Precept 7: Multiple Regression

Soc 400: Applied Social Statistics

Ziyao Tian[1]

Princeton University

October 25, 2018

[1]This draws material from Shay O'Brien, Simone Zhang and Matt Blackwell.

Today's Agenda

- Slides
  - Sampling distribution & standard error
  - Matrix notation for linear regression
  - R-squared & F-test
  - Bootstrap
- RStudio
  - Basic matrix operations
  - Interpretating multiple regression
  - F-test
  - Bootstrap

Estimands, Estimators, and Estimates

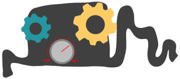
Estimands, Estimators, and Estimates

The goal of statistical inference is to learn about the unobserved population distribution, which can be characterized by **parameters**.

Estimands, Estimators, and Estimates

The goal of statistical inference is to learn about the unobserved population distribution, which can be characterized by **parameters**.

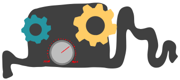
- **Estimands** are the parameters that we aim to estimate. Often written with greek letters (e.g. μ, θ, population mean) : $\frac{1}{N}\sum_{i=1}^{N} y_i$



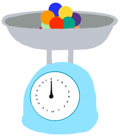
Estimands, Estimators, and Estimates

The goal of statistical inference is to learn about the unobserved population distribution, which can be characterized by **parameters**.

- **Estimands** are the parameters that we aim to estimate. Often written with greek letters (e.g. μ, θ, population mean) : $\frac{1}{N}\sum_{i=1}^{N} y_i$



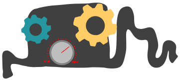
- **Estimators** are functions of sample data (i.e. **statistics**) which we use to learn about the estimands. Often denoted with a "hat" (e.g. $\hat{\mu}, \hat{\theta}$)



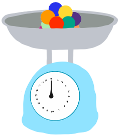
Estimands, Estimators, and Estimates

The goal of statistical inference is to learn about the unobserved population distribution, which can be characterized by **parameters**.

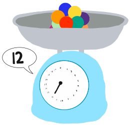
- **Estimands** are the parameters that we aim to estimate. Often written with greek letters (e.g. μ, θ, population mean) : $\frac{1}{N}\sum_{i=1}^{N} y_i$



- **Estimators** are functions of sample data (i.e. **statistics**) which we use to learn about the estimands. Often denoted with a "hat" (e.g. $\hat{\mu}, \hat{\theta}$)



- **Estimates** are particular values of estimators that are realized in a given sample (e.g. sample mean): $\frac{1}{n}\sum_{i=1}^{n} y_i$



Why Study Estimators?

Why Study Estimators?

- Two Goals:
  1. **Inference**: How much uncertainty do we have in this estimate?

Why Study Estimators?

- Two Goals:
  1. **Inference**: How much uncertainty do we have in this estimate?
  2. **Evaluate Estimators:** How do we choose which estimator to use?

Why Study Estimators?

- Two Goals:
  1. **Inference**: How much uncertainty do we have in this estimate?
  2. **Evaluate Estimators:** How do we choose which estimator to use?
- We will consider the hypothetical **sampling distribution** of estimates we would have obtained if we had drawn **repeated samples** of size n from the population.

Why Study Estimators?

- Two Goals:
  1. **Inference**: How much uncertainty do we have in this estimate?
  2. **Evaluate Estimators:** How do we choose which estimator to use?
- We will consider the hypothetical **sampling distribution** of estimates we would have obtained if we had drawn **repeated samples** of size n from the population.
- In real applications, we cannot draw repeated samples, so we attempt to approximate the sampling distribution (either by resampling or by mathematical formulas)

Why Study Estimators?

- Two Goals:
  1. **Inference**: How much uncertainty do we have in this estimate?
  2. **Evaluate Estimators:** How do we choose which estimator to use?
- We will consider the hypothetical **sampling distribution** of estimates we would have obtained if we had drawn **repeated samples** of size n from the population.
- In real applications, we cannot draw repeated samples, so we attempt to approximate the sampling distribution (either by resampling or by mathematical formulas)

Standard Error

We refer to the standard deviation of a sampling distribution as a **standard error**.

