Association Studies

Calculations and Interpretations

Session 7

Genetics Podcasts of Note

- Freakonomics interview with 23andMe founder:
  - https://freakonomics.com/podcast/23andme/

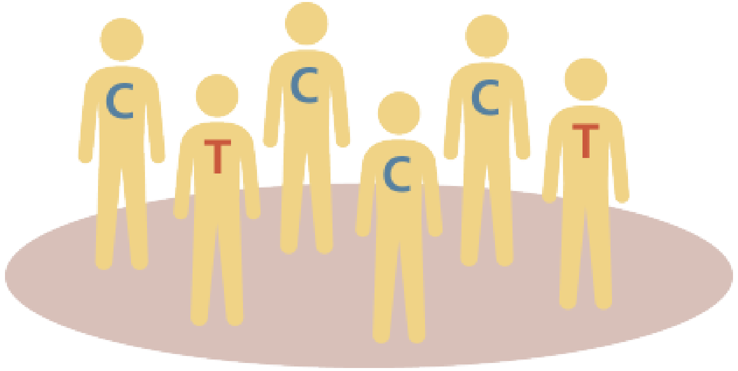
- Future of Everything interview with Carlos Bustamonte:
  - https://scopeblog.stanford.edu/2019/04/18/the-future-of-genomics-a-podcast-featuring-stanford-geneticists/

Learning objectives

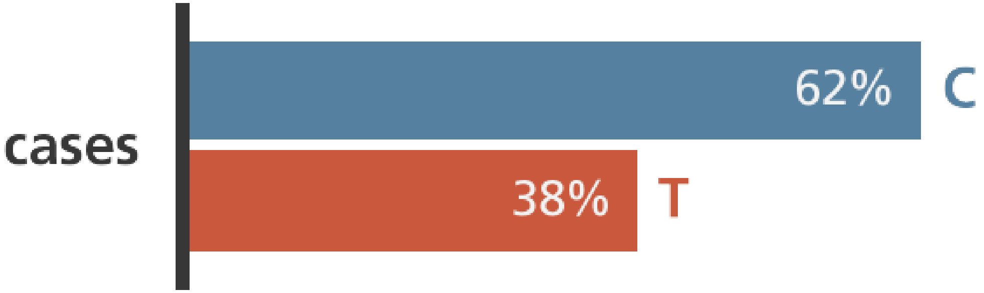
- Calculate and interpret odds ratios in case/control genetic association studies.
- Interpret quantitative trait association studies.

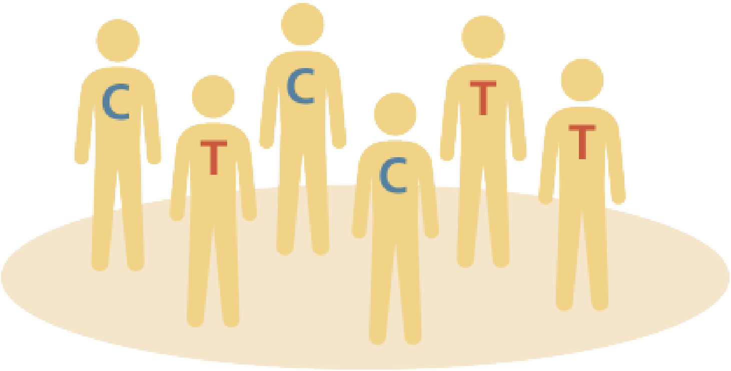
Association studies

- Determine if a particular genetic feature (exposure) co-occurs with a trait (disease) more often than would be expected by chance.
- Binary: Calculate ‘odds’ of an outcome occurring.
  - Framed as an ‘odds ratio’, the odds of an outcome with an exposure (genotype) in relation to the odds of an outcome without the exposure (reference genotype).
- Quantitative: calculate change in an outcome for every unit increase of an exposure.

