

From Bootstrap to Asymptotic Inference

Standard errors, the CLT, and confidence intervals

Introduction

- We're going to show how you can approximate non-parametric bootstrap inference by using the Normal distribution
- Our running example will be **hedonic pricing**: predicting the value of housing based on its amenities (square feet, bathrooms, total rooms, area, age)

Roadmap

1. Review: the sampling distribution and the standard error
2. Bootstrapping a regression
3. Why the bell curve keeps showing up: the CLT
4. Confidence intervals and tests
5. When it breaks
6. Our inference routine

every interval today is: estimate $\pm$ c $\times$ SE

The sampling distribution and standard error

Review: The Sampling Distribution

- Imagine running the same study over and over, with a new sample each time
- Each study computes the same statistic $\hat{\theta}$ (a mean, a coefficient, a median) and gets a different number
- The distribution of those numbers is the **sampling distribution** of $\hat{\theta}$
- The **standard error** is the standard deviation of the sampling distribution:

$$SE(\hat{\theta}) = \sqrt{\mathbb{V}[\hat{\theta}]}$$

- It is not the standard deviation of the data. It is the standard deviation of the **estimate**

The **standard error** is the
standard deviation of the
sampling distribution

The **standard error is the
standard deviation of the
sampling distribution!**

**The standard error is the
standard deviation of the
sampling distribution!!**

The **standard error is the
standard deviation of the
sampling distribution!!!**

Let's work out an example with a very small population

- Population {1, 3, 5}
- What is μ ? What is σ^2 ?
- Draw $n = 2$ with replacement. Calculate the mean for 9 possible samples:

sample	(1,1)	(1,3)	(1,5)	(3,1)	(3,3)	(3,5)	(5,1)	(5,3)	(5,5)
$\bar{x}$	1	2	3	2	3	4	3	4	5

- What is $E[\bar{x}]$? What is $V[\bar{x}]$?
- What is the standard error of the mean?
- This is the whole sampling distribution. Usually, we don't get to see it. How do we calculate SE when we don't see the whole sampling distribution?

Two Ways to Get the Standard Error

- **1. Know the whole sampling distribution.** List every possible sample, compute $\bar{x}$ for each, and take the variance of the $\bar{x}$'s. For $\{1, 3, 5\}$: $V[\bar{x}]$ (sampling distribution variance) = $4/3$, $SE = \sqrt{V[\bar{x}]} = 1.15$. Only possible for tiny populations or simulations
 - With real data: the **bootstrap** is this done on our sample instead of the population
- **2. Use the formula.** The sampling distribution's variance comes from the population's: $V[\bar{x}] = \sigma^2/n$, so $SE = \sigma/\sqrt{n}$. No list needed, but it needs σ , the population SD
 - $se(\bar{x}) = \sigma/\sqrt{n}$ is **exact but useless as written**: σ is unknown. Every practical SE fills that gap the same way
 - With real data: we know neither the sampling distribution nor σ . So we **plug in s**, the SD of our one sample: estimated $SE = s/\sqrt{n}$
- For $\{1, 3, 5\}$: population $\sigma^2 = 8/3$, so $\sigma = 1.63$ and $SE = \sigma/\sqrt{2} = 1.15$, the same answer as route 1

real data: estimated $SE = s/\sqrt{n}$, or the bootstrap

Estimating the Standard Error

- Route 2 needs σ , which we don't know. We estimate it from our one sample with s
- From one sample, say (3, 5): $\bar{x} = 4$ and

$$s^2 = \frac{(3-4)^2 + (5-4)^2}{2-1} = 2, \quad \widehat{\text{SE}} = \sqrt{s^2/n} = 1$$

- Why $n - 1$ when calculating s^2 , our estimate of σ^2 that comes from the sample? The deviations are measured from $\bar{x}$, which was fit to the same data and sits **closer to them than μ does**. From $\mu = 3$ the squares would sum to 4, not 2:

1. Split each deviation:
$$x_i - \mu = (x_i - \bar{x}) + (\bar{x} - \mu)$$

2. Square and add up:
$$\sum (x_i - \mu)^2 = \sum (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

Cross term is 0 (deviations from $\bar{x}$ sum to 0). Spread around μ = spread around $\bar{x}$ + a penalty. For (3, 5): $4 = 2 + 2$

3. Rearrange step 2:
$$\sum (x_i - \bar{x})^2 = \sum (x_i - \mu)^2 - n(\bar{x} - \mu)^2$$

Average each piece:
$$\mathbb{E}\left[\sum (x_i - \bar{x})^2\right] = n\sigma^2 - n \cdot \mathbb{V}[\bar{x}] = n\sigma^2 - n \cdot \frac{\sigma^2}{n} = (n - 1)\sigma^2$$

- Two denominators, two jobs: **$n - 1$** estimates σ^2 ; the **n** in σ^2/n comes from averaging

Exercise

- Data: {2, 4, 5, 6, 8}
- Compute $\bar{x}$, the sum of squared deviations, s^2 , and the estimated SE of $\bar{x}$
- What do you get if you divide by n instead of $n - 1$?
- Which one answers: "how spread out are these five numbers?" and which answers "how spread out is the population?"

Bootstrapping

Review: Sampling Distribution vs. Bootstrap Distribution

- The sampling distribution resamples from the **population**
- The bootstrap resamples from **our sample**

	Sampling distribution	Bootstrap distribution
Samples come from	the population F (unknown)	our sample, treated as the population
How many exist	one	one per dataset
Centered at	the true θ	our estimate $\hat{\theta}$
Spread	the true SE	an estimate of the SE

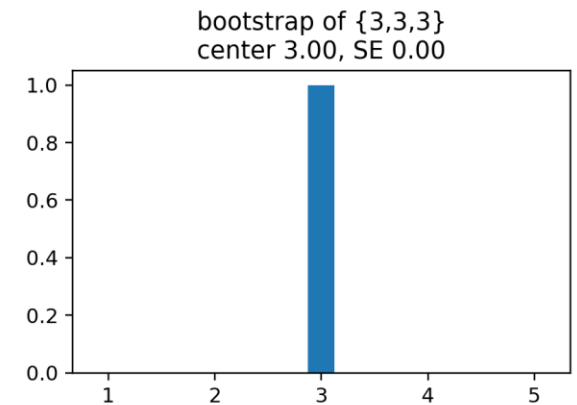
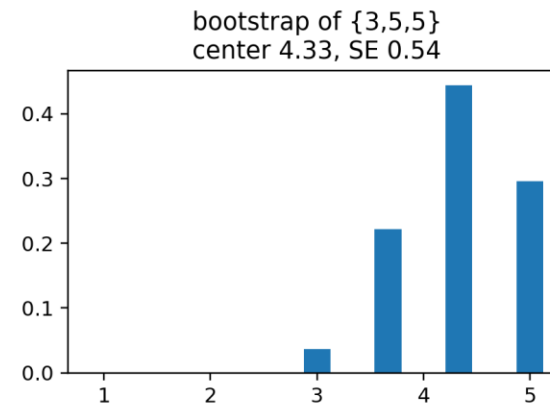
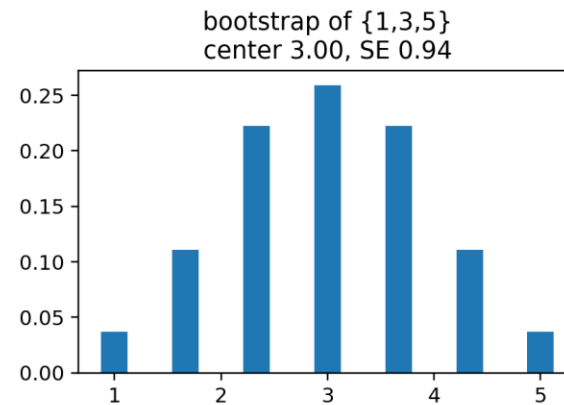
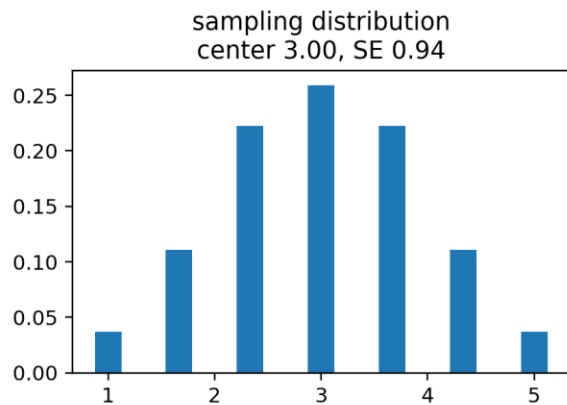
- The bet: $\hat{\theta}_R^* - \hat{\theta}$ behaves like $\hat{\theta} - \theta$

the bootstrap copies the spread and shape, never the location

Three Students, One Population

- Population $\{1, 3, 5\}$, samples of size 3. True sampling distribution: centered at 3, SE = 0.94
- Each student bootstraps their own sample:

sample	bootstrap center	bootstrap SE
$\{1, 3, 5\}$	3	0.94
$\{3, 5, 5\}$	4.33	0.54
$\{3, 3, 3\}$	3	0



- The point:** each bootstrap is centered at that student's own $\bar{x}$, not at μ (4.33 vs. 3), and its spread is only as good as their sample's ($\{3, 5, 5\}$ too narrow, $\{3, 3, 3\}$ zero)

Bootstrapping a regression

Hedonic Pricing: Ames, Iowa

- You have covered OLS/linear regression in ML1, so we can play with it here
- 2,930 house sales in Ames, Iowa
- Regress $\text{arcsinh}(\text{price})$ on $\text{arcsinh}(\text{total rooms})$, $\text{arcsinh}(\text{age})$, $\text{arcsinh}(\text{area})$, and full-bath dummies, with an intercept
- Bootstrap: resample **whole rows** $R = 1,000$ times and refit the regression each time
- Each refit gives one draw of every coefficient: an estimate of each coefficient's sampling distribution

