Quick Review of Last Week's Computational Concepts

- Objects
- Functions manipulating objects
- Data frames
- Vectors
- Numeric, Character, Boolean, and Factor objects...

Extra Credit

Goal: change all of the "EE" entries in the ponds data to "E"
So, after last class, you may try to do this:

```
##### Extra Credit

# this won't work
ponds$site[which(ponds$site == "EE")] <- "E"

## Warning:  invalid factor level, NAs generated
```

What is a factor?

```
ponds$site[1:20]

##  [1] A  B  C  D  EE A  B  C  D  EE A  B  C  D  EE A  B  C
## [19] D  EE
## Levels: A B C D EE
```

What is a factor?

```
ponds$site[1:20]

##  [1] A  B  C  D  EE A  B  C  D  EE A  B  C  D  EE A  B  C
## [19] D  EE
## Levels: A B C D EE
```

- A factor is made up of text strings
- A factor has levels

Numerics, Characters, and Factors

```r
# a numeric vector
c(1, 2, 3)

## [1] 1 2 3
```

Numerics, Characters, and Factors

```
# a numeric vector
c(1, 2, 3)

## [1] 1 2 3
```

```
# a character vector
c("1", "2", "3")

## [1] "1" "2" "3"
```

Numerics, Characters, and Factors

```
# a numeric vector
c(1, 2, 3)

## [1] 1 2 3
```

```
# a character vector
c("1", "2", "3")

## [1] "1" "2" "3"
```

```
# a character vector -> factor
factor(c("1", "2", "3"))

## [1] 1 2 3
## Levels: 1 2 3
```

From Factors to Characters and Back

```
# instead, you need to do this...
ponds$site <- as.character(ponds$site)

ponds$site[which(ponds$site == "EE")] <- "E"

ponds$site <- factor(ponds$site)
```

Or Just Change Factor Levels...

```
# Or, just change the levels
levels(ponds$site)

## [1] "A"  "B"  "C"  "D"  "EE"


levels(ponds$site) <- c("A", "B", "C", "D", "E")

levels(ponds$site)

## [1] "A" "B" "C" "D" "E"
```

A More Foolproof Level Change...

```
# alternative approach
levels(ponds$site) <- c(levels(ponds$site)[1:4], "E")

ponds$site[1:10]

##  [1] A B C D E A B C D E
## Levels: A B C D E
```

You could use which

Exercise

- Create a factor vector of the letters A thorugh D that repeats 10 times (use rep)
- Do the same thing, but with the strings A1, B1, ...D1
- Merge these two into a single vector.

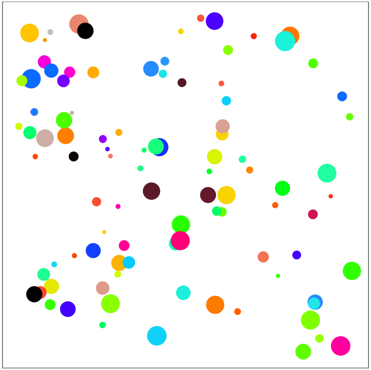
Exercise

```
# alternative approach
v1 <- factor(rep(c("A", "B", "C", "D"), 10))

v2 <- factor(rep(c("A1", "B1", "C1", "D1"), 10))

v3 <- factor(c(as.character(v1), as.character(v2)))

v3[1:20]

##  [1] A B C D A B C D A B C D A B C D A B C D
## Levels: A A1 B B1 C C1 D D1
```

Questions?

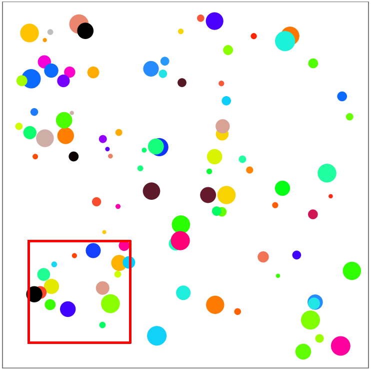
Sampling Populations

What is a population?