Standard Error

We refer to the standard deviation of a sampling distribution as a **standard error**.

Two Points of Potential Confusion:

- Each sampling distribution has its own standard deviation, and therefore its own standard error.

Standard Error

We refer to the standard deviation of a sampling distribution as a **standard error**.

Two Points of Potential Confusion:

- Each sampling distribution has its own standard deviation, and therefore its own standard error.
- Some people refer to an estimated standard error as the standard error.

What we've learned (I): Population and Sample Mean

Let X_1, X_2, ... , X_n be a random sample of size n from a distribution (population) with mean μ and variance σ^2

Estimand	**Estimator**	**Sampling Dist.**
Population Mean μ	Sample Mean $\bar{X}$	$\bar{X} \overset{\text{approx.}}{\sim} N(\mu, \frac{\sigma^2}{n})$
Population Variance σ^2	$S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$	$E[S^2] = \sigma^2; S^2 \xrightarrow{n \to \infty} \sigma^2$
$SE[\bar{X}]$	$\widehat{SE}[\bar{X}] = \sqrt{\frac{S^2}{n}}$	

When σ^2 is unknown, $\bar{X} \sim N(\mu, \frac{S^2}{n})$ or $\bar{X} \sim N(\mu, \hat{SE}[\bar{X}]^2)$.
When X_i is independently drawn from a Normal distribution, $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$

What we've learned (II): Population and Sample Regression

Consider the population regression model: $Y = \beta_0 + \beta_1 X + u$

Estimand	**Estimator**	**Sampling Dist.**
β_1	$\hat{\beta}_1$	$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma_u^2}{\sum(X_i - \bar{X})^2})$
Error u	Residual $\hat{u} = y_i - \hat{y}_i$	
Error Variance σ_u^2	$\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i{}^2}{n-2}$	$E[\hat{\sigma}_u^2] = \sigma_u^2$
$SE[\hat{\beta}_1]$	$\widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}_u^2}{\sum(X_i - \bar{X})^2}}$	

What we've learned (II): Population and Sample Regression

Consider the population regression model: $Y = \beta_0 + \beta_1 X + u$

Estimand	Estimator	Sampling Dist.
β_1	$\hat{\beta}_1$	$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma_u^2}{\sum(X_i - \bar{X})^2})$
Error u	Residual $\hat{u} = y_i - \hat{y}_i$	
Error Variance σ_u^2	$\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i{}^2}{n-2}$	$E[\hat{\sigma}_u^2] = \sigma_u^2$
$SE[\hat{\beta}_1]$	$\widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}_u^2}{\sum(X_i - \bar{X})^2}}$	

...under certain conditions...

What we've learned (III): Simple OLS Assumptions and Sampling Distribution

When σ_u^2 is unknown, $\widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma_u^2}}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{\sum \hat{u_i}^2}{(n-2)\sum(X_i - \bar{X})^2}}$

Information	**Assumptions**	**Sampling Dist.**
σ_u^2 is known	1-6	$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma_u^2}{\sum(X_i - \bar{X})^2})$
σ_u^2 is unknown	1-6	$\hat{\beta}_1 \sim t_{n-2}(\beta_1, \widehat{SE}(\hat{\beta}_1)^2)$
σ_u^2 is unknown	1-5 and n is large	$\hat{\beta}_1 \sim t_{n-2}(\beta_1, \widehat{SE}(\hat{\beta}_1)^2)$