Regress $\text{arcsinh}(\text{price})$ on total rooms, $\text{arcsinh}(\text{age})$, $\text{arcsinh}(\text{area})$, and full-bath dummies, with an intercept

```
outcome_var = ['price']
num_vars = ['TotRms.AbvGrd', ]
num_ihs_vars = [ 'age',
                  'area']
cat_vars = ['Full.Bath']

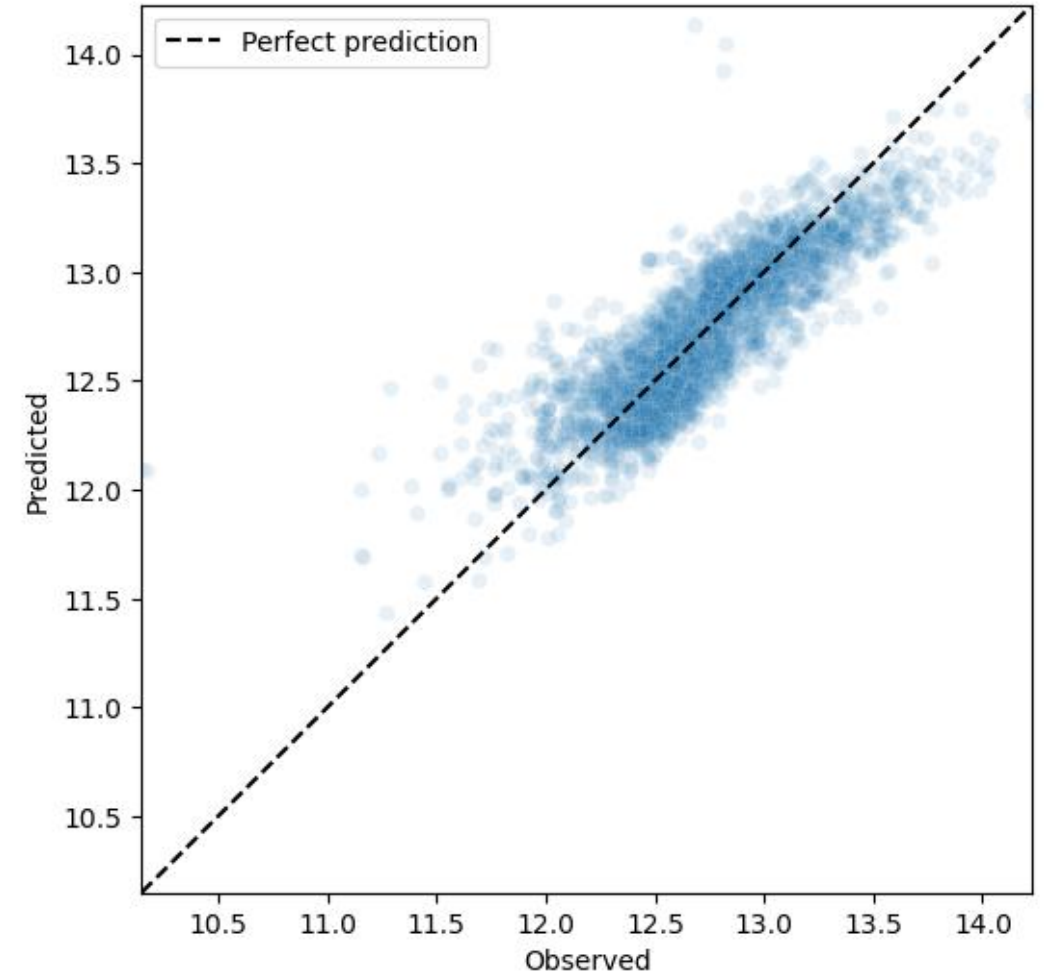
rdf = df.loc[:,outcome_var + num_vars + num_ihs_vars + cat_vars].dropna()

X = rdf.loc[:,num_vars]
for var in num_ihs_vars:
    X = pd.concat([ X, np.arcsinh(rdf[var])], axis = 1)
for var in cat_vars:
    X = pd.concat([ X, pd.get_dummies(rdf[var], dtype = int, drop_first = True, prefix=var)], axis = 1 )

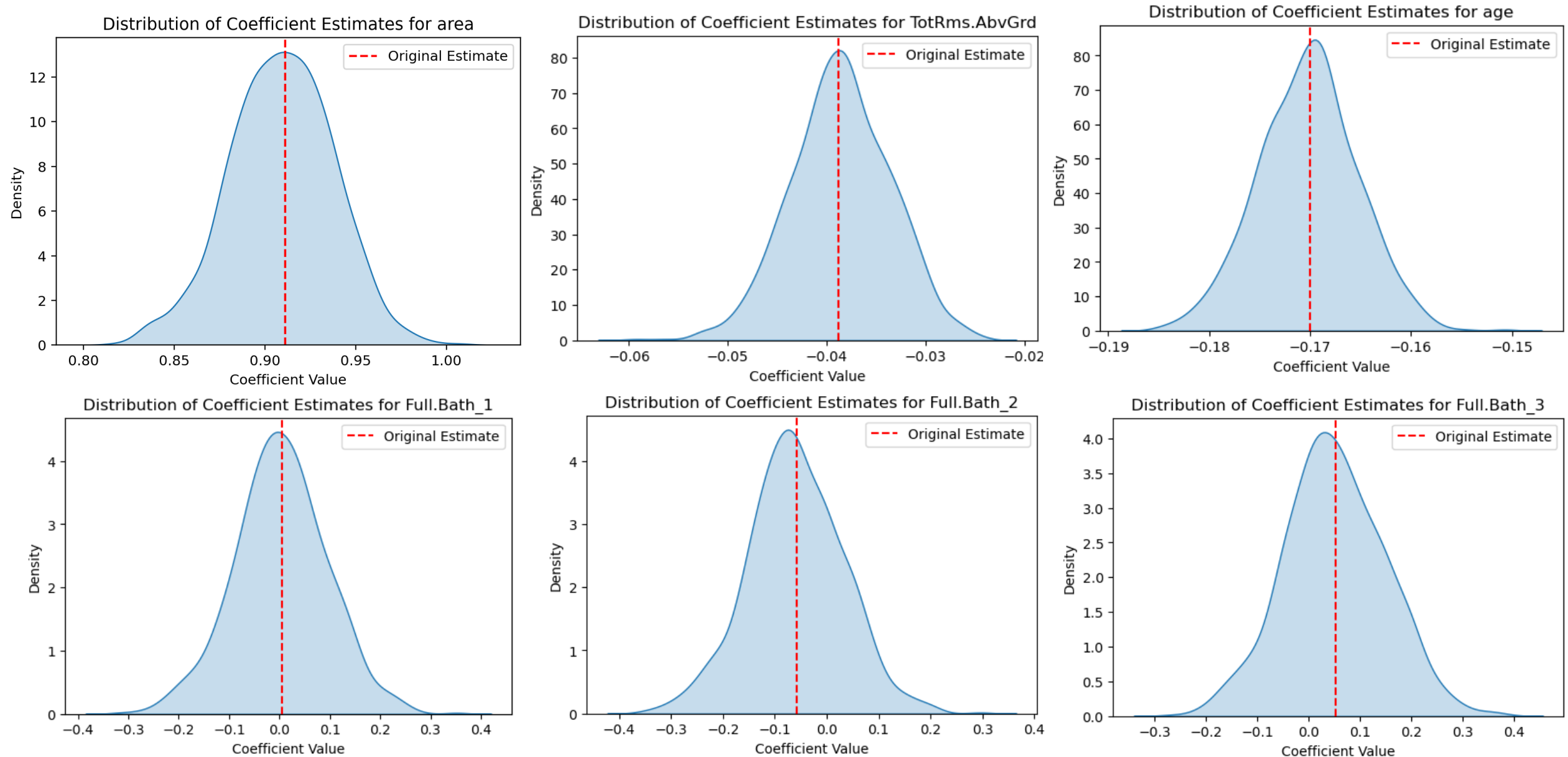
### Add intercept:
X.insert(0, 'const', 1.0)
y = np.arcsinh( rdf['price'] )
```

```
def mlr(X,y):
    reg = np.linalg.lstsq(X, y, rcond=None)
    beta_hat = reg[0]
    return beta_hat
beta_hat = mlr(X,y)
print(beta_hat)
```

```
y_hat = X @ beta_hat # Make predictions
```



Bootstrapped Coefficients



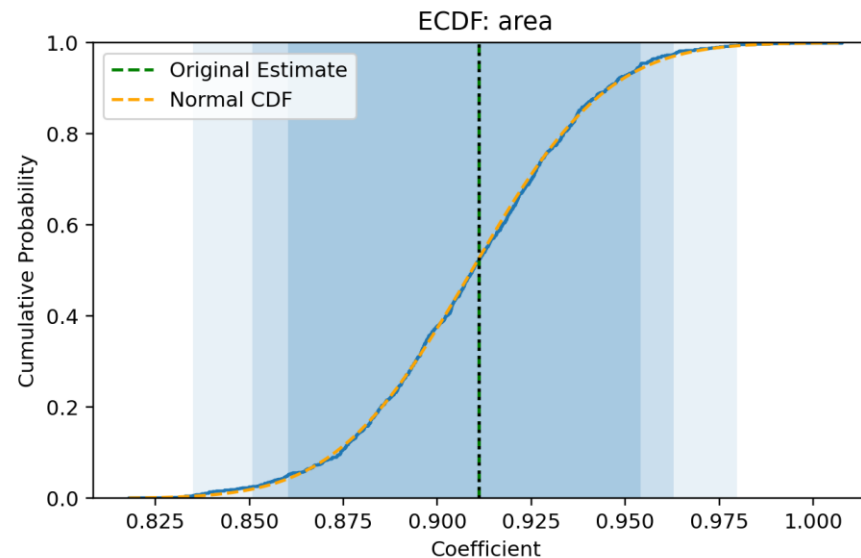
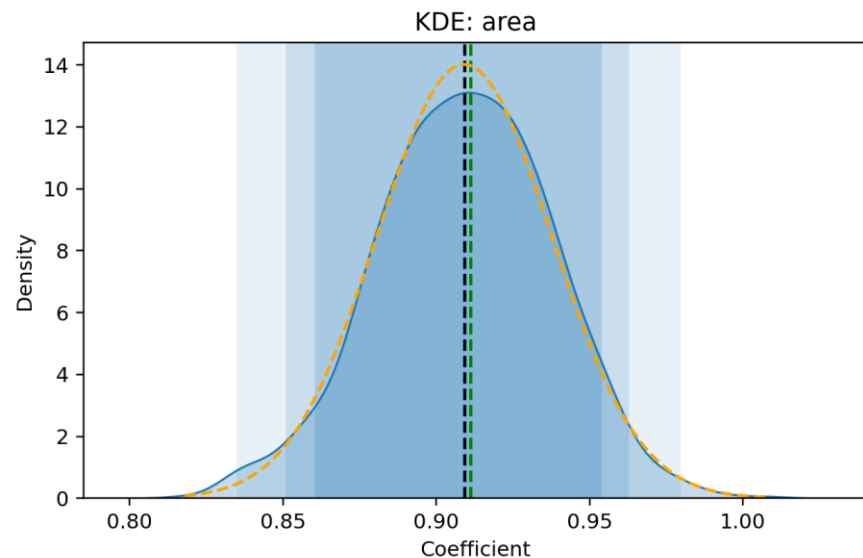
Fitting the Bootstrapped Sampling Distribution

