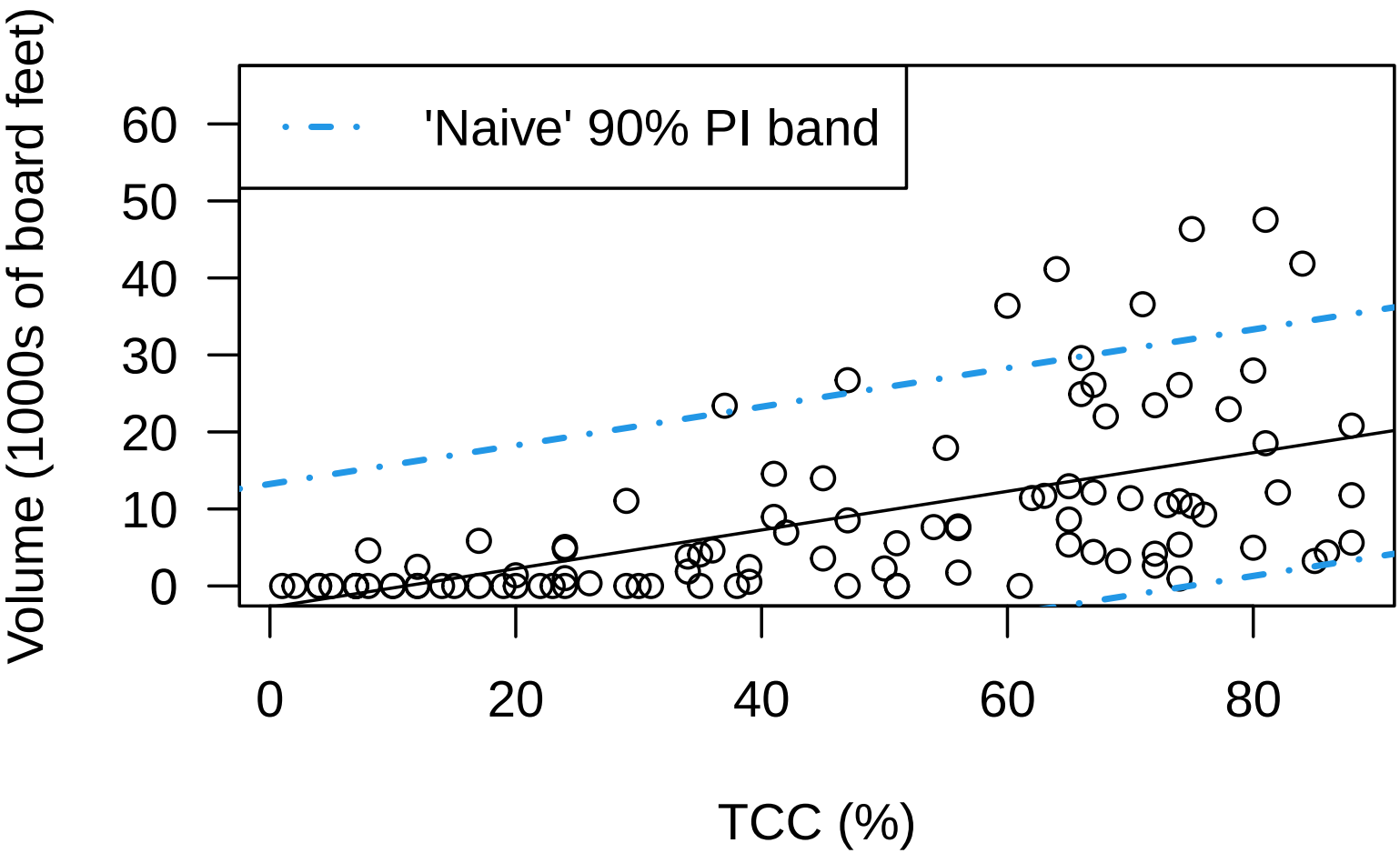
Consider the absolute residuals $|r_i| = |Y_i - \hat{Y}_i|$, for example:



Conformal prediction: intuition

Let $\hat{q}_{n,0.90} = \lceil 0.9n \rceil$ smallest value in $\{|r_1|, \ldots, |r_n|\}$.

Prob$(|r_{n+1}| \leq \hat{q}_{n,0.90}) \approx 0.90$, but this is **only approximate**. . .



Conformal prediction: intuition

"Conformal quantile lemma":

- First, *imagine we already have* the next observation!
  Let $\hat{q}_{n+1,0.90} = \lceil 0.9(n+1) \rceil$ smallest value in $\{|r_1|, \ldots, |r_{n+1}|\}$. Since they are in random order, $\text{Prob}(|r_{n+1}| \leq \hat{q}_{n+1,0.90}) = 0.90$ **exactly**—no approximation.