Population = All Individuals

What is a sample?

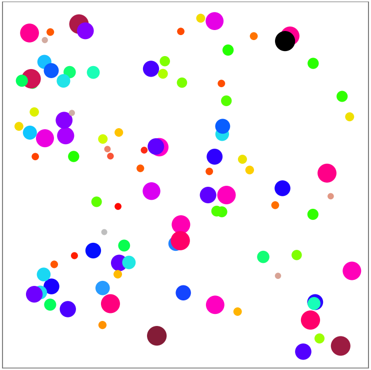


A **sample** of individuals in a randomly distributed population.

How can sampling a population go awry?

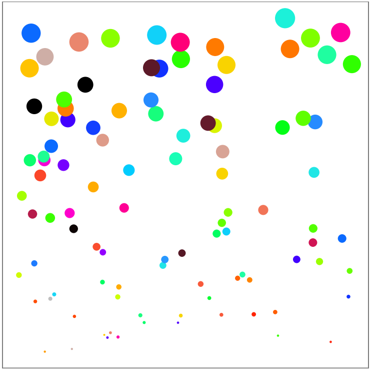
- Sample is not **representative**
- Replicates do not have **equal chance** of being sampled
- Replicates are not is not **independent**

Bias from Unequal Representation

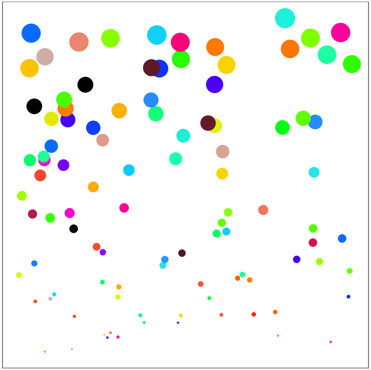


If you only chose one color, you would only get one range of sizes.

Bias from Unequal Change of Sampling

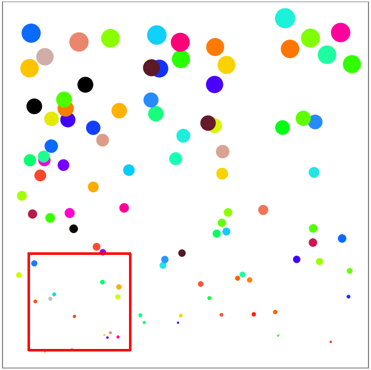


Bias from Unequal Change of Sampling



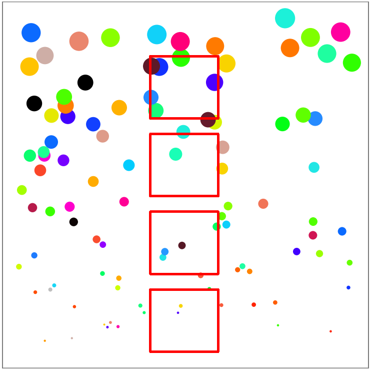
Spatial gradient in size

Bias from Unequal Change of Sampling



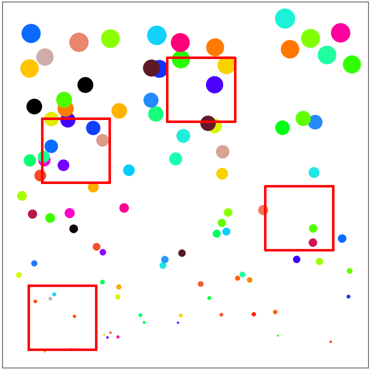
Oh, I'll just grab those individuals closest to me...