- What if we overlay a normal with the **bootstrap mean** and the standard deviation of the bootstrap sample (our estimate of the **standard error**)?

$$\bar{\beta}_{\text{boot}} = \frac{1}{R} \sum_{b=1}^R \hat{\beta}_b, \quad \widehat{\text{SE}}(\hat{\beta}) = \sqrt{\frac{1}{R-1} \sum_{b=1}^R (\hat{\beta}_b - \bar{\beta}_{\text{boot}})^2}$$

```
vars = X.columns.to_list()

for idx, var in enumerate(vars):
    sample = results[:, idx]
    mean = np.mean(sample)
    standard_error = np.std(sample, ddof=1)
```



```
ax[0].plot(
    x_grid,
    norm.pdf(x_grid, loc=mean, scale=standard_error),
    color='orange',
    linestyle='--',
    label='Normal PDF'
)
```

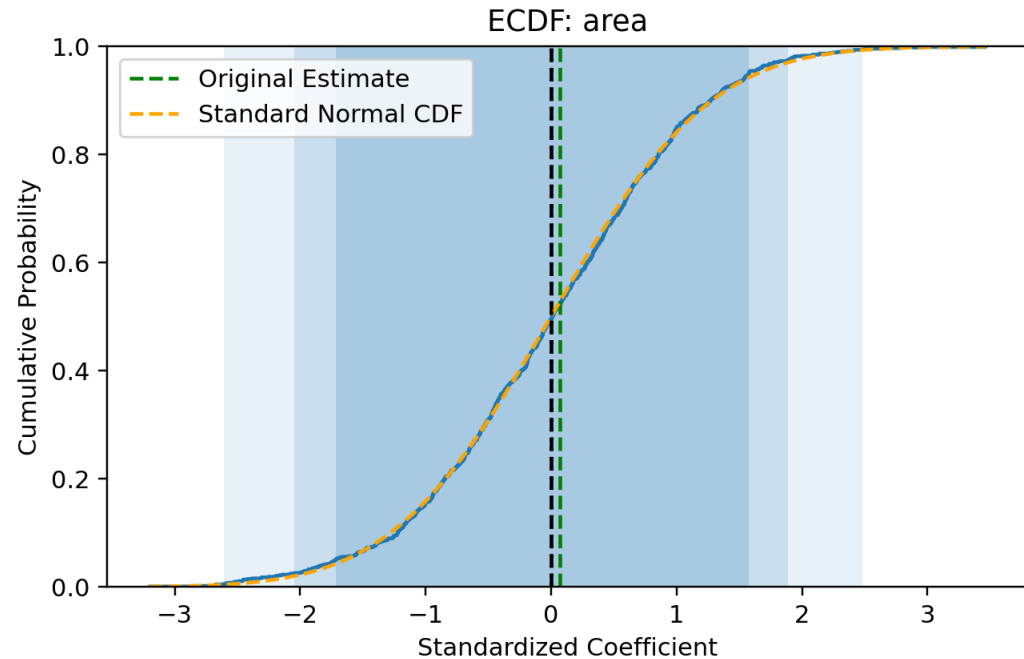
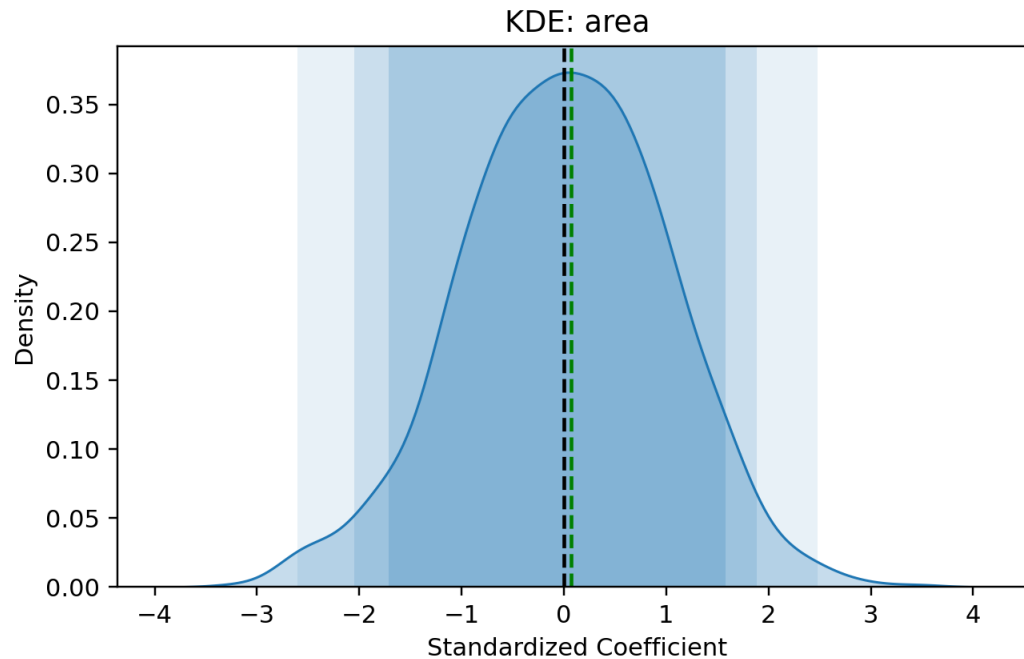
```
ax[1].plot(
    x_grid,
    norm.cdf(x_grid, loc=mean, scale=standard_error),
    color='orange',
    linestyle='--',
    label='Normal CDF'
)
```

z-Score Normalization

- Let's go a step further and z-score normalize our estimates:

$$Z_b = \frac{\hat{\beta}_b - \hat{\beta}}{\widehat{\text{SE}}(\hat{\beta})}$$

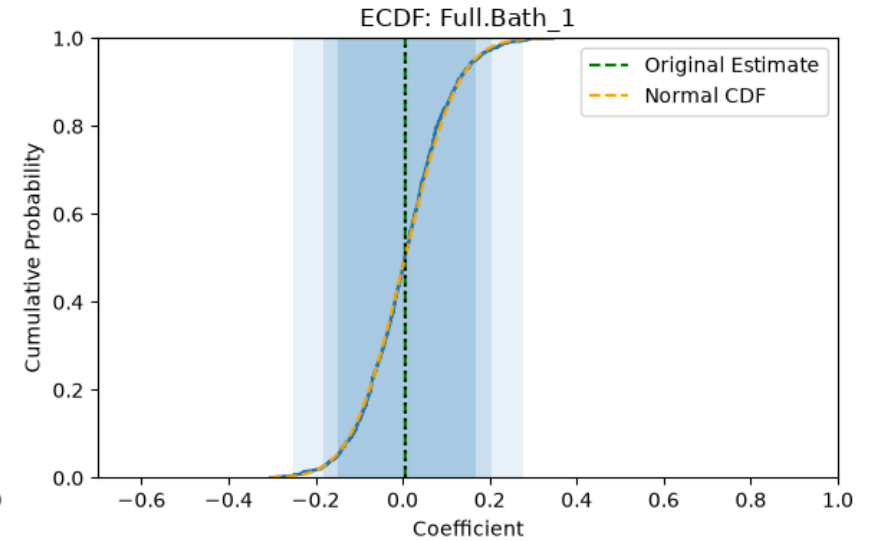
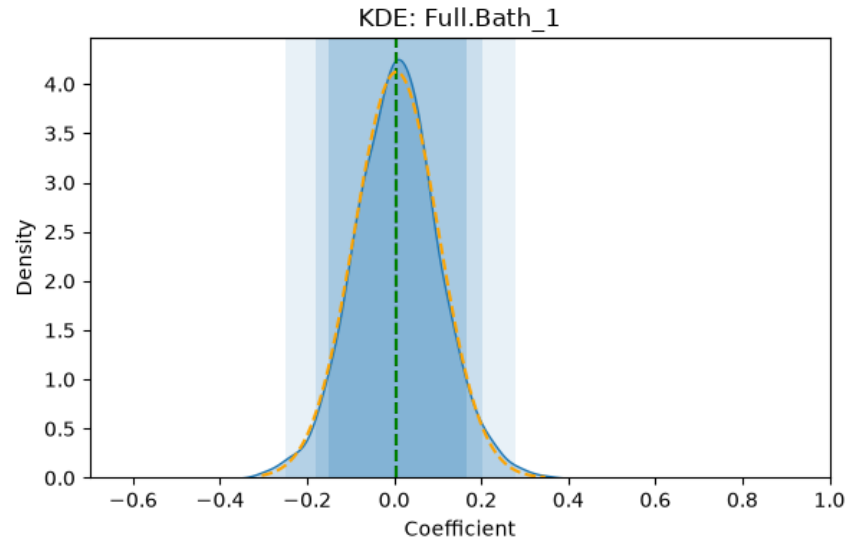
- Are the Z_b values approximately standard normal? Look at what happens to the confidence intervals