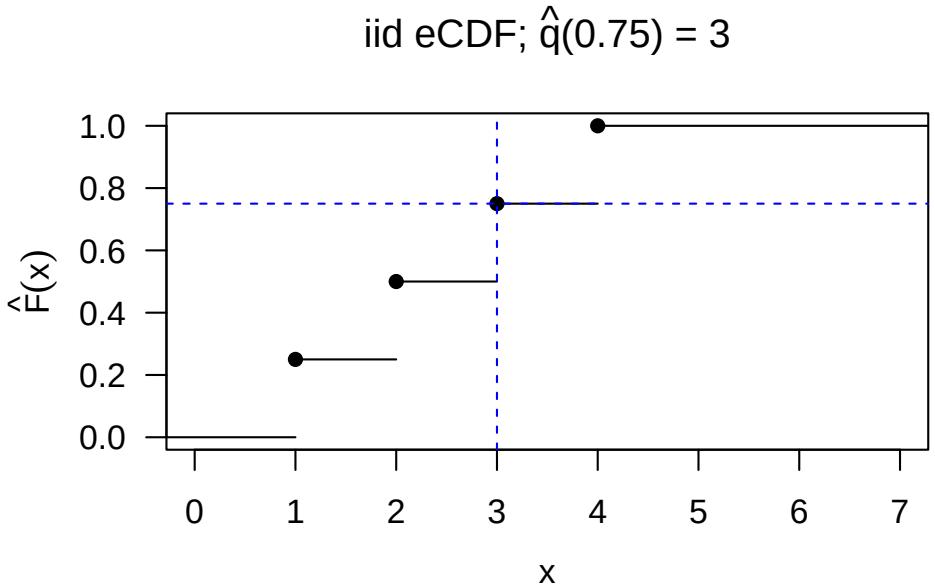
Conformal prediction: intuition

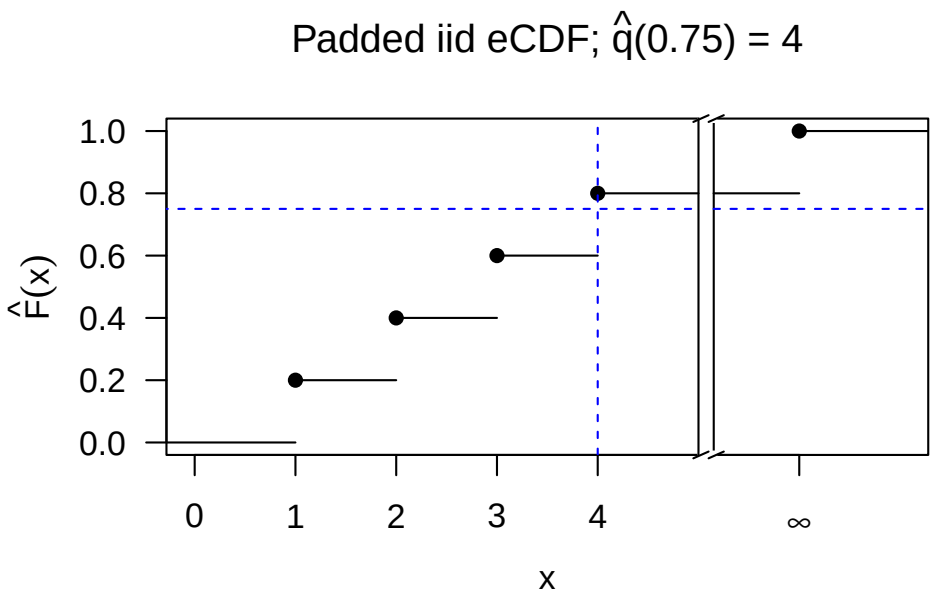
"Conformal quantile lemma":

- First, *imagine we already have* the next observation!
  Let $\hat{q}_{n+1,0.90} = \lceil 0.9(n+1) \rceil$ smallest value in $\{|r_1|, \ldots, |r_{n+1}|\}$. Since they are in random order, $\text{Prob}(|r_{n+1}| \leq \hat{q}_{n+1,0.90}) = 0.90$ **exactly**—no approximation.

- Second, this still works if we replace the unknown $|r_{n+1}|$ with any larger number, eg ∞.
  Let $\hat{q}_{conf,0.90} = \lceil 0.9(n+1) \rceil$ smallest value in $\{|r_1|, \ldots, |r_n|, \infty\}$. Since $\hat{q}_{conf,0.90} \geq \hat{q}_{n+1,0.90}$, we have $\text{Prob}(|r_{n+1}| \leq \hat{q}_{conf,0.90}) \geq 0.90$ **guaranteed**.

Conformal prediction: intuition





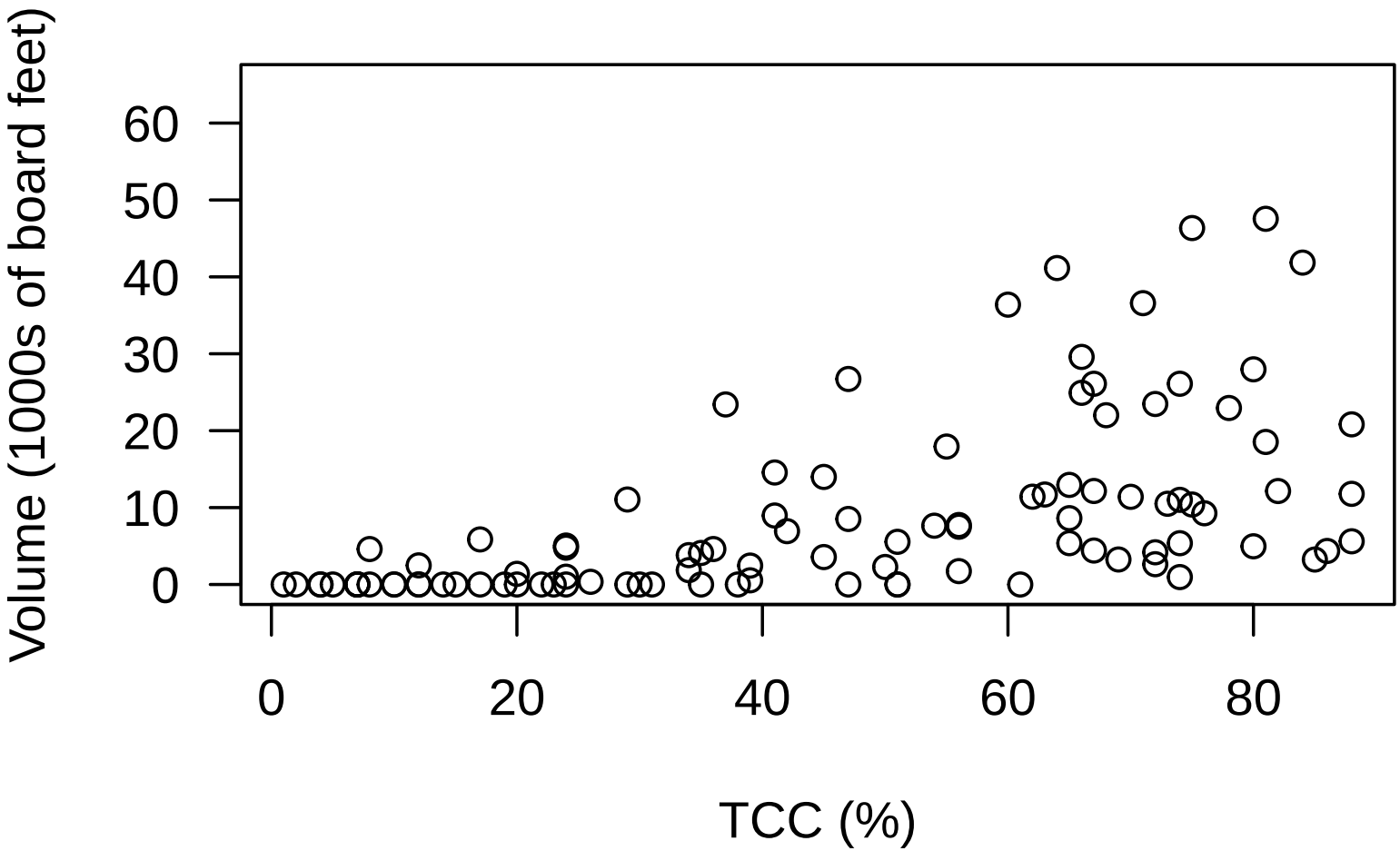
Conformal prediction: intuition

"Split conformal prediction" for SRS of size $m + n$:

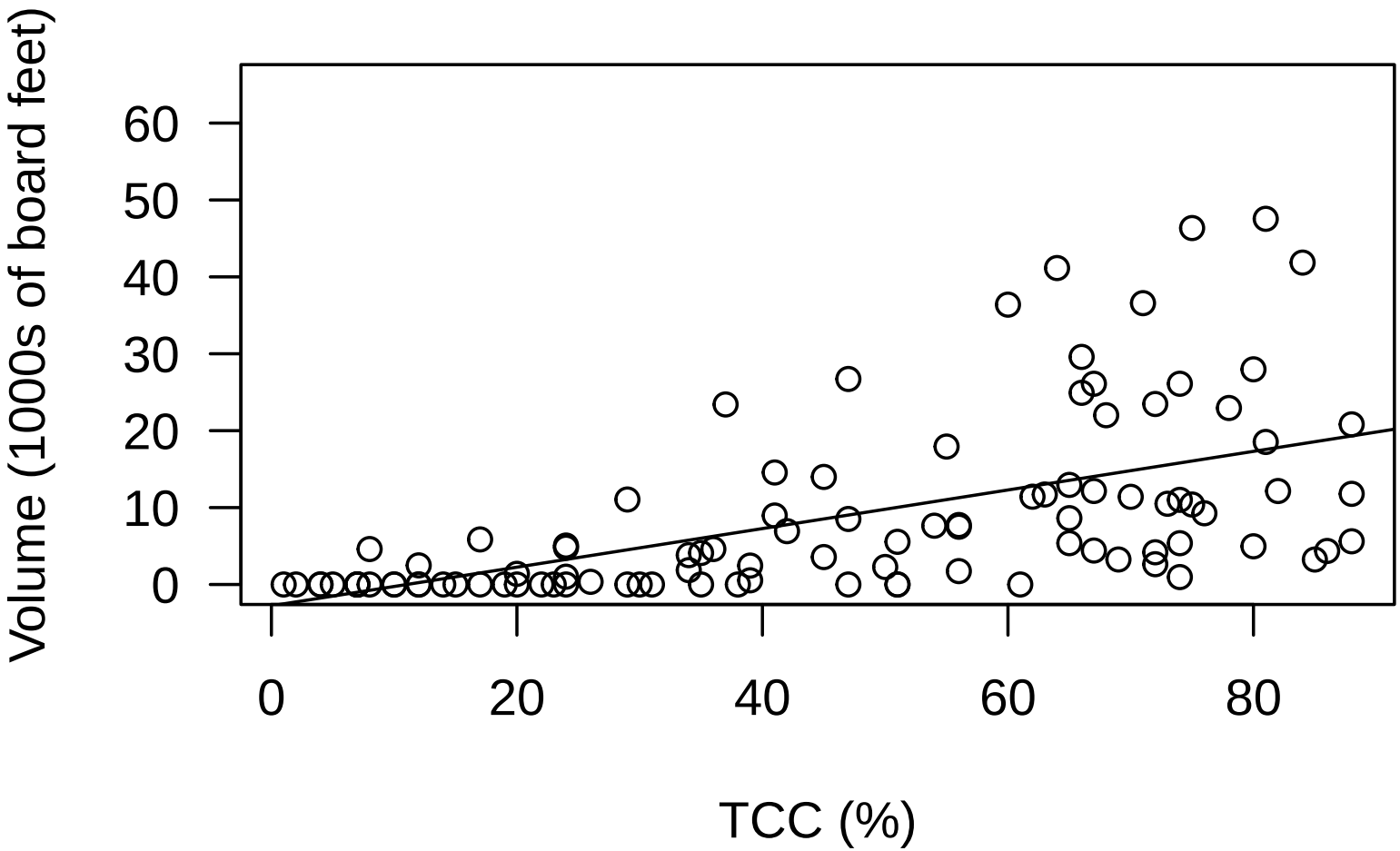
- Fit a regression to "training sample" of size m
- Using the rest of the data
  ("calibration sample" of size n),
  find the absolute residuals $|r_i| = |Y_i - \hat{Y}_i|$
- Find $\hat{q}_{conf,.90}$ on the r_i values

Then [regression line $\pm$ $\hat{q}_{conf,.90}$] is a conformal 90% PI band,
"90% guaranteed" to cover next observation Y_{n+1}

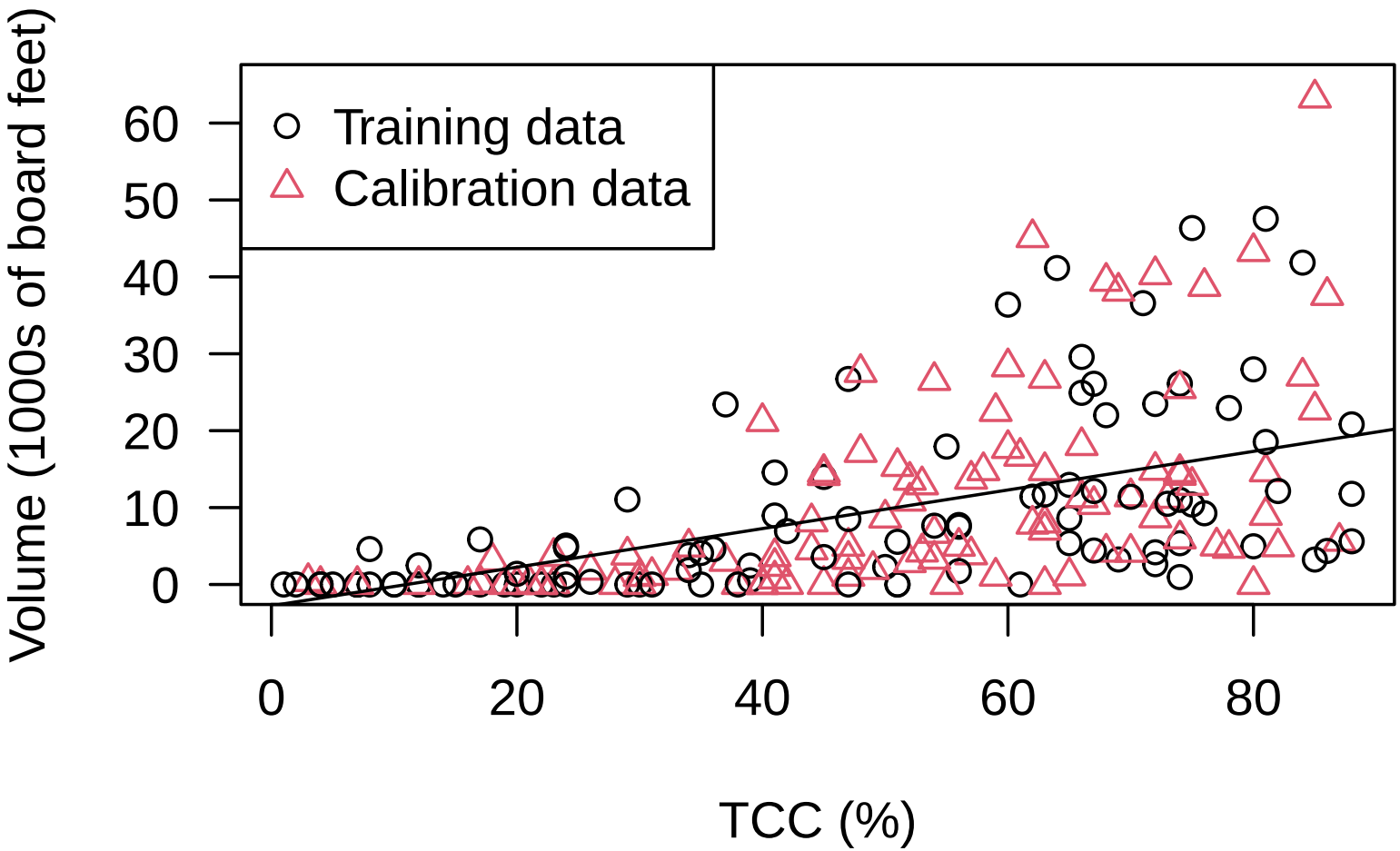
Conformal prediction for Volume using TCC