Matrix Notation Overview

	Old notation (for univariate regression)	Matrix notation
Linear model	$Y_i = \beta_0 + \beta_1 X_i + u$	$y = X\beta + u$
Coefficient	$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$	$\hat{\beta} = (X'X)^{-1} X' y$
Homoskedasticity assumption	$Var[u \mid X] = \sigma_u^2$	$Var[u \mid X] = \sigma_u^2 I_n$
Variance of coefficient	$\frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$	$\sigma_u^2 (X'X)^{-1}$
Error variance	$\frac{\sum_{i=1}^n \hat{u}_i^2}{n-2}$	$\hat{\sigma}_u^2 = \frac{\hat{u}'\hat{u}}{n-k-1}$ $\quad (y - X\hat{\beta})'(y - X\hat{\beta})$
SS_{tot}	$\sum_{i=1}^n (Y_i - \bar{Y})^2$	$(y - \bar{y})'(y - \bar{y})$
SS_{res}	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$\hat{u}'\hat{u}$

Matrix Notation

- $\mathbf{X}$ is the $n \times (K+1)$ design matrix of independent variables
- β be the $(K+1) \times 1$ column vector of coefficients.
- $\mathbf{X}\beta$ will be $n \times 1$:

- We can compactly write the linear model as the following:

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times 1)}{\mathbf{X}\beta} + \underset{(n\times 1)}{\mathbf{u}}$$

$$\underset{(n\times 1)}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \underset{(n\times(K+1))}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \quad \underset{((K+1)\times 1)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

OLS Estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- What's the intuition here?
- "Numerator" $\mathbf{X}'\mathbf{y}$: is roughly composed of the covariances between the columns of $\mathbf{X}$ and $\mathbf{y}$
- "Denominator" $\mathbf{X}'\mathbf{X}$ is roughly composed of the sample variances and covariances of variables within $\mathbf{X}$
- Thus, we have something like:

$$\widehat{\boldsymbol{\beta}} \approx (\text{variance of } \mathbf{X})^{-1}(\text{covariance of } \mathbf{X} \text{ \& } \mathbf{y})$$

- This is a rough sketch and isn't strictly true, but it can provide intuition.

Variance-Covariance Matrix

- The homoskedasticity assumption is different: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- In order to investigate this, we need to know what the variance of a vector is.
- The variance of a vector is actually a matrix:

$$\text{var}[\mathbf{u}] = \Sigma_u = \begin{bmatrix} \text{var}(u_1) & \text{cov}(u_1, u_2) & \dots & \text{cov}(u_1, u_n) \\ \text{cov}(u_2, u_1) & \text{var}(u_2) & \dots & \text{cov}(u_2, u_n) \\ \vdots & & \ddots & \\ \text{cov}(u_n, u_1) & \text{cov}(u_n, u_2) & \dots & \text{var}(u_n) \end{bmatrix}$$

- This matrix is **symmetric** since $\text{cov}(u_i, u_j) = \text{cov}(u_j, u_i)$

Matrix Version of Homoskedasticity

- Once again: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- $\mathbf{I}_n$ is the $n \times n$ identity matrix
- Visually:

$$\text{var}[\mathbf{u}|\mathbf{X}] = \sigma_u^2 \mathbf{I}_n = \begin{bmatrix} \sigma_u^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_u^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_u^2 \end{bmatrix}$$

- In less matrix notation:
  - $\text{var}(u_i|X) = \sigma_u^2$ for all i (constant variance)
  - $\text{cov}(u_i, u_j) = 0$ for all $i \neq j$ (implied by iid)

Sampling Variance for OLS Estimates

- Under assumptions 1-5, the sampling variance of the OLS estimator can be written in matrix form as the following:

$$\text{var}[\widehat{\beta}] = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$$

- This matrix looks like this:

	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\cdots$	$\widehat{\beta}_K$
$\widehat{\beta}_0$	$\text{var}[\widehat{\beta}_0]$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_1]$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_2]$	$\cdots$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_K]$
$\widehat{\beta}_1$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_1]$	$\text{var}[\widehat{\beta}_1]$	$\text{cov}[\widehat{\beta}_1, \widehat{\beta}_2]$	$\cdots$	$\text{cov}[\widehat{\beta}_1, \widehat{\beta}_K]$
$\widehat{\beta}_2$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_2]$	$\text{cov}[\widehat{\beta}_1, \widehat{\beta}_2]$	$\text{var}[\widehat{\beta}_2]$	$\cdots$	$\text{cov}[\widehat{\beta}_2, \widehat{\beta}_K]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\widehat{\beta}_K$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_K]$	$\text{cov}[\widehat{\beta}_K, \widehat{\beta}_1]$	$\text{cov}[\widehat{\beta}_K, \widehat{\beta}_2]$	$\cdots$	$\text{var}[\widehat{\beta}_K]$

