Precept 3: Random Samples

Soc 400: Applied Social Statistics

Ziyao Tian[1]

Princeton University

September 27, 2018

[1]This set of slides draws on material from former preceptors Shay O'Brien, Simone Zhang, Matt Blackwell, Justin Grimmer and Jens Hainmueller.

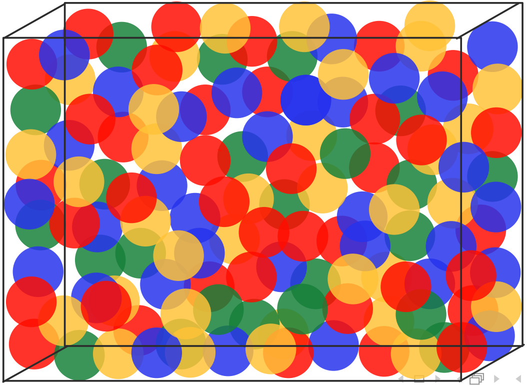
Today's Tasks

- Review material presented in lecture
  - sampling
  - estimators (and their properties)
  - CLT
  - confidence intervals
- Cover computational examples
  - rnorm(), pnorm(), qnorm()
  - drawing random samples
  - generating CIs
- All of this will help you on the problem set!

The Big Picture

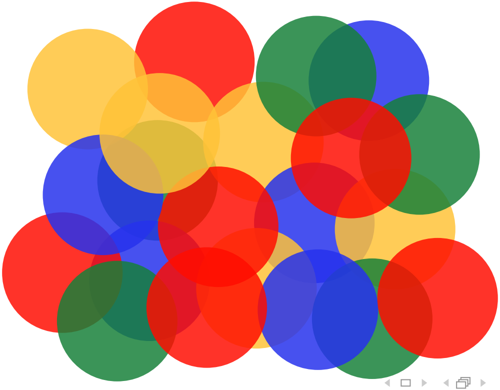
In studying the world, we usually run into the following challenge:

- There's some quantity of interest we want to know about a population, the **estimand**, which we consider to have a "true" value
- *e.g. What percentage of balls in this ball pit are red?*



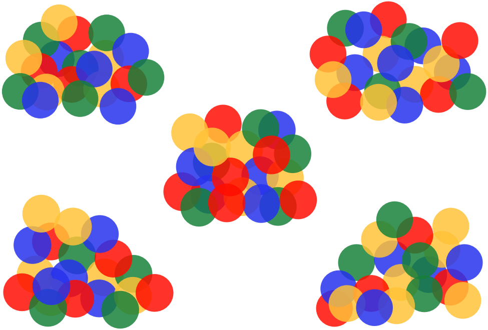
The Big Picture

- Ideally, we'd like to collect information on every member of the population. But usually, that's not possible. Instead we collect data on a random sample drawn from the population.
- *e.g. Randomly pull out a bunch of balls and count how many are red*

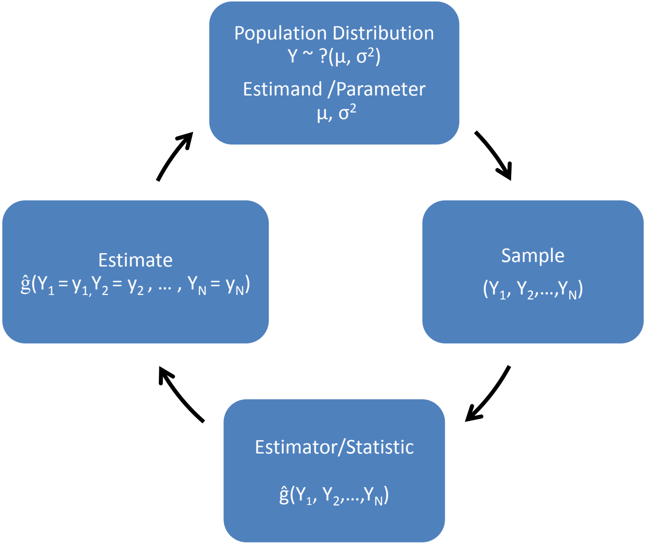


The Big Picture

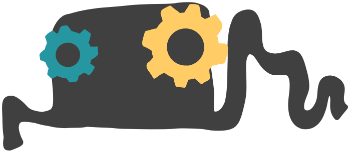
- This week is about understanding how to infer the "true" population-level distribution from the data we do have in a sample.
- *e.g. Calculate the percentage of red balls in the sample and extrapolate from that information to a lot of hypothetical samples*