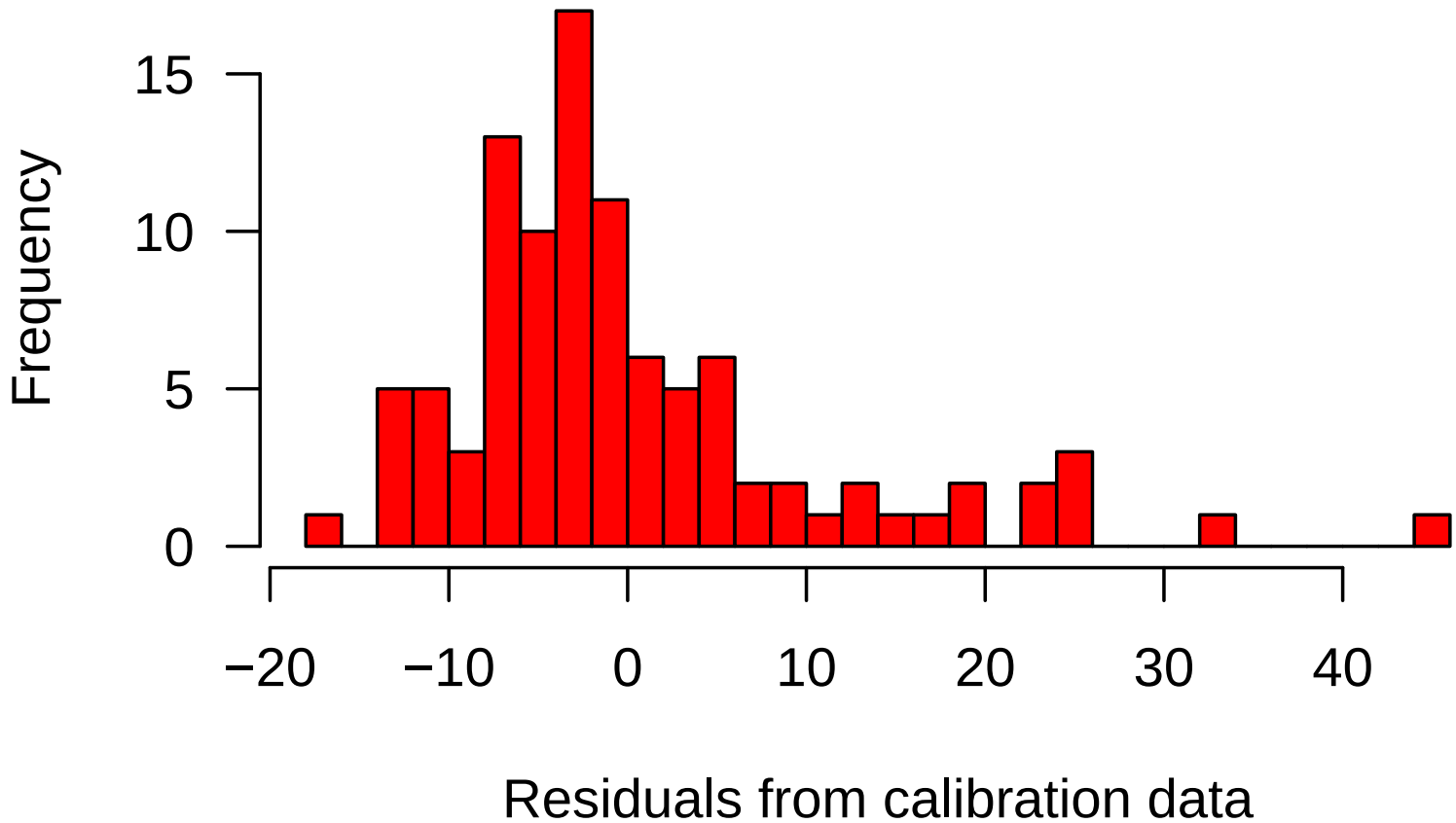
Conformal prediction for Volume using TCC