Estimating Error Variance

Note that we never observe the true error variance, σ_u^2. We can estimate it with the following:

$$\widehat{\sigma}_u^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-(k+1)}$$

where $n-(k+1)$ = residual degrees of freedom and

$$\widehat{\mathbf{u}}'\widehat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\widehat{\beta})'(\mathbf{y} - \mathbf{X}\widehat{\beta})$$

Prediction error

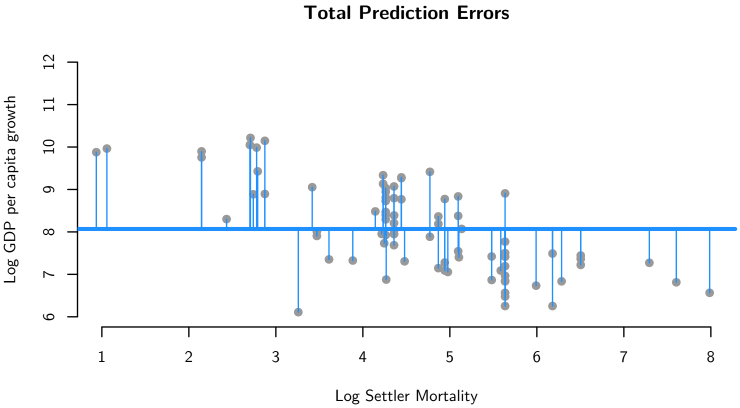
- Prediction errors without **X**: best prediction is the mean, so our squared errors, or the **total sum of squares** (SS_{tot}) would be:

$$SS_{tot} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = (\mathbf{y} - \overline{y})'(\mathbf{y} - \overline{y})$$

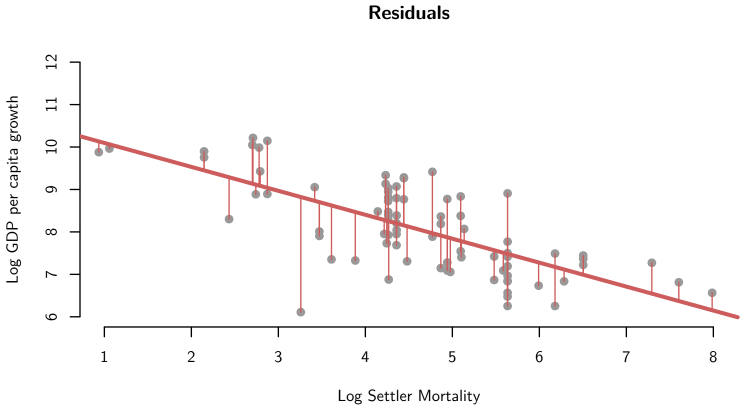
- Once we have estimated our model, we have new prediction errors, which are the **sum of the squared residuals** (SS_{res}):

$$SS_{res} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \widehat{\mathbf{u}}'\widehat{\mathbf{u}}$$

Sum of Squares



Sum of Squares



R-square

- **Coefficient of determination** or R^2:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- This is the fraction of the total prediction error eliminated by providing information in $\mathbf{X}$.

F Test Procedure

The **F statistic** can be calculated by the following procedure:

1. Fit the **Unrestricted Model (UR)** which *does not* impose H_0
2. Fit the **Restricted Model (R)** which *does* impose H_0
3. From the two results, compute the **F Statistic:**

$$F_0 = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

where **SSR**=sum of squared residuals, **q**=number of restrictions, k=number of predictors in the unrestricted model, and n= # of observations.