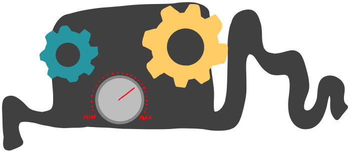
An Overview


Population Distribution
Y ~ ?(μ, σ²)
Estimand /Parameter
μ, σ²
Sample
(Y₁, Y₂,…,Y_N)
Estimator/Statistic
ĝ(Y₁, Y₂,…,Y_N)
Estimate
ĝ(Y₁ = y₁, Y₂ = y₂ , … , Y_N = y_N)

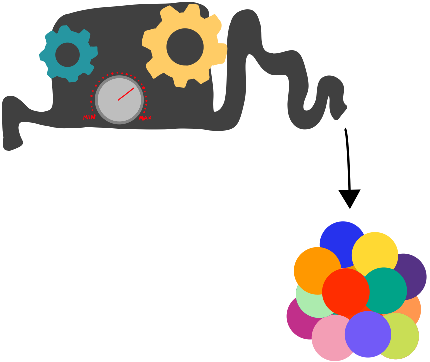
An Overview



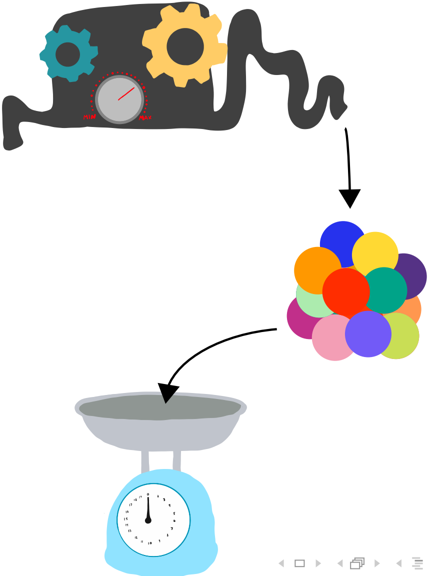
An Overview



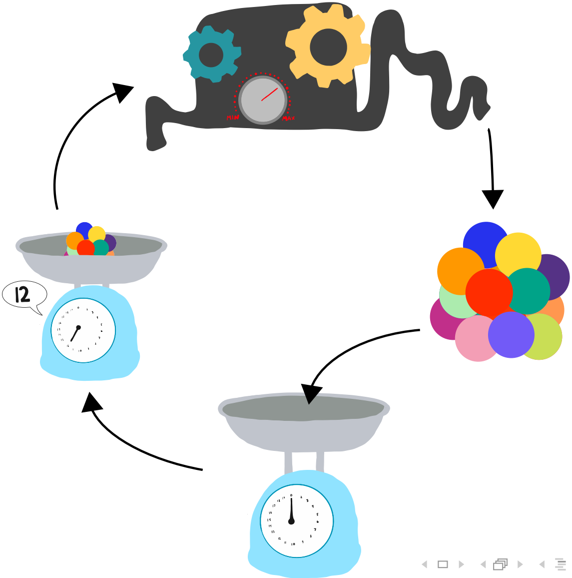
An Overview



An Overview



An Overview



Estimands, Estimators, and Estimates

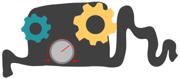
Estimands, Estimators, and Estimates

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**.

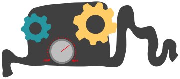
- **Estimands** are the parameters that we aim to estimate. Often written with greek letters (e.g. μ, θ, population mean) : $\frac{1}{N}\sum_{i=1}^{N} y_i$



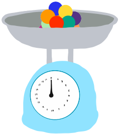
Estimands, Estimators, and Estimates

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) : $\frac{1}{N}\sum_{i=1}^{N} y_i$



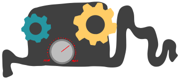
- **Estimators** are functions of sample data (i.e. **statistics**) which we use to learn about the estimands. Often denoted with a "hat" (e.g. $\hat{\mu}, \hat{\theta}$)