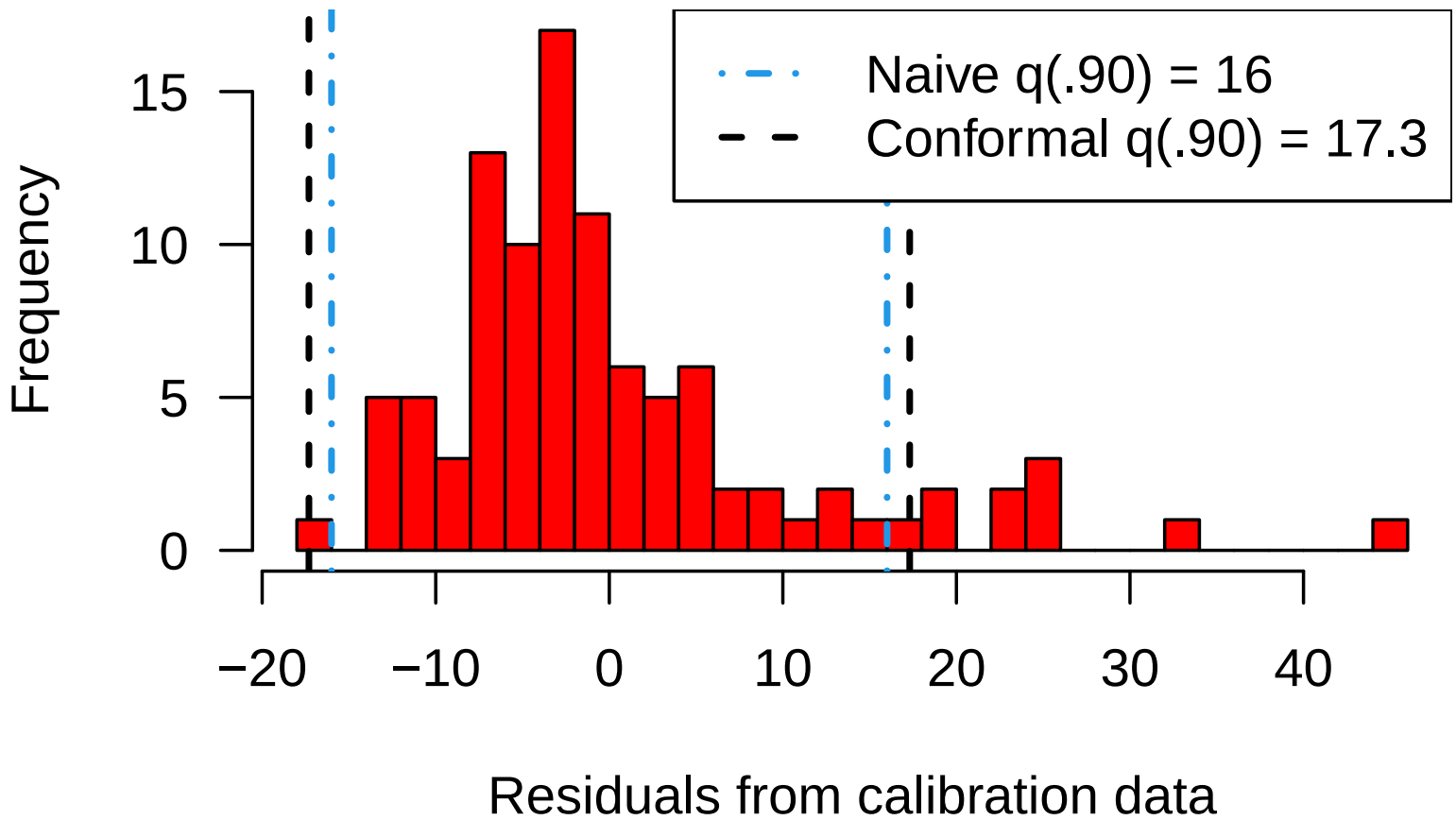
Conformal prediction for Volume using TCC



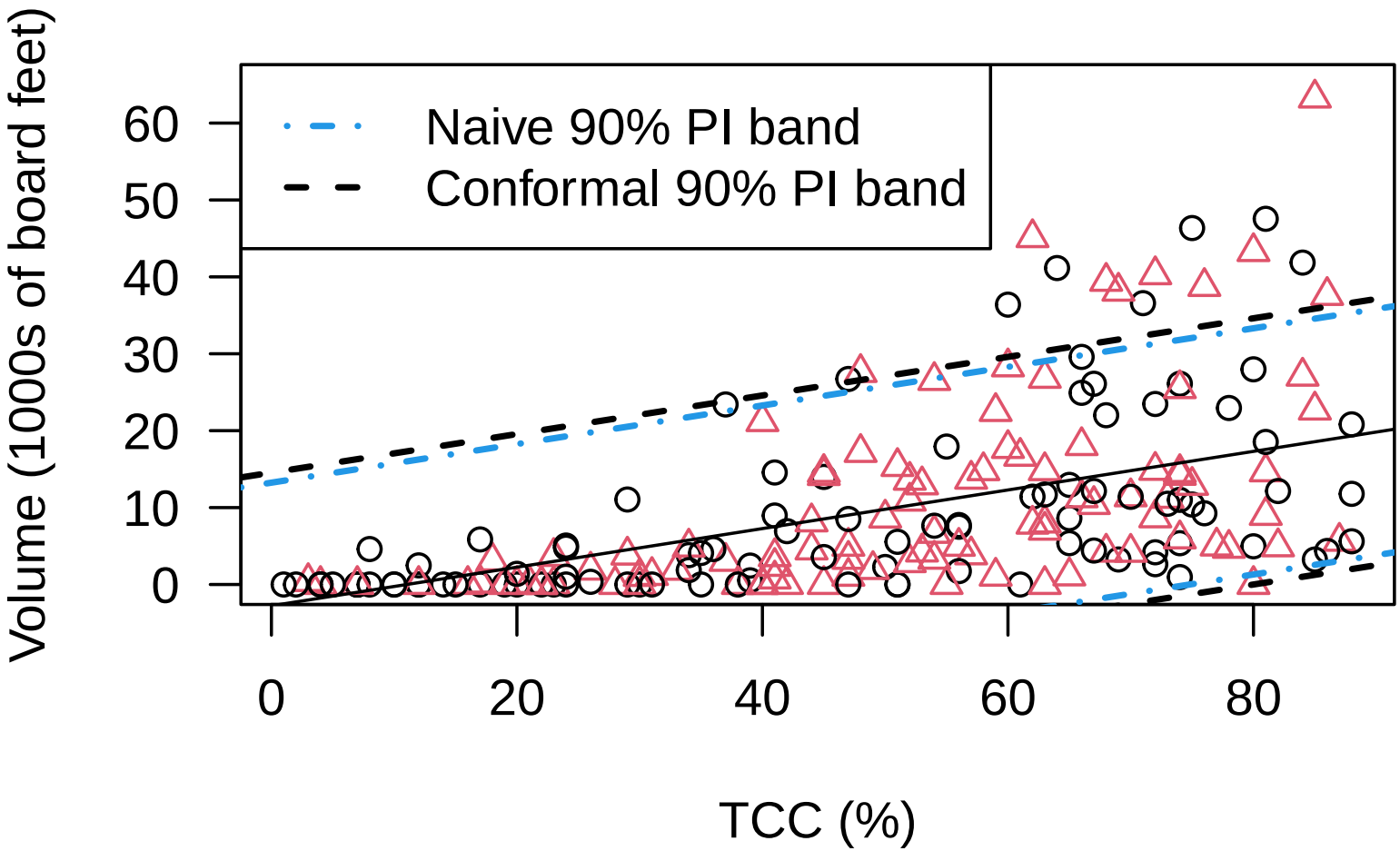
Conformal prediction for Volume using TCC