Intuition:

$$\frac{\text{increase in prediction error}}{\text{original prediction error}}$$

Matrix Notation Overview

	Old notation (for univariate regression)	Matrix notation
Linear model	$Y_i = \beta_0 + \beta_1 X_i + u$	$y = X\beta + u$
Coefficient	$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$	$\hat{\beta} = (X'X)^{-1}X'y$
Homoskedasticity assumption	$Var[u\|X] = \sigma_u^2$	$Var[u\|X] = \sigma_u^2 I_n$
Variance of coefficient	$\frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$	$\sigma_u^2 (X'X)^{-1}$
Error variance	$\frac{\sum_{i=1}^{n}\hat{u}_i^2}{n-2}$	$\hat{\sigma}_u^2 = \frac{\hat{u}'\hat{u}}{n-k-1}$ $\quad (y - X\hat{\beta})'(y - X\hat{\beta})$
SS_{tot}	$\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2$	$(y - \bar{y})'(y - \bar{y})$
SS_{res}	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$\hat{u}'\hat{u}$

What we've learned (IV): Multiple Regression

Consider the population regression model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$

Estimand	**Estimator**	**Sampling Dist.**
β_j	$\hat{\beta}_j$	$\hat{\beta}_j \sim N(\beta_j, \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}_{jj})$
Error $\mathbf{u}$	Residual vector $\hat{\mathbf{u}}$	
Error Variance σ_u^2	$\hat{\sigma}_u^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-(k+1)}$	$\hat{\sigma}_u^2 \sim \chi^2_{n-(k+1)}$
Standard Error of $\hat{\beta}_j$	$\widehat{SE}[\hat{\beta}_j] = \sqrt{\widehat{var}[\hat{\beta}]_{jj}} = \sqrt{\hat{\sigma}_u^2(\mathbf{X}'\mathbf{X})^{-1}_{jj}}$	

What we've learned (V): Multiple OLS Assumptions and Sampling Distribution

When σ_u^2 is unknown,

- $\hat{\sigma_u^2} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-(k+1)}$
- $\widehat{var}[\hat{\beta}] = \hat{\sigma_u^2}(\mathbf{X}'\mathbf{X})^{-1}$
- $\widehat{SE}[\hat{\beta}_j] = \sqrt{\widehat{var}[\hat{\beta}]_{jj}}$

Information	Assumptions	Sampling Dist.
σ_u^2 is known	1-6	$\hat{\beta}_j \sim N(\beta_j, \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1})_{jj})$
σ_u^2 is unknown	1-6	$\hat{\beta}_j \sim t_{n-(k+1)}(\beta_j, \widehat{SE}(\hat{\beta}_j)^2)$
σ_u^2 is unknown	1-5 and n is large	$\hat{\beta}_j \sim t_{n-(k+1)}(\beta_j, \widehat{SE}(\hat{\beta}_j)^2)$

Key Takeaway: Why Study Estimators

- We as social scientists want to communicate our understanding of the world, but with caution. We use estimate/estimator to show our theory. Are our answers good (evaluation)? How certain we are (inference)? To answer that, we need to learn about properties of estimators that we want to use and learn the sampling distribution of the estimate.

Key Takeaway: Why Study Estimators

- We as social scientists want to communicate our understanding of the world, but with caution. We use estimate/estimator to show our theory. Are our answers good (evaluation)? How certain we are (inference)? To answer that, we need to learn about properties of estimators that we want to use and learn the sampling distribution of the estimate.
- Properties and sampling distribution of estimators depend on conditions/assumptions.

Key Takeaway: Why Study Estimators

- We as social scientists want to communicate our understanding of the world, but with caution. We use estimate/estimator to show our theory. Are our answers good (evaluation)? How certain we are (inference)? To answer that, we need to learn about properties of estimators that we want to use and learn the sampling distribution of the estimate.
- Properties and sampling distribution of estimators depend on conditions/assumptions.
- Standard error is a way to describe how disperse the sampling distribution of our estimator is and thus one way to show how uncertain we are.