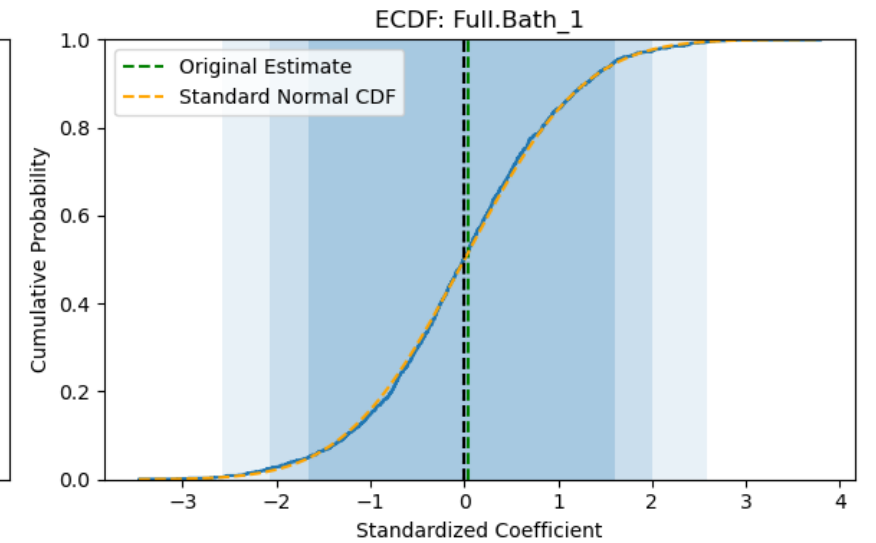
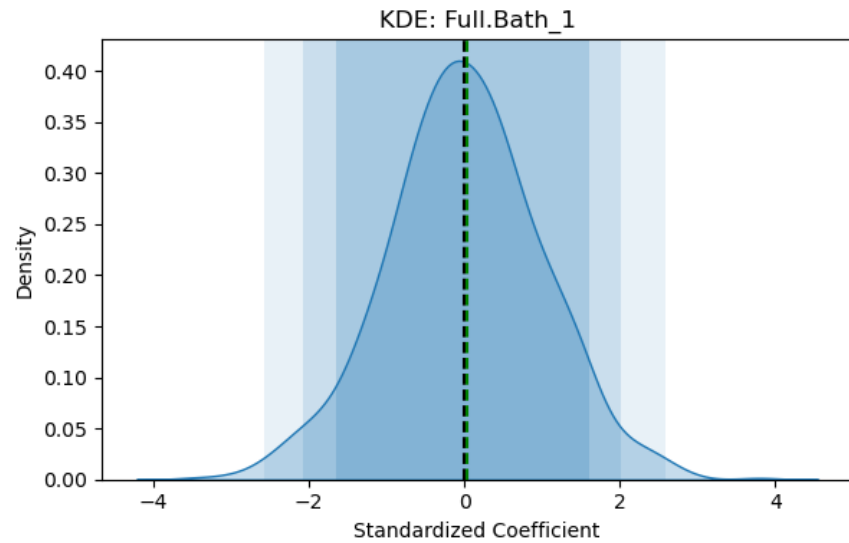
Raw vs. z-Scored

Bootstrap
coefficients

Full.Bath_1



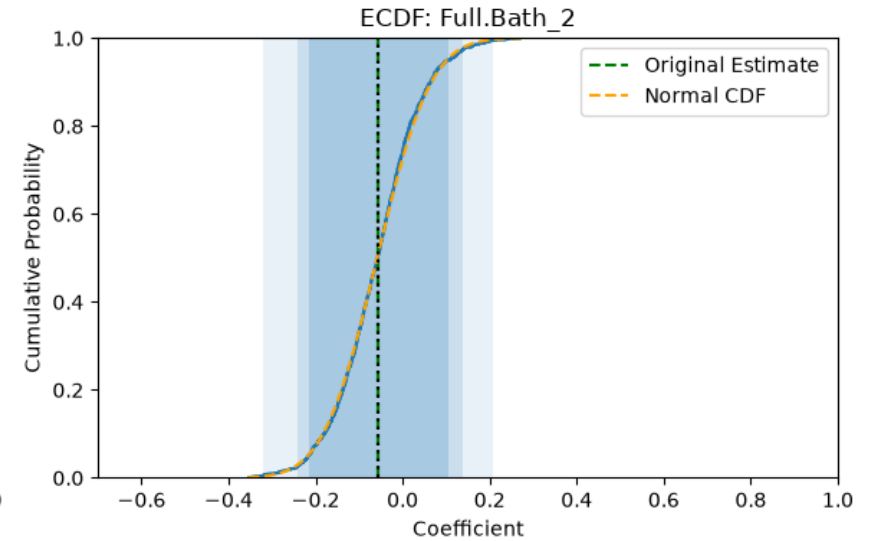
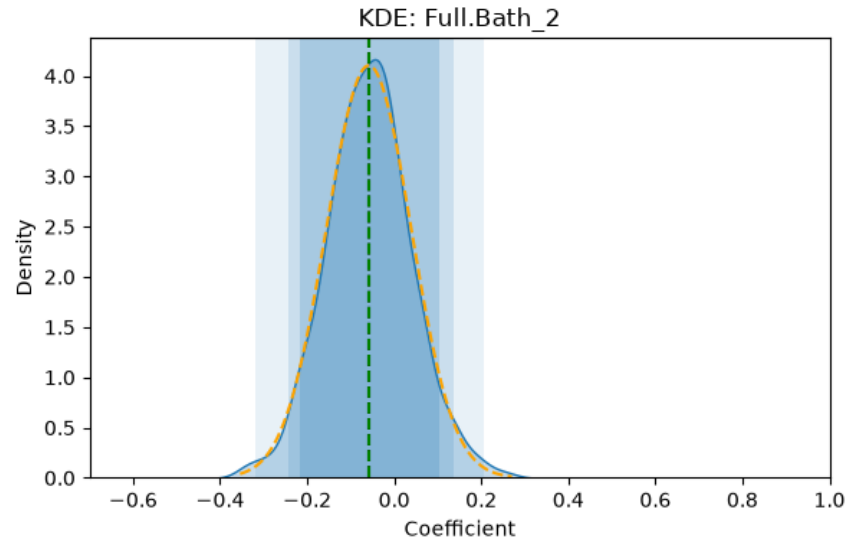
z-scored
 $(b - \text{mean}) / \text{SE}$



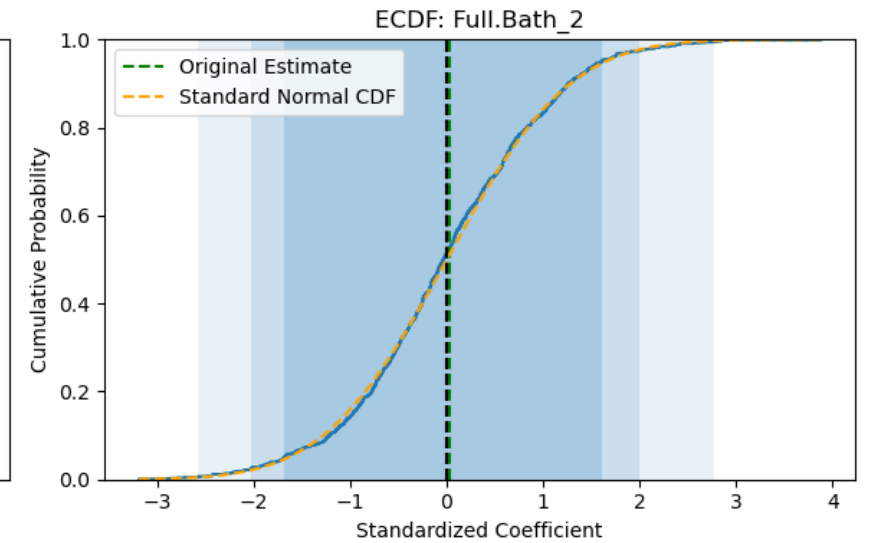
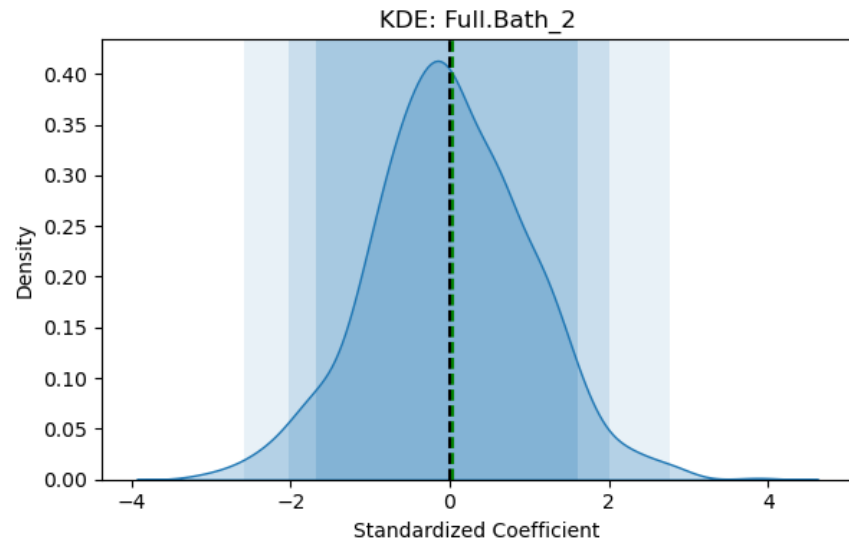
Raw vs. z-Scored

Bootstrap
coefficients

Full.Bath_2



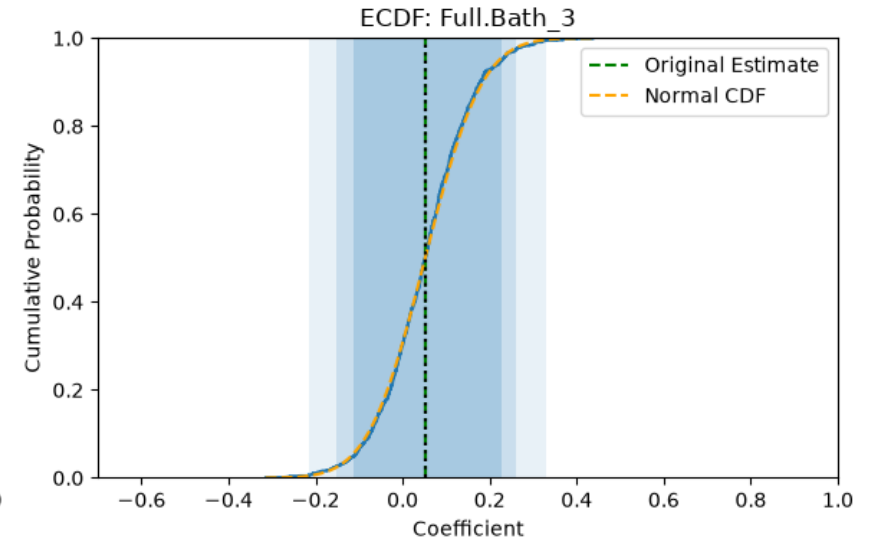
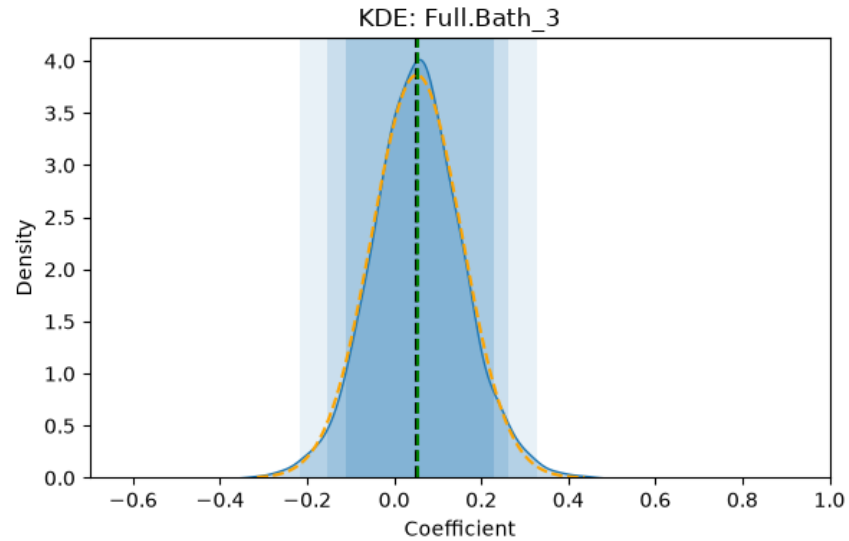
z-scored
 $(b - \text{mean}) / \text{SE}$