Conformal prediction for Volume using TCC

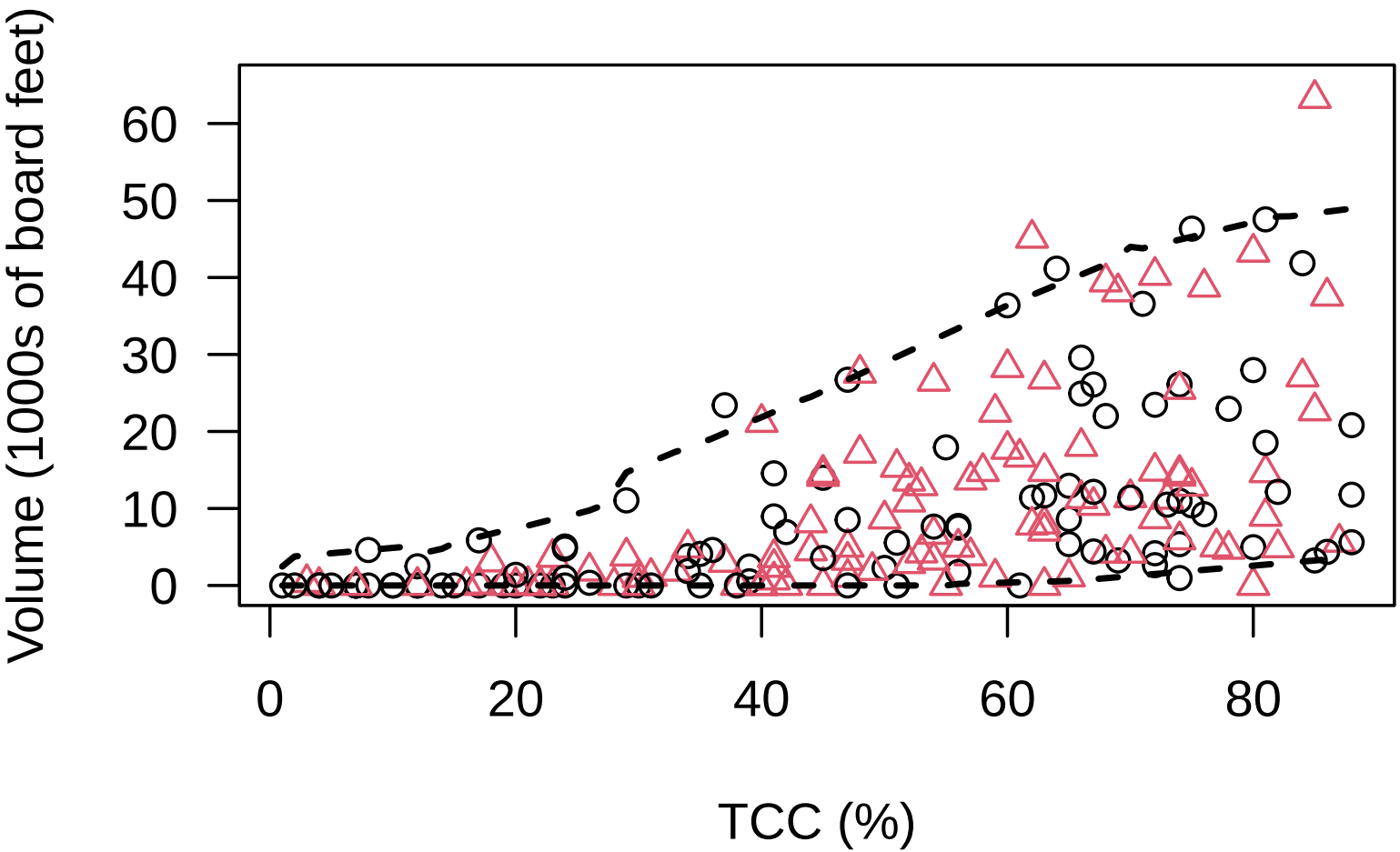


Conformal prediction for Volume using TCC



Conformalized Quantile Regression for Volume using TCC

Rather than constant-width PIs, we may prefer "adaptive" methods like Conformalized Quantile Regression (Romano et al. 2019): wider PIs at Xs whose Var(Y) is larger



Conformal prediction: what does it guarantee?

- **"Marginal 90% coverage"**:
  Conditional on the training set...
  Across all SRS of size $n+1$, where first n are the calibration set and last is the test observation...
  Y_{n+1} will be in the conformal PI built using the first n observations **for 90% of such samples**.

Conformal prediction: what does it guarantee?

- **"Marginal 90% coverage"**:
  Conditional on the training set...
  Across all SRS of size $n+1$, where first n are the calibration set and last is the test observation...
  Y_{n+1} will be in the conformal PI built using the first n observations **for 90% of such samples**.

- This may not be what we really want!
  Often we want the PI to cover next Y at a particular X, not at a randomly chosen X.

Conformal prediction: what does it guarantee?

- **"Marginal 90% coverage"**:
  Conditional on the training set...
  Across all SRS of size $n+1$, where first n are the calibration set and last is the test observation...
  Y_{n+1} will be in the conformal PI built using the first n observations **for 90% of such samples**.

- This may not be what we really want!
  Often we want the PI to cover next Y at a particular X, not at a randomly chosen X.

- But it's better than nothing. It works "out of the box" for pretty much any predictive algorithm, including ones where statisticians haven't worked out "native" PIs yet.

Design-based conformal prediction

Design-based conformal prediction

Conformal methods and design-based inference both...

- provide exact, finite-sample guarantees,
  not approximate or asymptotic
- rely only on sampling design,
  not on distribution of data in the population
- "work" even if the assisting models
  do not fit the population well

Design-based conformal prediction

Earlier conformal methods required **exchangeable sampling** (iid, SRS, etc)

. . . until Tibshirani et al. (2019),
"Conformal prediction under covariate shift":
they define a condition of "weighted exchangeability" which applies to many types of non-iid / non-SRS data

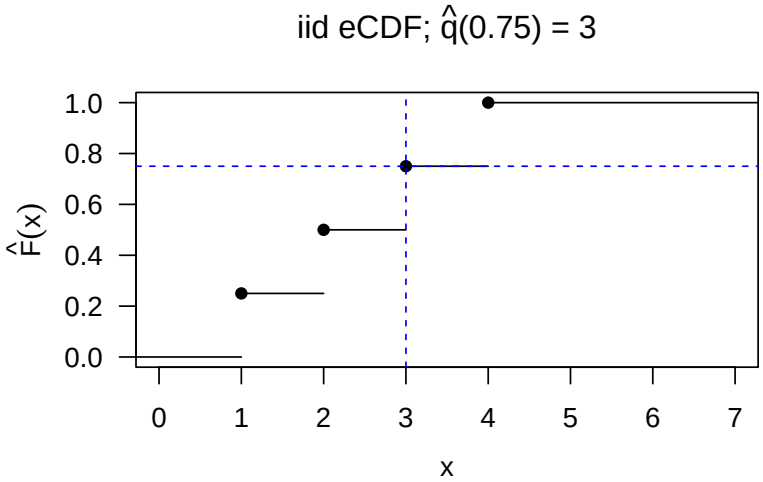
Spurred new conformal methods for dependent data: time series, designed experiments, networks. . . and survey samples!