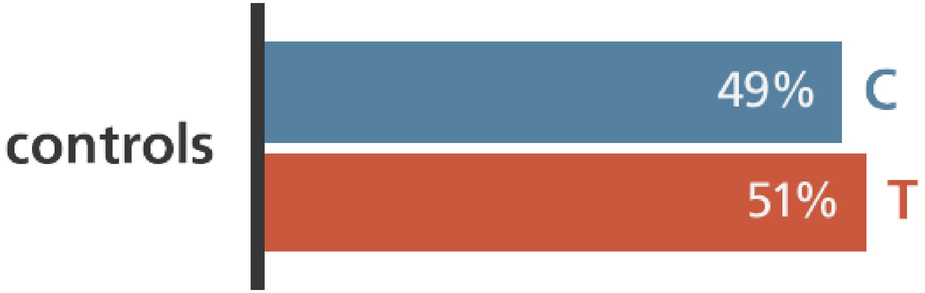


cases (n=1,000)
people with heart disease





controls (n=1,000)
people without heart disease



*Still probabilistic – with a C allele, you are MORE LIKELY to develop heart disease, but it is not guaranteed that you will, or that you won't if you have the T allele.

“Odds”

the chances or likelihood of something happening

“Odds Ratio”

the likelihood of something happening in one group in relation to the likelihood of something happening in another group

Odds ratio

The odds ratio is our measure of association for a case-control study. It tells us whether and how much an exposure increases the likelihood of our outcome of interest. We need to look at two things:

The estimate -- the odds ratio itself. How big in the connection between an exposure and an outcome? Are those with an exposure more likely to have the outcome?

The p-value -- how certain are we that the odds ratio didn’t just happen by chance?

Association testing in case-control studies

		Disease status		
		Cases	Controls	Total
Genotype	AA/AT	a	b	a+b
	TT	c	d	c+d
Total		a+c	b+d	

measure of events out of all possible events (Ratio) vs ratio of events to non-events (Odds)

$$RR = \frac{\text{Risk of event in the Treatment group}}{\text{Risk of event in the Control group}} = \frac{\mathrm{a}/(\mathrm{a}+\mathrm{b})}{c/(c+d)}$$

$$OR \;= \frac{\text{Odds of event in Treatment group}}{\text{Odds of event in Control group}} = \frac{\mathrm{a}/\mathrm{b}}{c/d} \quad :$$

If an outcome occurs 10 out of 100 times, the risk is 10%
But the odds is 10/90 = 11.1%

Association testing in case-control studies

		Disease status		
		Cases	Controls	Total
Genotype	AA/AT	a	b	a+b
	TT	c	d	c+d
Total		a+c	b+d	

Calculate Odds Ratio (OR) as the odds of being a case among genotype AA/AT divided by the odds of being a case among genotype TT.

$$\frac{{}^{a}/_{b}}{{}^{c}/_{d}} = \frac{ad}{bc}$$

$$\text{OR} = \frac{ad}{bc}$$

Association testing in case-control studies

		Disease status		
		Cases	Controls	Total
Genotype	AA/AT	a	b	a+b
	TT	c	d	c+d
Total		a+c	b+d	

H_0: OR = 1   (no association)

OR > 1   indicates increased odds

OR < 1   indicates decreased odds (protective)

Confidence intervals for odds ratios

		Disease status	
		Cases	Controls
Genotype	AA/AT	a	b
	TT	c	d

OR= $\frac{a/b}{c/d} = \frac{ad}{bc}$

s.e(log(OR))= $\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$

Confidence interval: $e^{\log(OR) \pm z_{\alpha/2} \times s.e(\log(OR))}$

Lower limit of 95% confidence interval: $e^{\log(OR) - 1.96 \times s.e}$

Upper limit of 95% confidence interval: $e^{\log(OR) + 1.96 \times s.e}$

Odds ratio - rs2233385 and neuropsychiatric adverse events

The odds of having a neuropsychiatric reaction to tamiflu for people with two variant rs2233385 alleles are 30x the odds of having a neuropsychiatric reaction for people without any copies of that variant.

Formula:
The odds of the outcome among people with the exposure are odds ratio times the odds of the outcome among people without the exposure.

Zoom breakout Q1:
Calculate– odds ratio and 95% confidence interval

	Cases	Controls	Total
TT+TC	158	392	550
CC	20	86	106
Total	178	478	1656

$$\text{OR}= \frac{ad}{bc}$$

$$\text{s.e}(\log(\text{OR}))=\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

Why do we even use odds and odds ratios???

The odds ratio allows us to calculate the associations between an exposure and an outcome without needing the frequency of the exposure in the general population

(very useful to rare exposures, such as rare diseases).