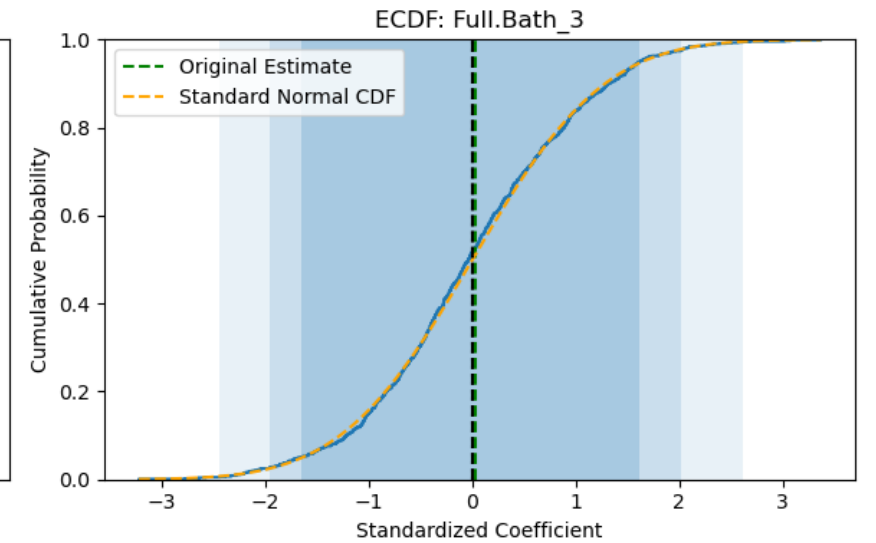
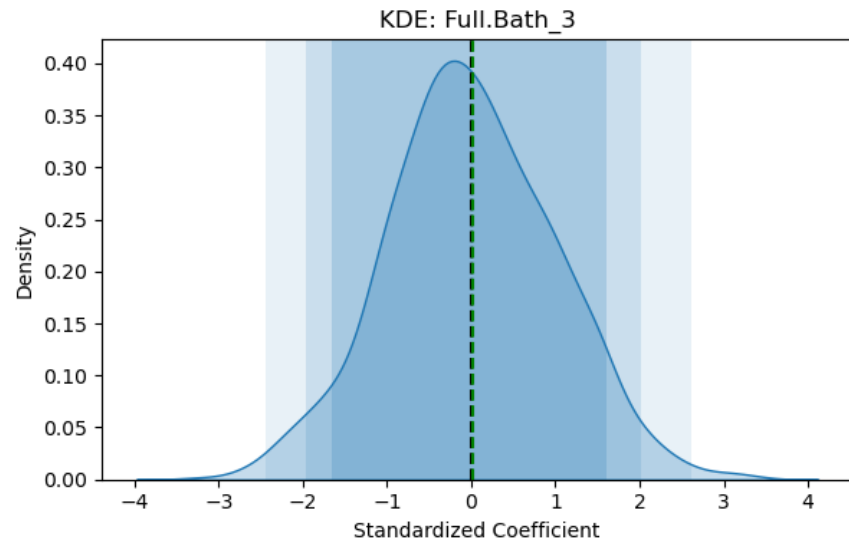
Raw vs. z-Scored

Bootstrap
coefficients

Full.Bath_3



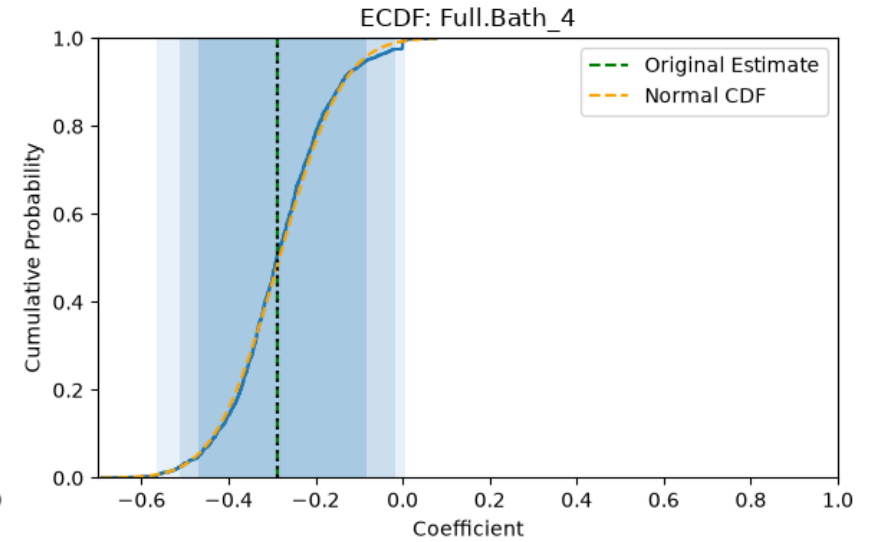
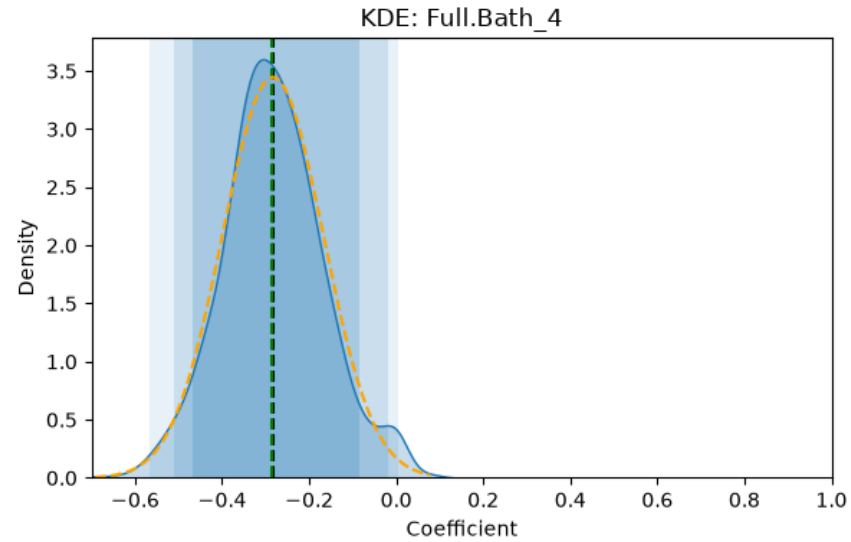
z-scored
 $(b - \text{mean}) / \text{SE}$



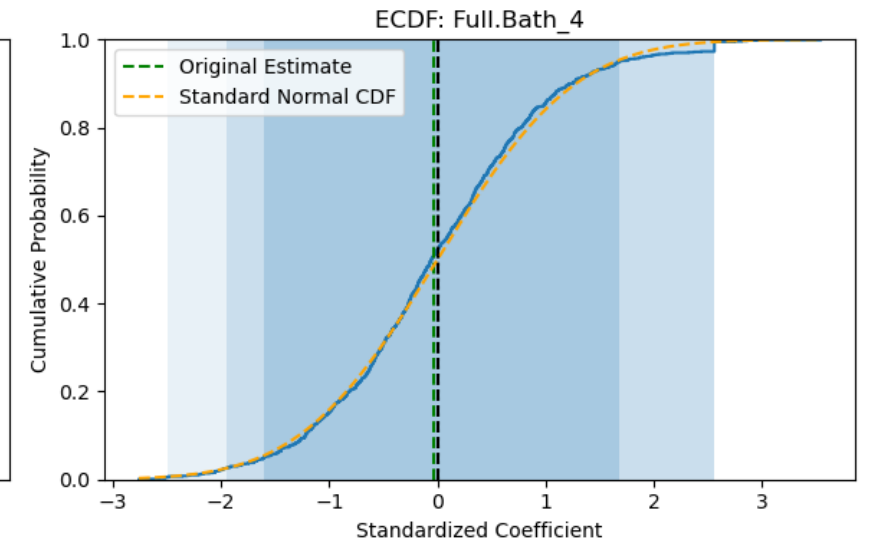
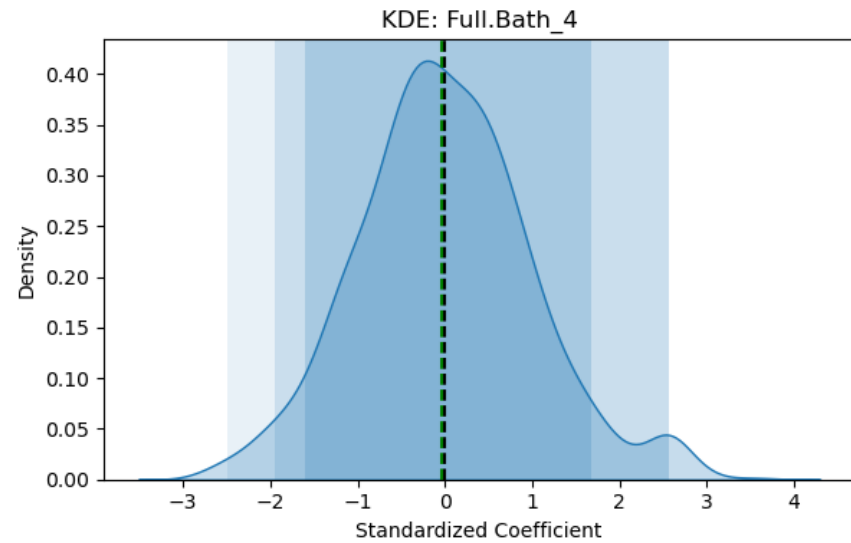
Raw vs. z-Scored