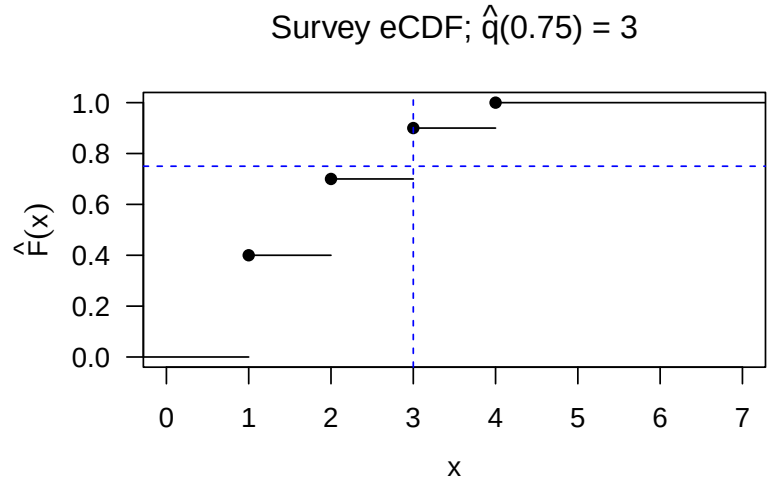
Design-based conformal prediction

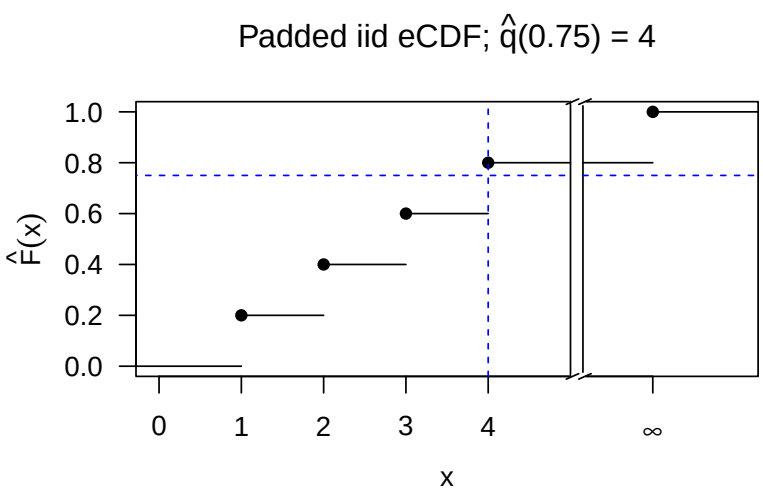
Wieczorek (2023), "Design-based conformal prediction," *Survey Methodology*, 49 (2)

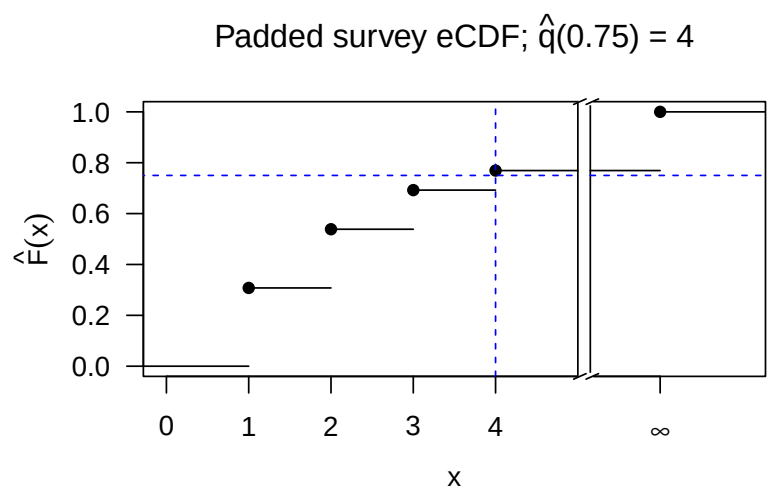
- Unequal prob. sampling (with replacement): use survey-weighted eCDF for the conformal quantiles

Design-based conformal prediction









Design-based conformal prediction

Methods available so far:

- Unequal-prob. sampling (with or without replacement)
  or post-stratification:
  use survey-weighted eCDF for the conformal quantiles
- Equal-prob. cluster samples:
  use hierarchical-data methods of Dunn et al. (2022)
- Stratified samples: separate conformal pred. in each stratum

Some open questions:

- Unequal-prob. cluster samples
- Combining data across strata
- Longitudinal / panel surveys
- Joint superpopulation + design-based framework

Design-based conformal prediction in R

Examples in a RMarkdown document at
https://github.com/ColbyStatSvyRsch/surveyConformal-CANSSI

Conclusion

- Conformal methods can provide useful PIs when building predictive models
- Many interesting open questions on conformal methods for survey sampling

Please reach out if you are interested in these methods or other statistical collaborations!

Contact: jerzy.wieczorek@colby.edu

Wieczorek (2023), "Design-based conformal prediction," *Survey Methodology*, 49 (2)

Supplemental slides

From covariate shift to survey weights

Tibshirani et al. (2019)'s "weighted exchangeability" is meant for **covariate shift**:
Training data from one distribution, test case from another.
Their "weights" are e.g. ratios of PDFs:
$w(x_i) = p_{\text{test}}(x_i)/p_{\text{train}}(x_i)$

For design-based inference, "training distr." is multinomial draws from finite pop, using the survey design's known sampling probs π_i, and "test distr." is SRS on entire finite pop: $w(x_i) \propto (1/N)/\pi_i$ or the usual inverse-prob sampling weight.
This "test distr." is how we formally guarantee marginal coverage across entire finite pop.

Practical concerns with conformal methods for surveys

Users need to know their test case's sampling probability under original sampling design. But we don't usually release sampling probs for every population unit: impractical, privacy risks, etc. Instead, survey orgs could:

- Release broad categories with similar sampling weights
- Release a GVF but for samp weight instead of variance
- Release a single "largest samp weight" to be used for all conformal predictions, conservatively

Otherwise, users could:

- Use the largest weight in the sample for all conformal predictions, conservatively
- Try several plausible weights and report a sensitivity analysis
- Find an in-sample case with similar X values to test case, and use their weight

Classification with conformal methods

- Fit a probabilistic classifier to training set, so that $\hat{f}(x)_y$ is estimated prob of being in class y when $X = x$.
- Instead of abs. residuals, find "scores" $1 - \hat{f}(X_i)_{Y_i}$ for each i in calibration set.
  Smaller score = more plausibly belongs to that class.
- Let $\hat{q}_{conf}$ be $\lceil (1-\alpha)(n+1) \rceil$ smallest score.
- For test case, find $1 - \hat{f}(X_{n+1})_y$ for each class y. Let prediction set be all the classes with scores below $\hat{q}_{conf}$.