Solution: **Stratified** Sampling

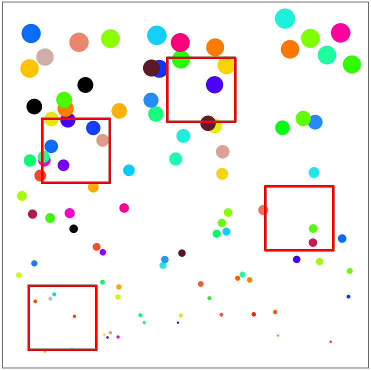


Sample over a known gradient, aka **cluster sampling**
Can incorporate multiple gradients

Solution: Random Sampling



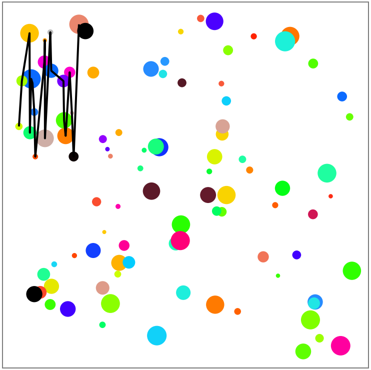
Solution: Random Sampling



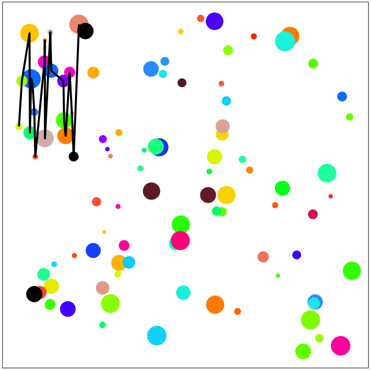
Two sampling schemes:

- **Random** - samples chosen using random numbers
- **Haphazard** - samples chosen without any system (careful!)

Non-Independence & Haphazard Sampling

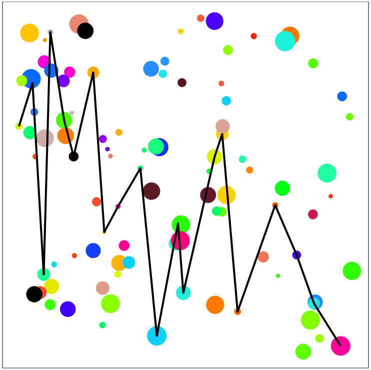


Non-Independence & Haphazard Sampling



What if there are interactions between individuals?

Solution: Chose Samples Randomly



Path chosen with random number generator

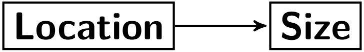
Deciding Sampling Design

What influences the measurement you are interested in?

$$\boxed{\textbf{Location}} \longrightarrow \boxed{\textbf{Size}}$$

Deciding Sampling Design

What influences the measurement you are interested in?



Causal Graph

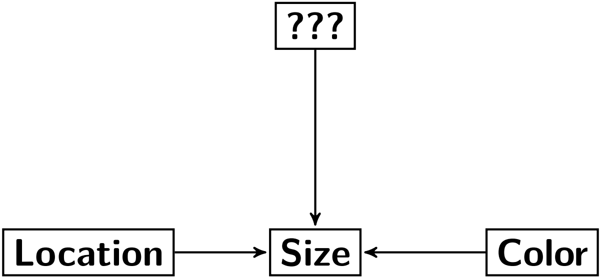
Stratified or Random?

Do you know all of the influences?

$$\boxed{\textbf{Location}} \longrightarrow \boxed{\textbf{Size}} \longleftarrow \boxed{\textbf{Color}}$$

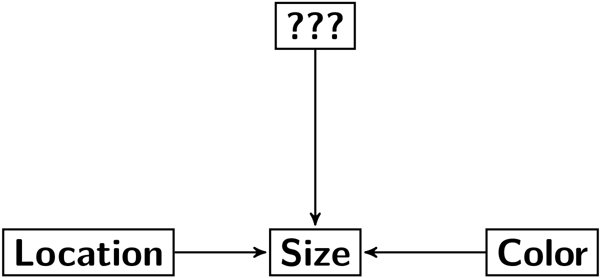
Stratified or Random?

Do you know all of the influences?



Stratified or Random?

Do you know all of the influences?