Bootstrap
coefficients

Full.Bath_4

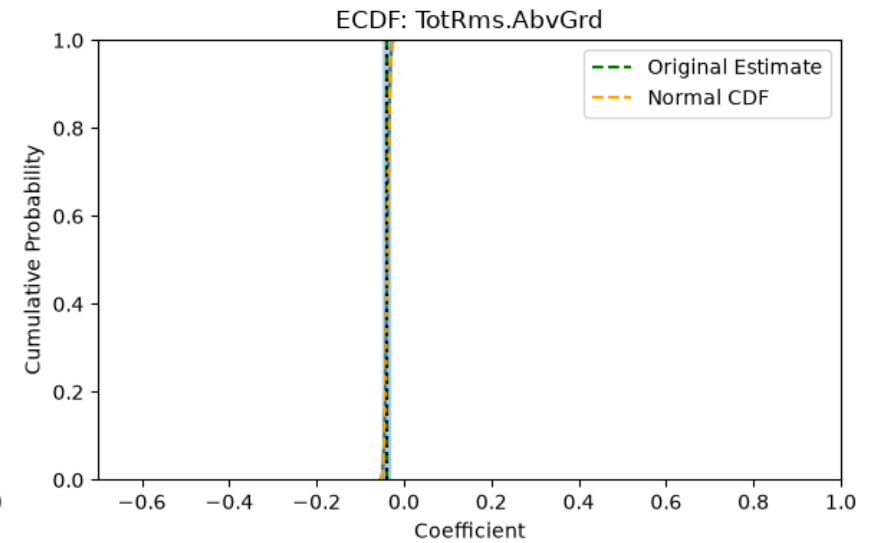
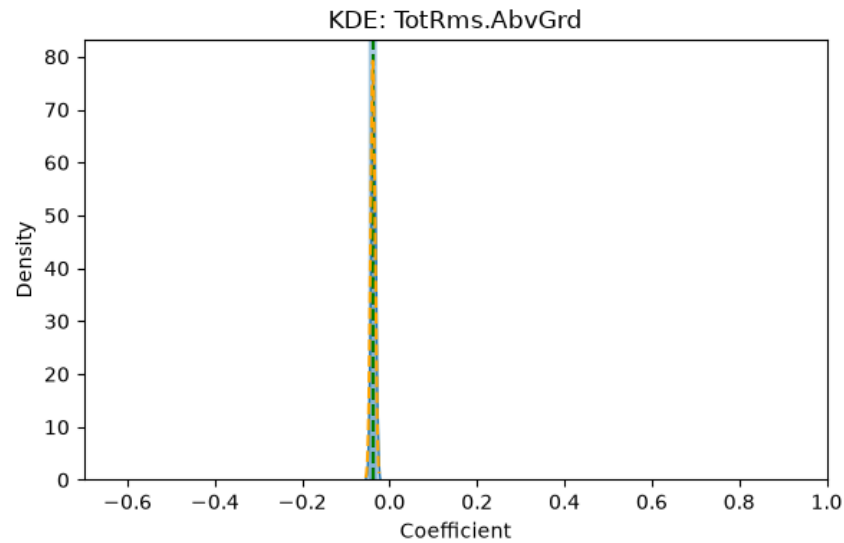


z-scored
 $(b - \text{mean}) / \text{SE}$

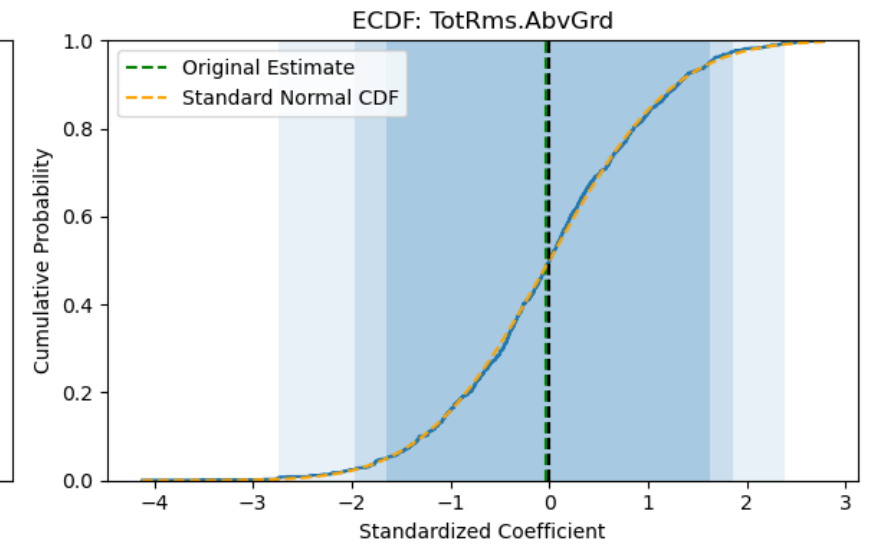
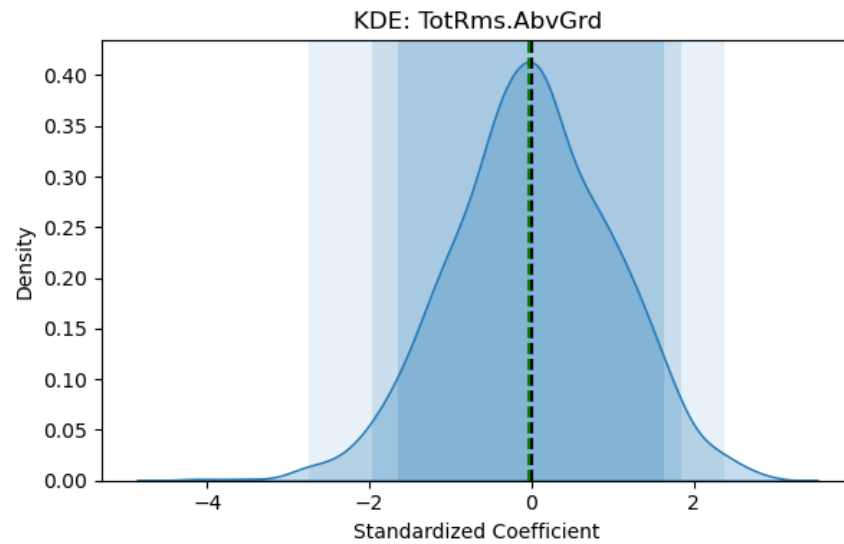


Raw vs. z-Scored

**Bootstrap
coefficients**
TotRms.AbvGrd

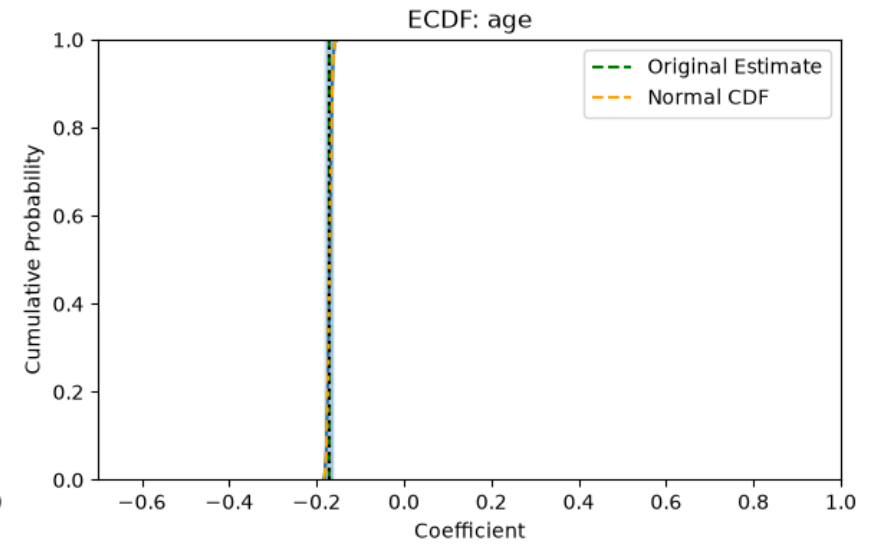
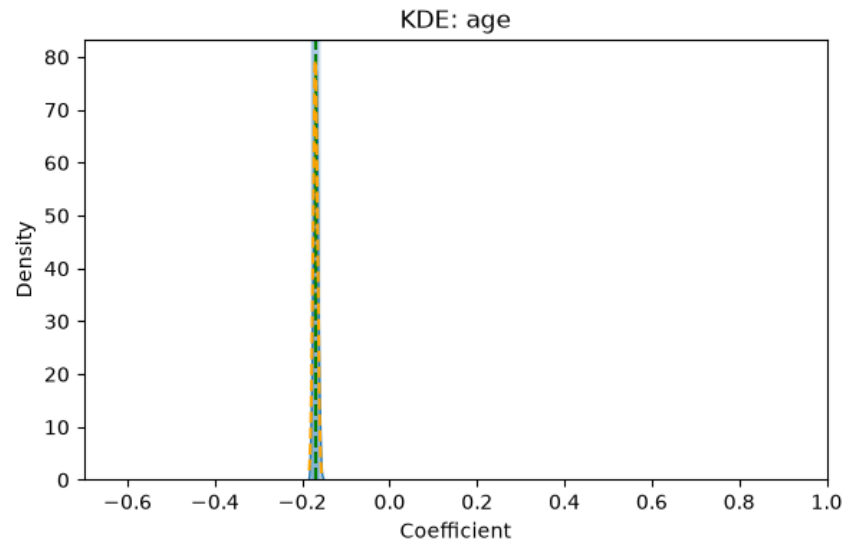