Key Takeaway: Why Study Estimators

- We as social scientists want to communicate our understanding of the world, but with caution. We use estimate/estimator to show our theory. Are our answers good (evaluation)? How certain we are (inference)? To answer that, we need to learn about properties of estimators that we want to use and learn the sampling distribution of the estimate.
- Properties and sampling distribution of estimators depend on conditions/assumptions.
- Standard error is a way to describe how disperse the sampling distribution of our estimator is and thus one way to show how uncertain we are.
- We have tools to learn about the SE of estimators.

Tools to learn about the Standard Error

- analytical: derive SE analytically, and estimate it from one sample $\widehat{SE}[som\hat{e}thing]$
- "omniscient": repeated sampling from a fake population (pedagogical simulation)

Tools to learn about the Standard Error

- analytical: derive SE analytically, and estimate it from one sample $\widehat{SE}[\hat{something}]$
- "omniscient": repeated sampling from a fake population (pedagogical simulation)
- resampling: draw repeated samples from the original data sample(s)

Tools to learn about the Standard Error

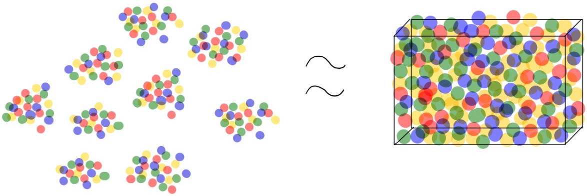
- analytical: derive SE analytically, and estimate it from one sample $\widehat{SE}[\hat{something}]$
- "omniscient": repeated sampling from a fake population (pedagogical simulation)
- resampling: draw repeated samples from the original data sample(s)
  - permutation test

Tools to learn about the Standard Error

- analytical: derive SE analytically, and estimate it from one sample $\widehat{SE}[\hat{something}]$
- "omniscient": repeated sampling from a fake population (pedagogical simulation)
- resampling: draw repeated samples from the original data sample(s)
    - permutation test
    - bootstrap: approxiamte sampling distribution by bootstrapping from one sample

Bootstrapping big picture

Lots of samples are kind of like the population



The Bootstrap

We see a single sample that is a draw from a population:

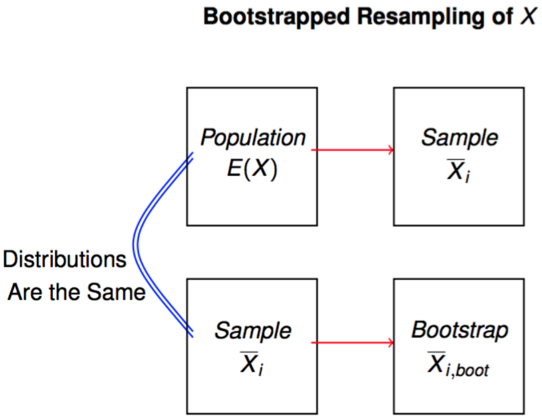
- There's a true mean loan amount; we only observe one sample

Since we cannot resample from the population, we resample from the sample!

Idea: Within a loop, generate a bootstrapped sample:

1. Sample from $\{1, 2, \ldots, N\}$ with replacement
2. Re-calculate the quantity of interest on each bootstrapped sample
3. Resampling from the sample *approximates* sampling again from the full population (giving us a sense of the sampling distribution)

Bootstrap: Intuition



Simple Example with Sample Means

Let $X_i = \{3, 7, 9, 11, 150\}$

Bootstrapped Samples:

						$\bar{X}_{\text{boot}}$
$\text{X}_{\text{boot},1}$	3	3	9	11	3	5.8
$\text{X}_{\text{boot},1}$	7	150	11	7	11	37.2
$\text{X}_{\text{boot},1}$	11	9	9	7	3	7.8
$\vdots$						

Bootstrapped Standard Error

- Bootstrapped Standard Error
  sd($\bar{X}_{\mathrm{boot}}$)
- Bootstrapped Confidence Interval:
  Take the 2.5% and 97.5% quantiles of $\bar{X}_{\mathrm{boot}}$

Questions?

Happy Fall Break!