You can represent this as an equation:

Size = Color + Location + ???

Stratified or Random?

- How is your population defined?
- What is the scale of your inference?
- What might influence the inclusion of a replicate?
- How important are external factors you know about?
- How important are external factors you cannot assess?

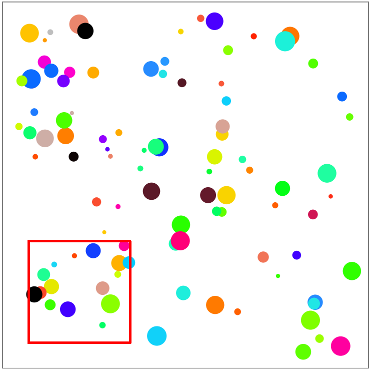
Exercise

Draw a causal graph of the influences on one thing you measure

How would you sample your population?

Describing a Sample

Sample Properties: Mean



What is the mean size of individuals in this population?

$$\bar{Y} = \frac{\sum y_i}{n}$$

Sample Properties: Mean

$$\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$\bar{Y}$ - The average value of a sample

x_i - The value of a measurement for a single individual

n - The number of individuals in a sample

Sample Properties: Mean

$$\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$\bar{Y}$ - The average value of a sample
x_i - The value of a measurement for a single individual
n - The number of individuals in a sample

μ - The average value of a population

Sample Properties: Mean

$$\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$\bar{Y}$ - The average value of a sample
x_i - The value of a measurement for a single individual
n - The number of individuals in a sample

μ - The average value of a population
(Greek = population, Latin = Sample)

R: Sample Size and Estimate Precision

As n increases, does your estimate get closer to the true mean?

1. Taking a mean

```
mean(c(1, 4, 5, 10, 15))

## [1] 7
```

R: Sample Size and Estimate Precision

As n increases, does your estimate get closer to the true mean?

2. Mean from a random population

```
mean(runif(n = 500, min = 0, max = 100))

## [1] 47.53
```

runif draws from a Uniform distribution

R: Sample Size and Estimate Precision

As n increases, does your estimate get closer to the true mean?

3. Sampling from a simulated population

```
set.seed(5000)
population <- runif(400, 0, 100)
mean(sample(population, size = 50))

## [1] 46.83
```

R: Sample Size and Estimate Precision

As n increases, does your estimate get closer to the true mean?

3. Sampling from a simulated population

```
set.seed(5000)
population <- runif(400, 0, 100)
mean(sample(population, size = 50))

## [1] 46.83
```

set.seed ensures that you get the same random number every time
sample draws a sample of a defined size from a vector

Exercise: Sample Size and Estimate Precision

As n increases, does your estimate get closer to the true mean?

1. Use runif (or rnorm, if you're feeling saucy) to simulate a population
2. How does the repeatability of the mean change as you change the sample size?

Exercise: Sample Size and Estimate Precision

As n increases, does your estimate get closer to the true mean?

```
set.seed(5000)
population <- runif(n = 400, min = 0, max = 100)
mean(sample(population, size = 3))

## [1] 64.52

mean(sample(population, size = 3))

## [1] 54.91
```

Exercise: Sample Size and Estimate Precision

As n increases, does your estimate get closer to the true mean?

```
mean(sample(population, size = 100))

## [1] 45.06

mean(sample(population, size = 100))

## [1] 45.96
```

Sample Properties: Variance

How variable was that population?

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

- **Sums of Squares** over n-1
- n-1 corrects for both sample size and sample bias
- σ^2 if describing the population
- Units in square of measurement...

Sample Properties: Standard Deviation

$$s = \sqrt{s^2}$$

- Units the same as the measurement
- If distribution is normal, 67% of data within 1 SD, 95% within 2
- σ if describing the population

Exercise: Sample Size and Estimated Sample Variation

1. Repeat the last exercise, but with the functions sd or var
2. Do you need as many samples for a precise estimate as for the mean?

Next time...