z-scored
 $(b - \text{mean}) / \text{SE}$

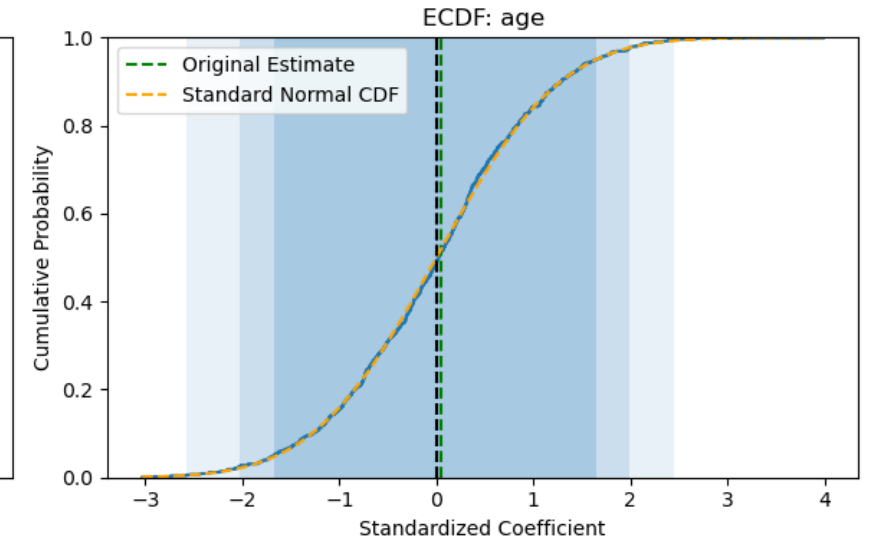
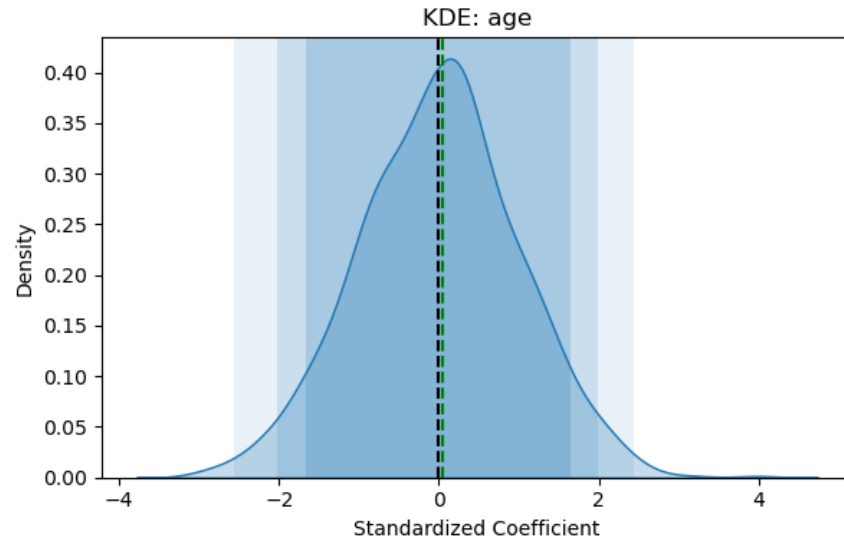


Raw vs. z-Scored

**Bootstrap
coefficients**
age

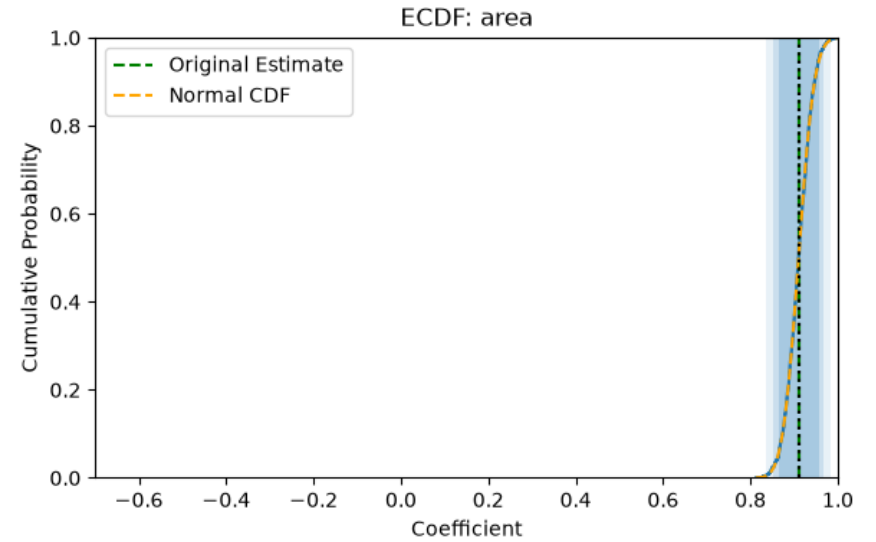
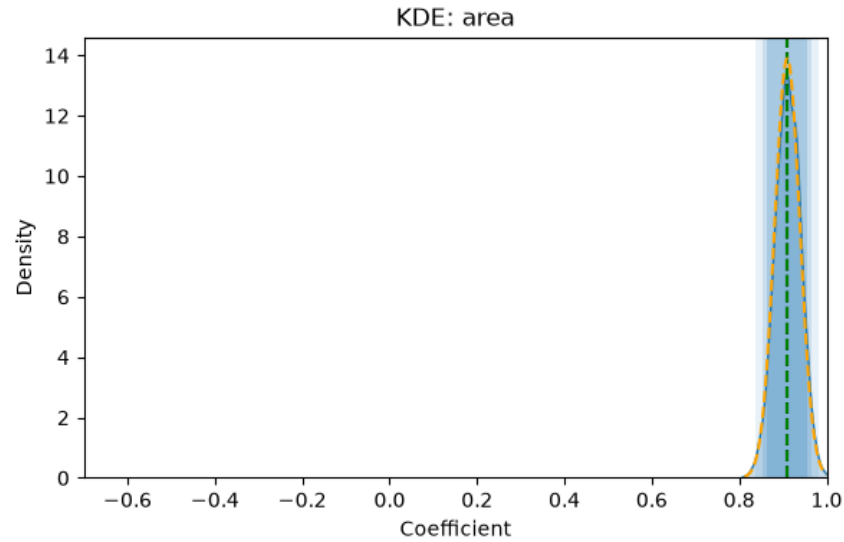


z-scored
 $(b - \text{mean}) / \text{SE}$

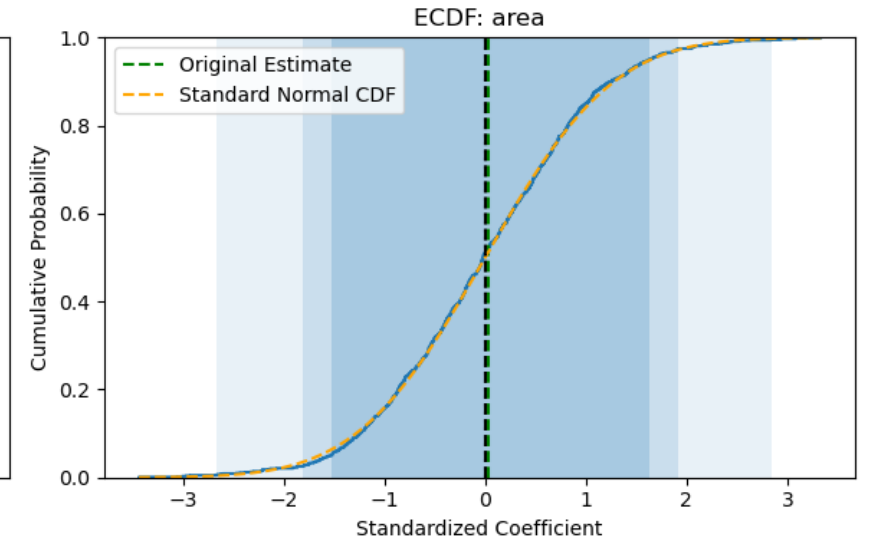
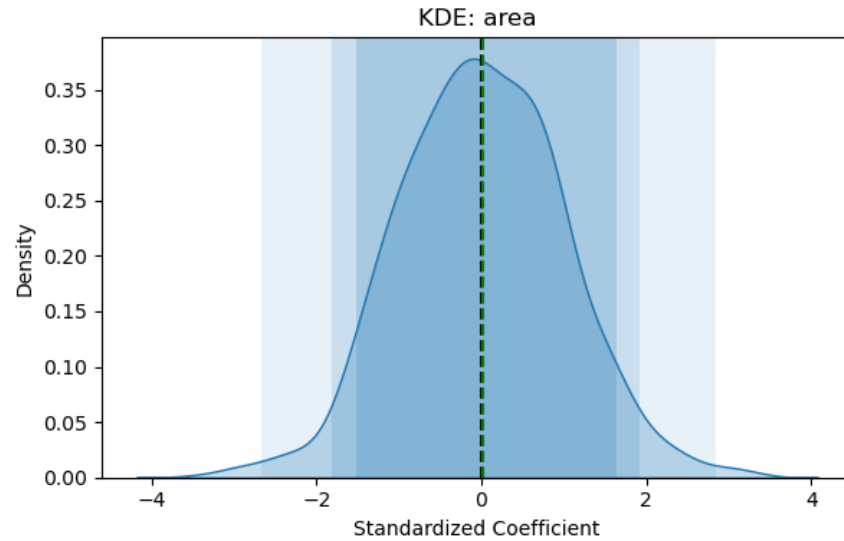


Raw vs. z-Scored

Bootstrap
coefficients
area



z-scored
 $(b - \text{mean}) / \text{SE}$



But Wait

- After z-score normalizing, we got roughly **the same picture, over and over**
- If we get basically the same picture every time, and the bootstrap ECDF and the standard Normal align almost perfectly, then can't we just approximate the sampling distribution of Z-scores by a standard Normal?
- This is the basic logic behind classical asymptotic inferential statistics

Checkpoint: Discuss with someone sitting next to you

- A colleague bootstraps the mean of 25 house prices: the bootstrap distribution is centered at \$182k with SD \$9k. They write "house prices are $\$182\text{k} \pm \$9\text{k}$." What's wrong? About how spread out are the prices themselves?
- With 100 houses instead of 25, about what SD would their bootstrap distribution have?
- Why do we resample whole rows (price and area together) instead of resampling price and area separately?
- Predict: every z-scored coefficient looked standard normal. Would the bootstrap still give that picture for (a) the maximum price, (b) the coefficient on a house style that only 8 houses in the data have, (c) the same regression fit to just 5 houses?