(we'd have to sample A LOT of people to get a true population picture and even pick up one or two cases of the disease)

The log(odds) allows us to transform this weird variable into a linear form, which is easier for us to fit to models, adjust for covariates, and interpret the output.

Often use logistic regression for case-control analyses

Allows you to adjust for relevant factors

- Population stratification, age, sex, matching variables etc

$\ln\left(\frac{p}{1-p}\right)$= α + β_1**g** + β_2x$_1$ + ….+β_{k+1}x$_k$ (g is genotype, x$_1$,…x$_k$ are covariates)

Coefficients are estimated using maximum likelihood estimation (MLE)

- $\ln\left(\frac{p}{1-p}\right)$ = log odds of an outcome
- Test H$_0$: β_1 =0 (likelihood ratio test, wald test, score test)
- The odds ratio is OR=e^{β_1}
- β_1 = SNP effect (log(OR)) ➔ e $^{\beta_1}$ = OR

Logistic regression model and interpretation

Logistic regression works on likelihoods. The likelihood of an outcome given a change in exposure. The dependent variable is the log of the odds. The log(odds) lets us do our analyses easier.

$$\text{logit(p)} = \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 \cdot x_{i2} \cdot$$

This part of the equation is what we are testing: how much does the likelihood of an outcome change when we change our exposure “x”

Logistic regression model and interpretation

$$\text{logit(p)} = \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 \cdot x_{i2} \cdot$$

For additive model, “x” is 0 when the allele of interest is not present, 1 when there is one copy, and 2 when there are two copies of the allele
*depends on suspected mode of inheritance

Logistic regression model and interpretation

$$\text{logit(p)} = \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 \cdot x_{i2} \cdot$$

Log (odds of reaction when one copy of the allele is present)= B_o +B_1*1

Log (odds of reaction outcome when zero copies of the allele are present)= B_o +B_1*0 = B_o

Log (odds of reaction with one copy of the allele=Log (odds of reaction outcome with zero copies of the allele)+B_1

Log(odds of reaction with one copy of the allele) - log(odds of reaction outcome with zero copies of the allele)=B_1

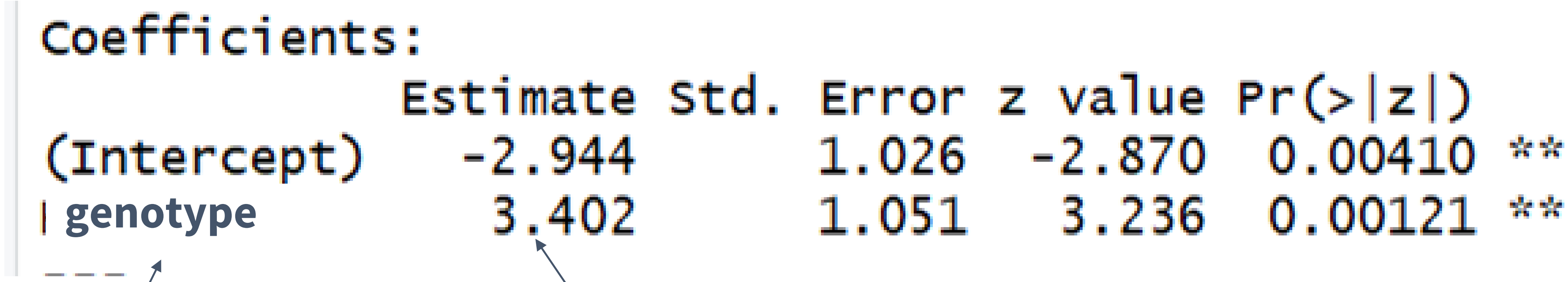
Math principle: log(x) - log(y) = log(x/y)

Log(odds of reaction with one copy of the allele/odds of reaction outcome with zero copies of the allele)= B^1