This is "adaptive": For a difficult test-case to classify, X_{n+1} has moderately high prob for many classes y, and the set will be large. For an easy test case, only one class has a high prob, and the set will be small.

"Full conformal prediction"

Do not split data. We have n training cases, and one test case where only X_{n+1} is known, and we want a PI for Y_{n+1}.

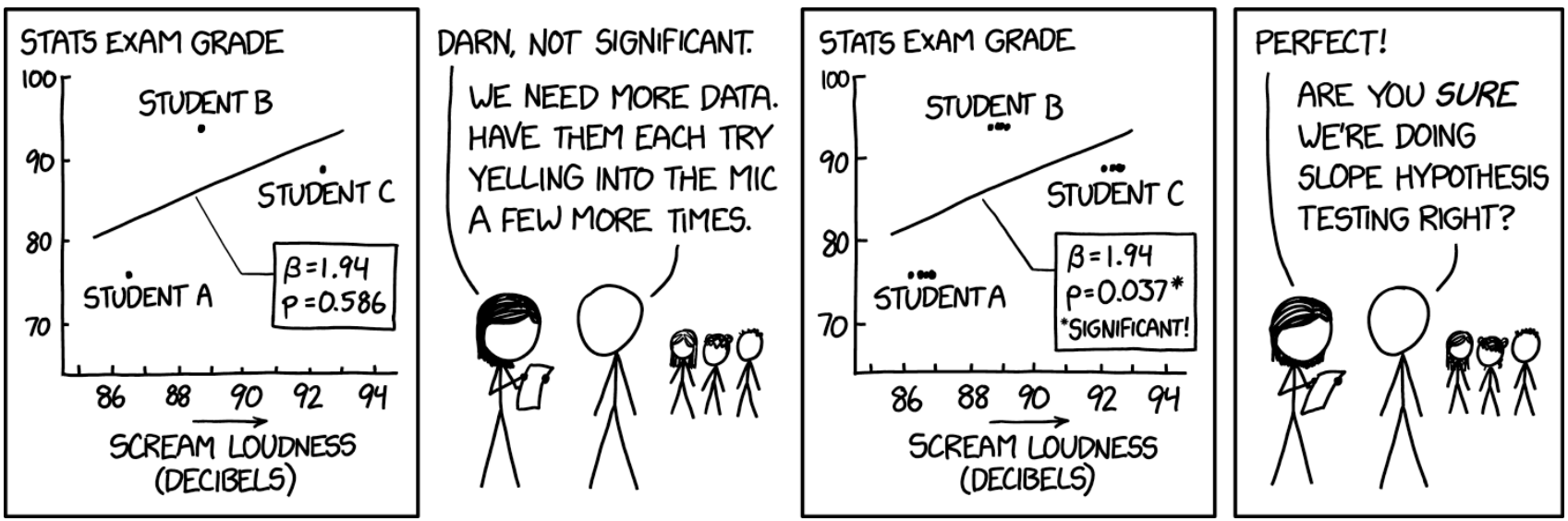
- Pick a hypothetical response value y
- Fit $\hat{f}_y$ to a dataset where we pretend y is correct: $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$
- Find the $n+1$ absolute residuals r_i
- Find their $1-\alpha$ quantile, $\hat{q}_{y,1-\alpha}$
- If $r_{n+1} \leq \hat{q}_{y,1-\alpha}$, we say y "conforms" to the data, and we include y in our PI

"Full conformal" vs. "split conformal"

- Full conformal avoids a random data split
- Full conformal is more stat. efficient:
  less variability in the PI endpoints
- Split conformal is much faster to compute:
  split = fit $\hat{f}$ and $\hat{q}_{conf}$ once;
  full= fit a new $\hat{f}_y$ for each y and a new PI for each X_{n+1}
- Empirically, split and full results are often similar
  (Lei et al., 2018)

Cross validation with complex sampling designs

Why not just always use leave-one-out CV?
Example from XKCD:



What is CV actually doing?

Instead of risk (expected loss $L(y, \hat{y})$) for the observed sample s,

$$Err_s(f) = \mathbb{E}_{(x_{new}, y_{new})} L(y_{new},\ \hat{f}_s(x_{new}))\,,$$

K-fold CV tries to estimate average risk over similar samples s^*

$$Err(f) = \mathbb{E}_{s^*}\left[\mathbb{E}_{(x_{new}, y_{new})} L(y_{new},\ \hat{f}_{s^*}(x_{new}))\right]$$

as empirical risk on K test sets after fitting f to K training sets:

$$\widehat{Err}_{CV}(f) = \frac{1}{K}\sum_{j=1}^{K}\left[\frac{1}{n_{test_j}}\sum_{i \in test_j} L(y_i,\ \hat{f}_{train_j}(x_i))\right].$$

The way CV selects train/test sets affects bias of $\widehat{Err}_{CV}(f)$.
For usual CV, bias is only from training set sizes: $n \times \frac{K-1}{K} < n$.

Why not use usual CV for complex survey designs?

- If s was iid sample of size n, usual CV's bias in $\widehat{Err}_{CV}(f)$ only comes from training set size $n \times \frac{K-1}{K} < n$. Often this bias is (a) small and (b) nearly constant across competitive models, so it should not affect model selection much.
- But for complex surveys, each $train_j$ should be formed in a way that reflects **actual sampling design** of s. Otherwise, the bias in $\widehat{Err}_{CV}(f)$ could be (a) large and (b) very different across competitive models, causing poor model selection.
- For complex surveys, when survey respondents don't all have the same sampling probability, bias can also come from taking a simple mean of the loss over test cases.

How **can** we do design-based CV?

Wieczorek, Guerin, and McMahon (2022),
"K-fold cross-validation for complex sample surveys," *Stat*

1. Create complex-survey CV folds in the same way that we form "Random Groups" for variance estimation & for group jackknife

- For single-stage SRS, divide the sample at random into K folds (as in usual CV).
- For cluster sampling, sample the clusters as units: all elements from a given cluster should be placed in the same fold.
- For stratified sampling, make each fold a stratified sample of units from each stratum.
- For multi-stage sampling, combine these rules as necessary.

2. Account for strata, clusters, survey weights, etc. in calculating expected loss, e.g. use survey-weighted mean for $\widehat{MSE}$.