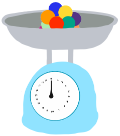
Estimands, Estimators, and Estimates

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

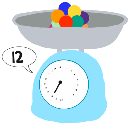
- **Estimands** are the parameters that we aim to estimate. Often written with greek letters (e.g. μ, θ, population mean) : $\frac{1}{N}\sum_{i=1}^{N} y_i$



- **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): $\frac{1}{n}\sum_{i=1}^{n} y_i$



Clarifying Notation and Terms You'll Encounter

- Estimand / Population Parameter (Theoretical)
  - Population mean: $\mu = E[X] = \frac{1}{N}\sum_{i=1}^{N} X_i$
  - Population variance: $\sigma^2 = E[(X - E(X))^2] = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$
- Estimator (Links Data to Estimand)
  - Estimator for population mean: $\hat{\mu}$
  - Estimator for population variance: $\hat{\sigma}^2$
- Estimate (Calculated from a Given Sample), e.g.
  - Sample mean: $\overline{X}_n = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Sample variance: $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X}_n)^2$

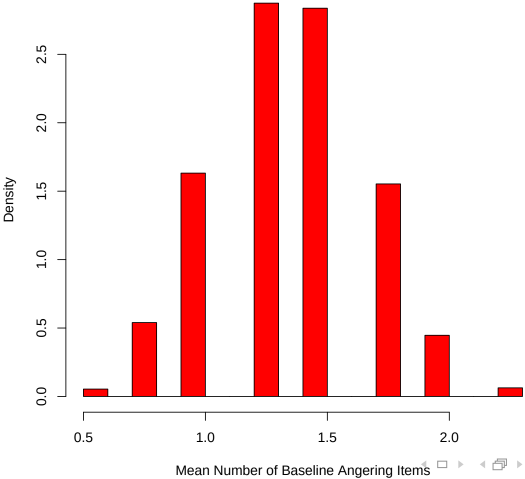
Sampling Distribution

Consider using the sample mean as an estimator for the "true" mean: $\hat{\mu} = \overline{X}_n$:

- Usually we only ever observe one sample of size n - so we get one value of $\overline{X}_n$
- But consider the hypothetical case that we got 10,000 random samples of size n. By random chance, the samples may look different from each other. Each sample would have its own $\overline{X}_n$
- The sampling distribution of $\overline{X}_n$ gives the probability density of the possible values of $\overline{X}_n$

Sampling Distribution of the Sample Mean

Example:



Other estimators (e.g. sample variance, or proportions) also have sampling distributions.
We can describe sampling distributions in terms of their center (i.e. mean) and spread (i.e. standard error).

The Central Limit Theorem

The Central Limit Theorem tells us something cool about sample means ($\overline{X}_n$).
From the lecture slides, as n increases, the sampling distribution of $\overline{X}_n$ becomes more bell-shaped. This is the basic implication of the **Central Limit Theorem**:

If $X_1, \ldots, X_n \sim_{i.i.d.} ?(\mu, \sigma^2)$ and n is large, then

$$\overline{X}_n \sim_{approx} N(\mu, \frac{\sigma^2}{n})$$

so

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

Summary of Properties of Estimators

Concept	Criteria	Intuition
Unbiasedness	$E[\hat{\mu}] = \mu$	Right on average
Efficiency	$V[\hat{\mu}_1] < V[\hat{\mu}_2]$	Low variance
Consistency	$\hat{\mu}_n \xrightarrow{p} \mu$	Converge to estimand as $n \to \infty$
Asymptotic Normality	$\hat{\mu}_n \overset{\text{approx.}}{\sim} N(\mu, \frac{\sigma^2}{n})$	Approximately normal in large n

Confidence Intervals

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(-z \leq \frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z \right) = (1 - \alpha)$$

Confidence Intervals

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(-z \leq \frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z \right) = (1 - \alpha)$$

We call z "critical value", and denote such z as $z_{\alpha/2}$.