Log(odds ratio of outcome with vs without allele) = B^1

Then, we can take the “e” exponent of each side to get our Odds Ratio = e^B

Logistic regression intercept in R

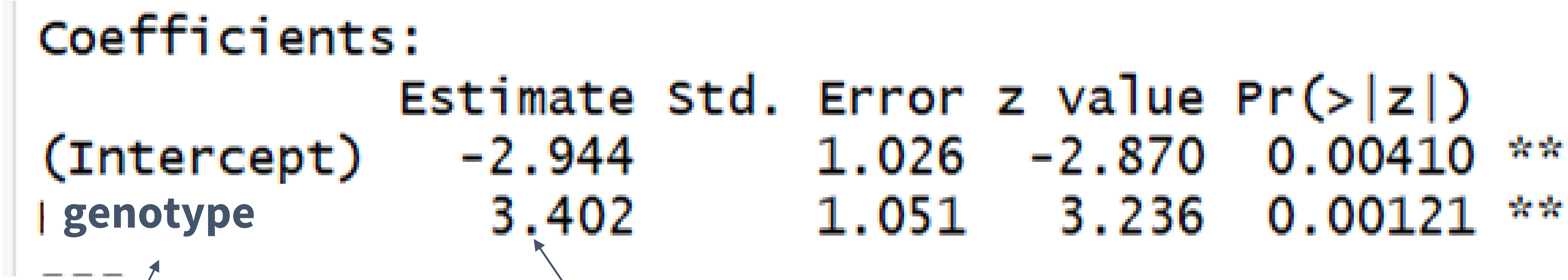


```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.944      1.026  -2.870  0.00410 **
genotype       3.402      1.051   3.236  0.00121 **
---
```

(x) This genotype variable is coded 0/1/2. We want to look at whether having this allele changes the odds of an outcome.

This estimate gives us the estimated difference in log odds between having a genotype that is different by one copy of the variant.