Why it works: the CLT

The Law of Large Numbers

- **Law of Large Numbers**: if $X_1, \dots, X_n$ are IID with mean μ , then for any $\epsilon > 0$

$$\Pr \left[|\bar{X}_n - \mu| > \epsilon \right] \rightarrow 0$$

- This is **consistency**: one long sample's average approaches μ
- It is not unbiasedness. $E[\bar{X}] = \mu$ is true for every n , even $n = 1$

Estimator	Unbiased?	Consistent?
$\bar{x}$	yes	yes
X_1 (first observation only)	yes	no
$\bar{x} + 1/n$	no	yes

- Unbiased plus variance going to zero gives consistency:

$$\Pr \left[|\bar{X}_n - \mu| > \epsilon \right] \leq \frac{\sigma^2}{n\epsilon^2} \rightarrow 0$$

Two Kinds of Getting Close

- **In probability**: the values get close. $\Pr[|X_n - X| > \varepsilon] \rightarrow 0$. This is the LLN
- **In distribution**: only the histogram (CDF) gets close. $\Pr[X_n \leq z] \rightarrow \Pr[X \leq z]$. This is the CLT
- Example: if $X \sim N(0, 1)$ and $X_n = -X$, then X_n has exactly the right distribution but is never close to X
- Probability implies distribution, but not the other way around (unless the limit is a constant)

The Central Limit Theorem

- Why is this happening?
- **Central Limit Theorem:** let $X_1, \dots, X_n$ be IID random variables, each with mean μ and variance σ^2 . Then as n goes to infinity,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1)$$

- This is what we observed above: for a big enough sample, that quantity is roughly normal
- We replace $\bar{x}$ by our parameter estimate, and $\sigma/\sqrt{n}$ by the standard error of the parameter, the SD of the sampling distribution (do not divide by $\sqrt{n}$ again; it's baked into the bootstrap estimate)
- For many of the quantities we study, there is a sum of IID random variables happening somewhere, to which the CLT applies

Confidence intervals and tests

Just Memorize These

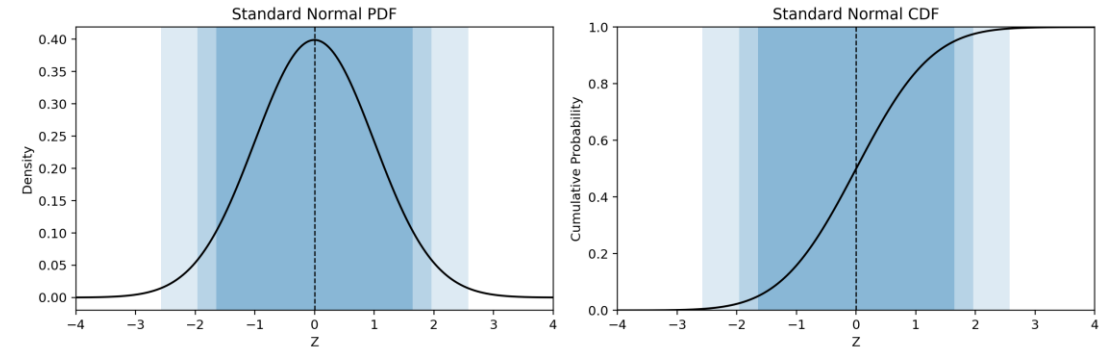
- Instead of taking the quantiles of our bootstrap sample, take the quantiles of the standard Normal, then invert the z-score normalization

Confidence	Lower	Upper	Level (α)
90%	-1.645	+1.645	.10
95%	-1.960	+1.960	.05
99%	-2.576	+2.576	.01

- So the 90% CI comes from this bound:

$$\left| \frac{\beta_0 - \hat{\beta}}{\widehat{SE}(\hat{\beta})} \right| \leq 1.645$$

$$\hat{\beta} - 1.645 \times \widehat{SE}(\hat{\beta}) \leq \beta_0 \leq \hat{\beta} + 1.645 \times \widehat{SE}(\hat{\beta})$$



Constructing Confidence Intervals

$$90\% : \quad (\hat{\beta} - 1.645 \times \widehat{SE}(\hat{\beta}), \quad \hat{\beta} + 1.645 \times \widehat{SE}(\hat{\beta}))$$

$$95\% : \quad (\hat{\beta} - 1.960 \times \widehat{SE}(\hat{\beta}), \quad \hat{\beta} + 1.960 \times \widehat{SE}(\hat{\beta}))$$

$$99\% : \quad (\hat{\beta} - 2.576 \times \widehat{SE}(\hat{\beta}), \quad \hat{\beta} + 2.576 \times \widehat{SE}(\hat{\beta}))$$

Thought Experiment: Would the CLT work in each case?

For each one: does "estimate ± 1.96 SE" work? If not, what breaks down each time?

- 1. You survey 100 students that come in 10 friend groups whose members answer the same way
- 2. You average the profit ratio (this year's profit $\div$ last year's) across 100 stores, where many stores roughly broke even last year
- 3. A condition affects 1 in 100 people. You test 100 people and find 0 cases
- 4. You bootstrap the largest sale price in the Ames data
- 5. You poll 10,000 people, but only those who answer a landline at 2 pm

Exercise on the board

- You run a regression and get a coefficient of 10 with a standard error of 5 for a key coefficient
- Compute 90/95/99% confidence intervals for the coefficient
- For which levels of confidence is a statistically significant difference from 0 detected?

What Is a Confidence Interval?

- Suppose our estimator is approximately normally distributed:

$$\frac{\hat{\beta} - \beta_0}{\widehat{SE}(\hat{\beta})} \sim N(0, 1)$$

- Choose a **critical value** c such that:

$$\Pr \left[\left| \frac{\hat{\beta} - \beta_0}{\widehat{SE}(\hat{\beta})} \right| \leq c \right] = 1 - \alpha$$

- What is random? The interval, because $\hat{\beta}$ changes with every sample. What is fixed? β_0 . **The 95% describes the procedure, not any one interval**
- Rearrange to get:

$$\Pr_{\hat{\beta}} \left[\hat{\beta} - c \times SE(\hat{\beta}) \leq \beta_0 \leq \hat{\beta} + c \times SE(\hat{\beta}) \right] = 1 - \alpha$$

- This is a statistical reliability claim, not a statement about the value of the true parameter β_0

CI to Hypothesis Test: TEST = INTERVAL

- If we hypothesize β_0 the take some value, usually zero, this becomes a **hypothesis test**: is the hypothesized value β_0 in the interval or not?

$$\Pr_{\hat{\beta}} \left[\hat{\beta} - c \times \text{SE}(\hat{\beta}) \leq 0 \leq \hat{\beta} + c \times \text{SE}(\hat{\beta}) \right] = 1 - \alpha$$

- Is β_0 (often 0) outside the CI? If so, reject H_0 ; if it is inside the CI, fail to reject H_0

The CI Is Not...

- This is just not $\Pr_{\beta_0}[|\hat{\beta} - \beta_0| \leq c]$
- It couldn't be this: we don't know the true β_0
- If β_0 was random, and had a distribution associated with it, then you'd use Bayes' Rule to observe data and update that distribution
- We often say things like, "Gravity is -9.81 m/s^2 ." We don't often say things like, "Gravity is normally distributed with mean -9.81 m/s^2 and standard deviation $.0051$ "
- If you **did** talk like that, you'd be a Bayesian statistician, and we'll talk about Bayesian statistics at the end of the course

The CI Is...

- If we did this kind of analysis many times, what range of values is plausible for our estimates?
- A set of parameter values that our data do not allow us to reject, using a procedure designed to exclude the true parameter only a specified fraction of the time (the Type 1 error probability).
- The probability that the **random** CI covers the true β_0 is $1 - \alpha$; we **expect** the CI to cover the true β_0 with probability $1 - \alpha$
- If we repeatedly sampled data and constructed a new confidence interval each time, approximately 95% of those intervals would contain the fixed, unknown parameter β_0 (the bootstrapping connection is clear here)

Coverage You Can Count

- **At the boards.** Population $\{1, 3, 5\}$, draw $n = 2$ with replacement, so $\mu = 3$
- 1. List all 9 possible samples and their $\bar{x}$
- 2. Our interval rule is $\bar{x} \pm 1.5$. For how many of the 9 samples does the interval contain $\mu = 3$? What is the coverage?
- 3. Repeat for $\bar{x} \pm 0.5$ and $\bar{x} \pm 2.5$
- 4. You drew $(1, 1)$, so your interval is $[-0.5, 2.5]$. What is the probability that this interval contains μ ?

Five Things to Think About

1. The parameter is fixed but unknown. We're trying to learn something about a property of the population.
2. Our estimates/confidence intervals are random variables of the data. Different samples produce different estimates, even for the same population doesn't change.
3. The standard error measures variability in our estimators, and bootstrapping gives us a practical way to estimate it.
4. The confidence interval translates variability into uncertainty about the parameter. Under the Normal approximation, that's $\pm 1.96 \times \text{SE}$. A hypothesized value outside the corresponding interval is rejected as implausible given the data and estimator.
5. The Bootstrap and the CLT are different approaches to understanding the underlying sampling distribution and making inferences about it, with their own strengths and weaknesses.

At the Boards: Does It Pay for Itself?

- A retail chain pilots an energy-management system in 5 stores. It costs **\$200 per store per month**. Monthly savings on each store's electric bill: **\$100, \$300, \$400, \$500, \$700**
- 1. Compute $\bar{x}$, s^2 , and the estimated SE
- 2. Build a 95% CI with 1.96. Does the system save money at all? (Test H_0 : average savings = 0)
- 3. Does it pay for itself? Is the \$200 break-even inside or outside your CI?
- 4. With only $n = 5$, the t critical value is 2.78 instead of 1.96. Redo questions 2 and 3. What should the CFO do?
- 5. The CFO writes: "There is a 95% chance the system saves between \$204 and \$596 per store." Rewrite it correctly
- 6. All 5 pilot stores are in Phoenix. What does your CI tell you about the stores in Maine?

Example

- Load the Metabric breast cancer dataset, and describe/plot `Overall Survival (Months)`.
- Describe a strategy for selecting variables that you'll use to predict `Overall Survival (Months)`, then implement your strategy and run your regression.
- Construct 90/95/99% non-parametric Bootstrap confidence intervals for the coefficients of your regression, as well as ones based on the Normal approximation using the standard deviation of the bootstrap sample as the standard error of the sampling distribution. How do they compare?
- Which variables are statistically significant, and at what levels?

Summary

Summary

- The **standard error** is the standard deviation of the sampling distribution
- The **bootstrap** estimates the spread and shape of the sampling distribution from one sample
- The **CLT** says that, standardized, the sampling distribution is approximately $N(0, 1)$: that's why the bootstrap pictures all looked the same
- So we can replace bootstrap quantiles with $\pm 1.645, 1.96, 2.576$:

$$\hat{\theta} \pm c \times \widehat{SE}(\hat{\theta})$$

- A CI is a reliability claim about the procedure. A test asks whether the null value is inside the CI
- It breaks with dependence, heavy tails, small samples of skewed data, and estimates that aren't averages