Confidence Intervals

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(-z \leq \frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z \right) = (1 - \alpha)$$

We call z "critical value", and denote such z as $z_{\alpha/2}$.
When $X \sim N(0,1)$
$P(X \leq z_{\alpha/2}) = 1 - \alpha/2$
$P(X \leq z_{\alpha}) = 1 - \alpha$

Confidence Intervals

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(-z_{\alpha/2} \leq \frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z_{\alpha/2} \right) = (1-\alpha)$$

Confidence Intervals

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(-z_{\alpha/2} \leq \frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z_{\alpha/2} \right) = (1 - \alpha)$$

We can rewrite this into

$$P\left(\hat{\mu} - z_{\alpha/2} * \hat{SE}[\hat{\mu}] \leq \mu \leq \hat{\mu} + z_{\alpha/2} * \hat{SE}[\hat{\mu}] \right) = (1 - \alpha)$$

Confidence Intervals

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(-z_{\alpha/2} \leq \frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z_{\alpha/2}\right) = (1-\alpha)$$

We can rewrite this into

$$P\left(\hat{\mu} - z_{\alpha/2} * \hat{SE}[\hat{\mu}] \leq \mu \leq \hat{\mu} + z_{\alpha/2} * \hat{SE}[\hat{\mu}]\right) = (1-\alpha)$$

α: significance level. More on this in future lectures. So far, all we need to know is: confidence level + $\alpha = 1$

Confidence Intervals

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\Bigg(- z_{\alpha/2} \leq \frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z_{\alpha/2} \Bigg) = (1-\alpha)$$

We can rewrite this into

$$P\Bigg(\hat{\mu} - z_{\alpha/2} * \hat{SE}[\hat{\mu}] \leq \mu \leq \hat{\mu} + z_{\alpha/2} * \hat{SE}[\hat{\mu}] \Bigg) = (1-\alpha)$$

α: significance level. More on this in future lectures. So far, all we need to know is: confidence level $+ \alpha = 1$

Confidence level?

Confidence Intervals

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(-z_{\alpha/2} \leq \frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z_{\alpha/2} \right) = (1 - \alpha)$$

We can rewrite this into

$$P\left(\hat{\mu} - z_{\alpha/2} * \hat{SE}[\hat{\mu}] \leq \mu \leq \hat{\mu} + z_{\alpha/2} * \hat{SE}[\hat{\mu}] \right) = (1 - \alpha)$$

α: significance level. More on this in future lectures. So far, all we need to know is: confidence level $+ \alpha = 1$
Confidence level? $100(1 - \alpha)\%$

Confidence Intervals

What is the formula for two-sided confidence intervals with confidence level of 100(1-α)%?

$$[\hat{\mu} - z_{\alpha/2} * \hat{SE}[\hat{\mu}], \hat{\mu} + z_{\alpha/2} * \hat{SE}[\hat{\mu}]]$$

Or

$$[\overline{X}_n - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{s}{\sqrt{n}}]$$

Confidence Intervals

What is the formula for two-sided confidence intervals with confidence level of 100(1-α)%?

$$[\hat{\mu} - z_{\alpha/2} * \hat{SE}[\hat{\mu}], \hat{\mu} + z_{\alpha/2} * \hat{SE}[\hat{\mu}]]$$

Or

$$[\overline{X}_n - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{s}{\sqrt{n}}]$$

How do we find $z_{\alpha/2}$? qnorm() : $F^{-1}(p)$
Returns z value at which CDF of Standard Normal equals p

Confidence Intervals

What is the formula for two-sided confidence intervals with confidence level of 100(1-α)%?

$$[\hat{\mu} - z_{\alpha/2} * \hat{SE}[\hat{\mu}], \hat{\mu} + z_{\alpha/2} * \hat{SE}[\hat{\mu}]]$$

Or

$$[\overline{X}_n - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{s}{\sqrt{n}}]$$

How do we find $z_{\alpha/2}$? qnorm() : $F^{-1}(p)$
Returns z value at which CDF of Standard Normal equals p
What is the width of the confidence interval? $2 * z_{\alpha/2}\frac{s}{\sqrt{n}}$

We can use our analytic samples to find a confidence interval

Our estimate

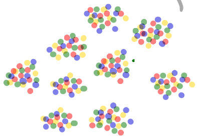
$$CI(\alpha) = [r - z_{\alpha/2} * SE, r + z_{\alpha/2} * SE]$$

Alpha

Standard error of our estimate

Critical value

$\alpha/2$ because we're looking for a two-sided interval



To use the confidence interval formula, we need to find:

1. The distribution
2. Confidence level
    - Alpha
3. Sidedness
4. Critical value(s)
5. Standard error of our estimate

```
##Calculating our critical value
cv <- qnorm(.975)
cv

## [1] 1.959964
```

```
##Finding the standard error of our estimate
se <- sqrt(red.sample*(1-red.sample)/n.samp)
se

## [1] 0.01966499
```

for a proportion, the formula is: $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Calculating the confidence interval

$$CI(\alpha) = [r - z_{\alpha/2} * SE, r + z_{\alpha/2} * SE]$$

```
##Finding and printing the confidence interval
c(red.sample - cv*se,
  red.sample + cv*se)
```

```
## [1] 0.2234573 0.3005427
```

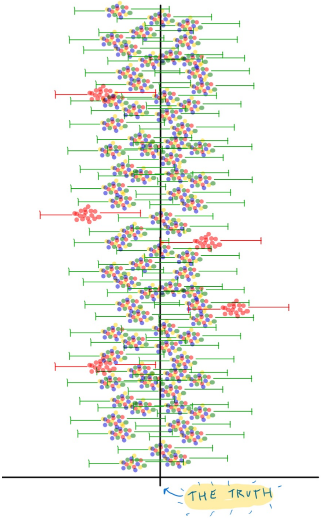
Our results

26.2% red with a 95 percent confidence interval of **[22.3, 30.1]**

We hope our sample is in the 95%

samples whose confidence intervals contain the truth (95%)

samples whose confidence intervals do not contain the truth (5%)



Fulton Data

- Election data from Fulton County, Georgia, aggregated to the precinct level

Table: Fulton Election Data

Variable	Description
precint	precint id
turnout	voter turnout rate
black	percent Black
sex	percent Female
age	mean age
dem	turnout in democratic primary
rep	turnout in republican primary
urban	is the precinct in Atlanta
school	school polling location

Questions?

Optional: One-sided Confidence Intervals

What about one-sided confidence intervals?

Optional: One-sided Confidence Intervals

What about one-sided confidence intervals?
An 100(1-α)% upper (one-sided) confidence bound

$$\overline{X}_n + z_\alpha \frac{s}{\sqrt{n}}$$

An 100(1-α)% lower (one-sided) confidence bound

$$\overline{X}_n - z_\alpha \frac{s}{\sqrt{n}}$$

Optional: One-sided Confidence Intervals

What about one-sided confidence intervals?
An 100(1-α)% upper (one-sided) confidence bound

$$\overline{X}_n + z_\alpha \frac{s}{\sqrt{n}}$$

An 100(1-α)% lower (one-sided) confidence bound

$$\overline{X}_n - z_\alpha \frac{s}{\sqrt{n}}$$

Optional: One-sided Confidence Intervals

Why? Take lower one-sided CI for example.
An 100(1-α)% lower (one-sided) confidence bound

$$\overline{X}_n - z_\alpha \frac{s}{\sqrt{n}}$$

Optional: One-sided Confidence Intervals

Why? Take lower one-sided CI for example.
An 100(1-α)% lower (one-sided) confidence bound

$$\overline{X}_n - z_\alpha \frac{s}{\sqrt{n}}$$

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(\frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z_\alpha\right) = (1 - \alpha)$$

Optional: One-sided Confidence Intervals

Why? Take lower one-sided CI for example.
An 100(1-α)% lower (one-sided) confidence bound

$$\overline{X}_n - z_\alpha \frac{s}{\sqrt{n}}$$

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(\frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z_\alpha\right) = (1-\alpha)$$

We can rewrite this into

$$P\left(\mu \geq \overline{X}_n - z_\alpha \frac{s}{\sqrt{n}}\right) = (1-\alpha)$$

Optional: One-sided Confidence Intervals

Why? Take lower one-sided CI for example.
An 100(1-α)% lower (one-sided) confidence bound

$$\overline{X}_n - z_\alpha \frac{s}{\sqrt{n}}$$

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(\frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z_\alpha\right) = (1-\alpha)$$

We can rewrite this into

$$P\left(\mu \geq \overline{X}_n - z_\alpha \frac{s}{\sqrt{n}}\right) = (1-\alpha)$$

Lower confidence bound $\mu \geq \overline{X}_n - z_\alpha \frac{s}{\sqrt{n}}$
Or in the form of an "interval" $(-\infty, \overline{X}_n - z_\alpha \frac{s}{\sqrt{n}})$