Logistic regression intercept



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.944      1.026  -2.870  0.00410 **
 genotype       3.402      1.051   3.236  0.00121 **
---
```

(x) This genotype variable is coded 0/1/2. We want to look at whether having this allele changes the odds of an outcome.

This estimate gives us the estimated difference in log odds between having a genotype that is different by one copy of the variant.

To find the odds ratio of outcome between those with 1 copy of an allele and zero copies of the allele, we calculate $e^{3.402}$

Resources to learn R coding for genetic analysis

- Packet from Sara on the course website
- Dr. Tim Thornton’s Course in SISG: R exercises http://faculty.washington.edu/tathornt/SISG2020.html
- edX: https://www.edx.org/learn/computer-programming
- Epidemiologist R Handbook: https://epirhandbook.com/index.html

Quantitative outcome genetic association

- Linear regression (instead of logistic)
- Additive coding of SNP (0,1,2) most common

$$Y = \alpha + \beta * SNP + X$$

- β = SNP effect (for every SNP, unit increase in outcome)
    - We do not need to use the exponent in quantitative outcomes
- SNP = covariate coded (0,1,2)
- X = additional covariates (e.g. sex, study, age, PCs from population stratification)

Importance of setting your reference allele

Odds ratio when AA is reference: ⅔ / ⅓ = ⅔ *3 = 2

The odds of the outcome are 2x more likely among those with CC genotype compared to among those with the AA genotype.

Odds ratio when CC is reference. ⅓ / ⅔ = ⅓ * $^{3}/_{2}$ = 0.5

The odds of the outcome are ½ as likely among those with AA genotype compared to among those with the CC genotype.

These are the saying the same thing! But the language matters.

Always know and be purposeful on your reference

In epidemiology, the reference group always matters.

Exposure (gene allele reference)

Outcome (some outcomes have no “direction”) brown vs black hair

Population (other factors are always involved, i.e. age, diet, access to care).

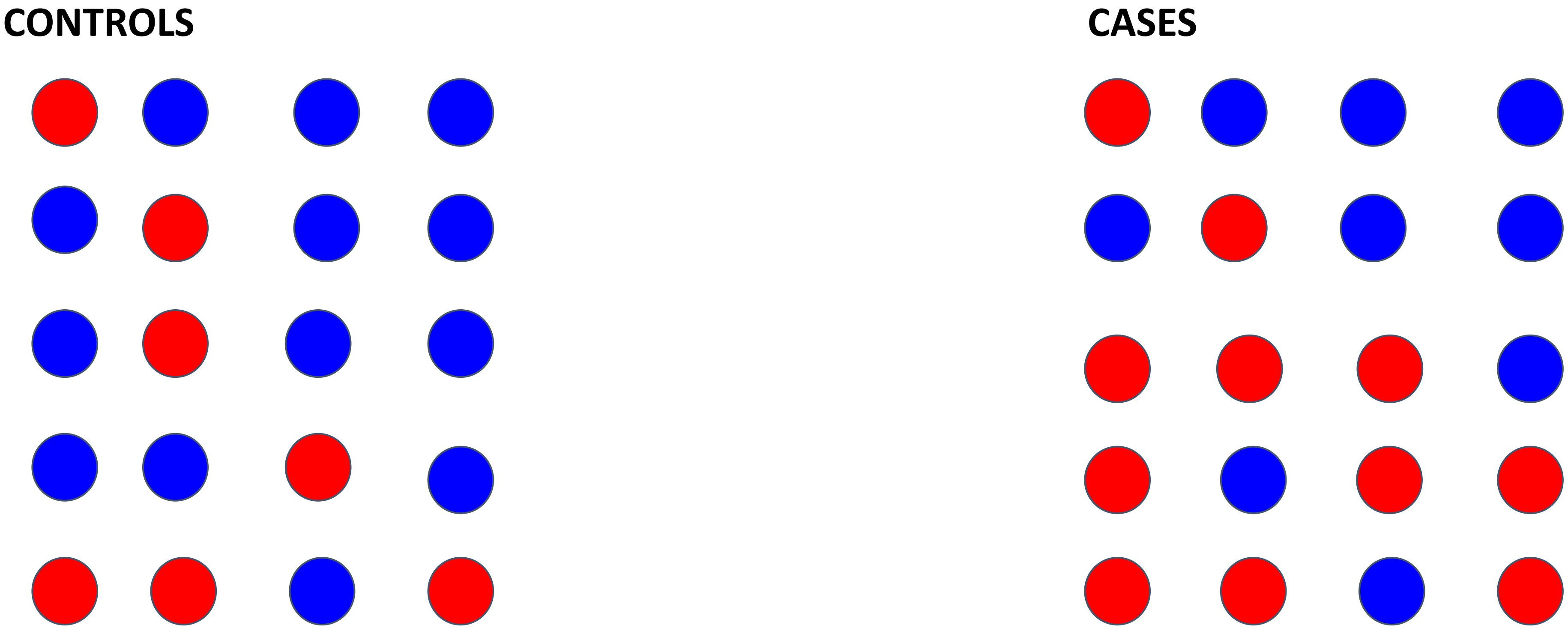
Zoom breakout #2 and #3

- Interpret the regression output from association studies (one binary and one quantitative trait)

Summary

- Odds ratios give the odds of an outcome in relation to a reference.
- Linear and logistic regression allow adjustment for other factors.

Case-control study - is red associated?

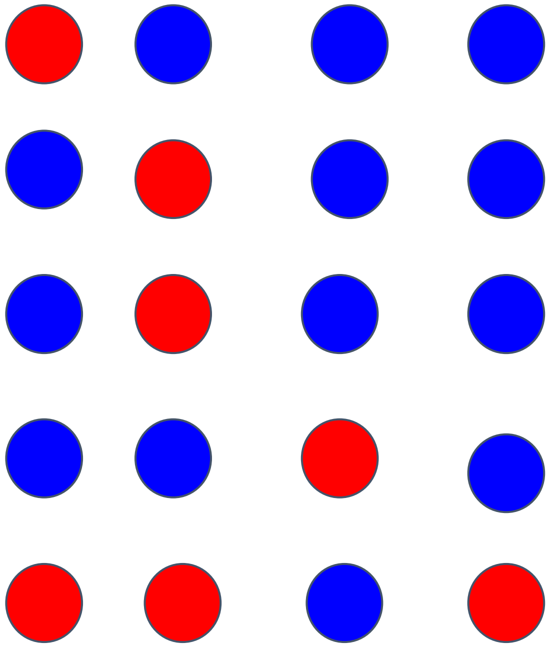


CONTROLS

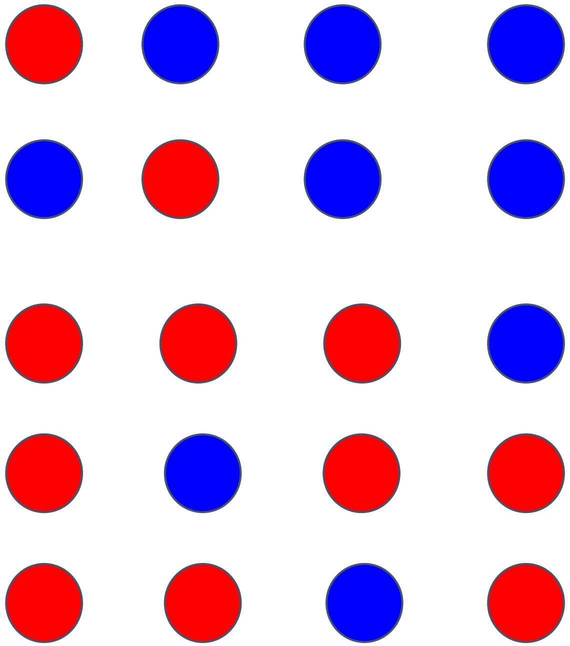
CASES

Case-control study - is red associated?

CONTROLS



CASES

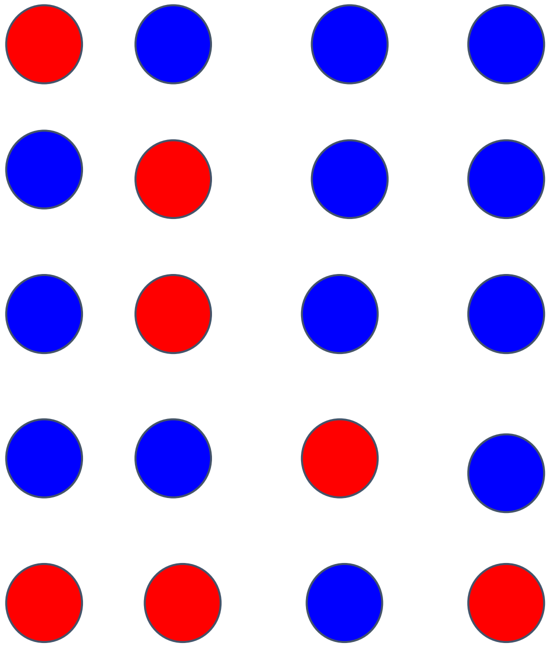


Zoom poll

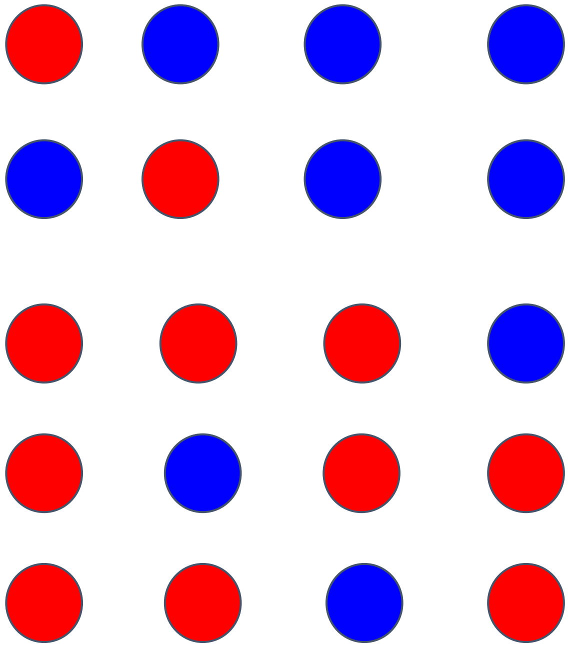
Calculate odds ratio of being red between cases and controls

Case-control study - is red associated?

CONTROLS



CASES



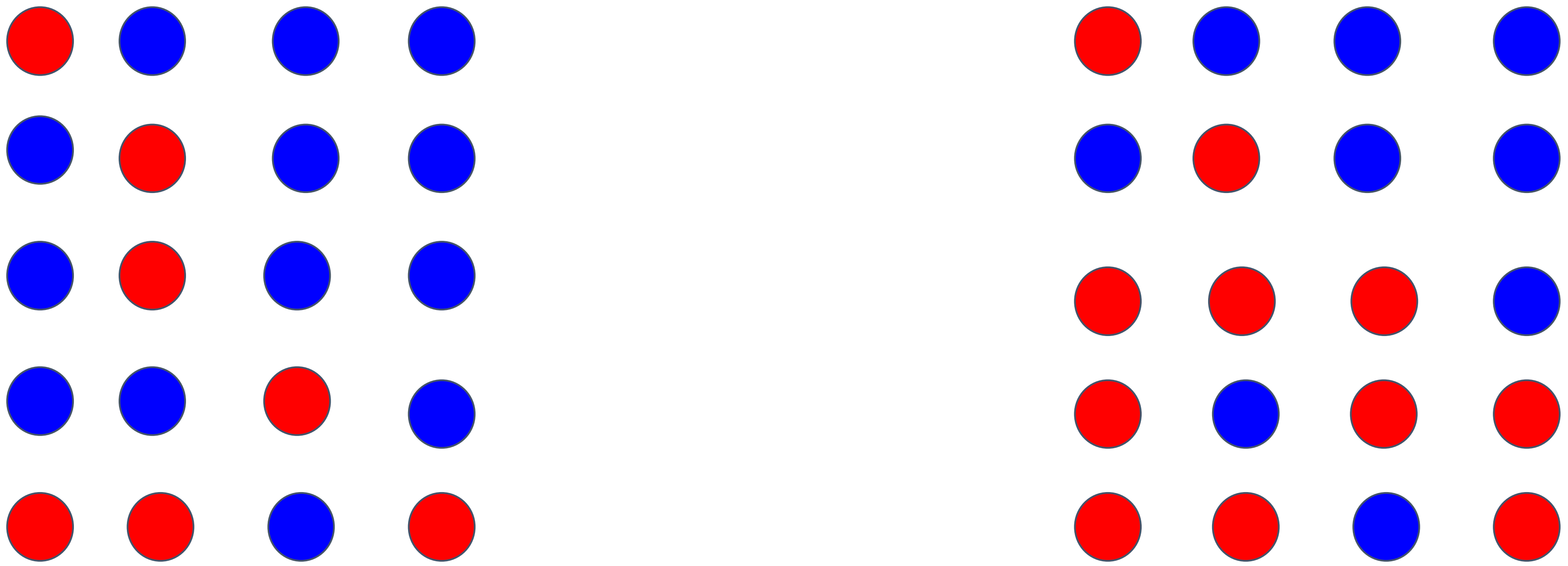
Calculate odds ratio of being red between cases and controls

Odds are 7/13 in controls

Odds are 11/9 in cases

Case-control study - is red associated?

CONTROLS **CASES**



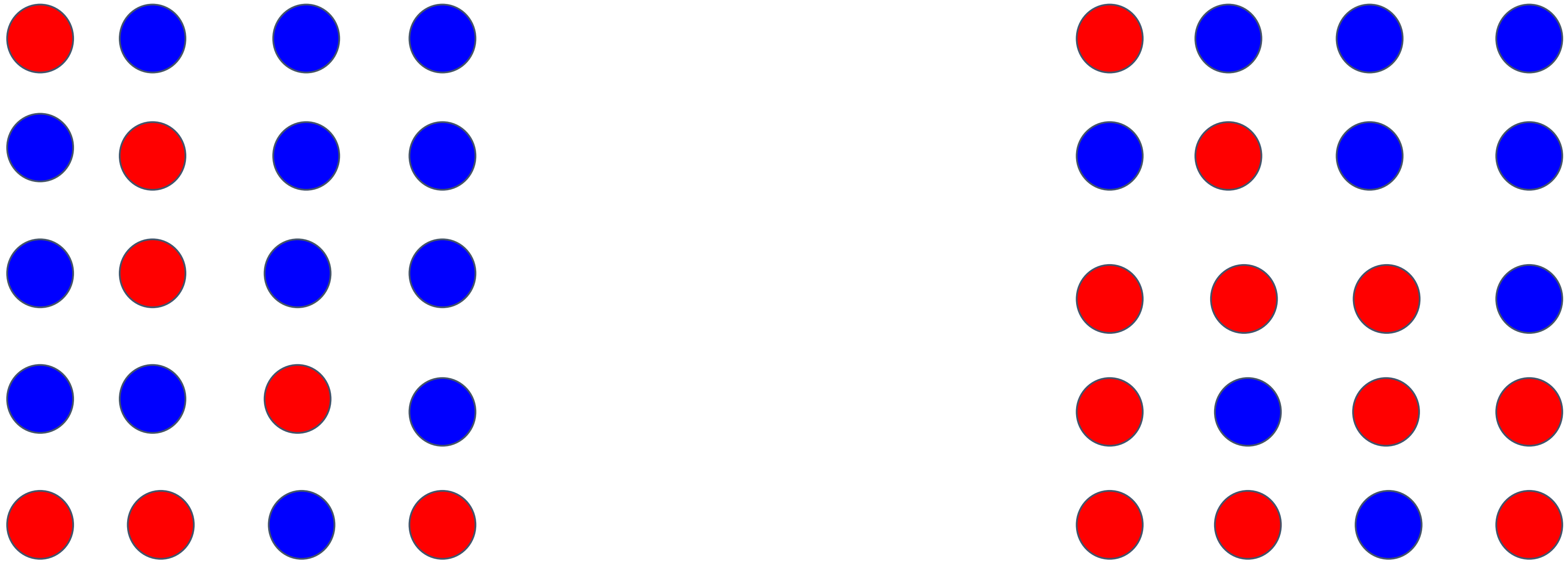
Calculate odds ratio of being red between cases and controls

Odds are 7/13 in controls Odds are 11/9 in cases

Odds ratio is (11/9) / (7/13) = 2.27!

Case-control study - is red associated?

Cases are 2.27x times as likely to be red as controls

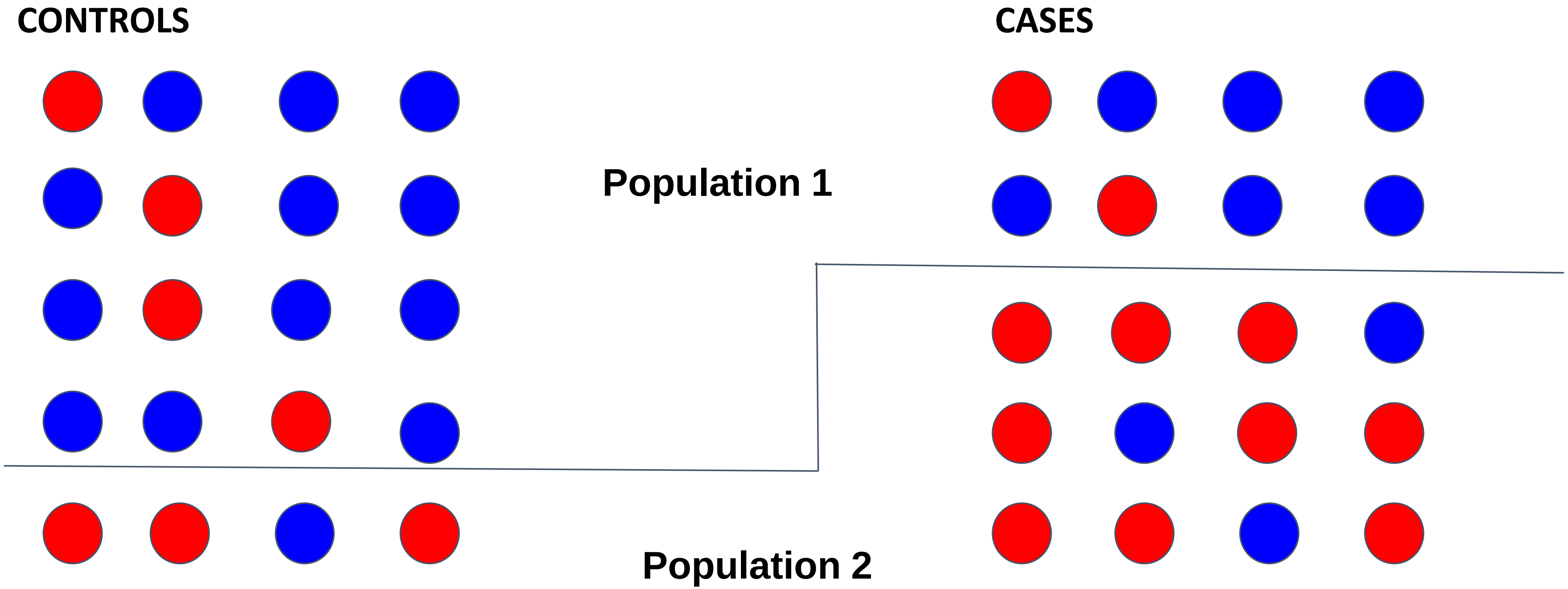


CONTROLS

CASES

Case-control study - is red associated?

Cases are 2.27x times as likely to be red as controls



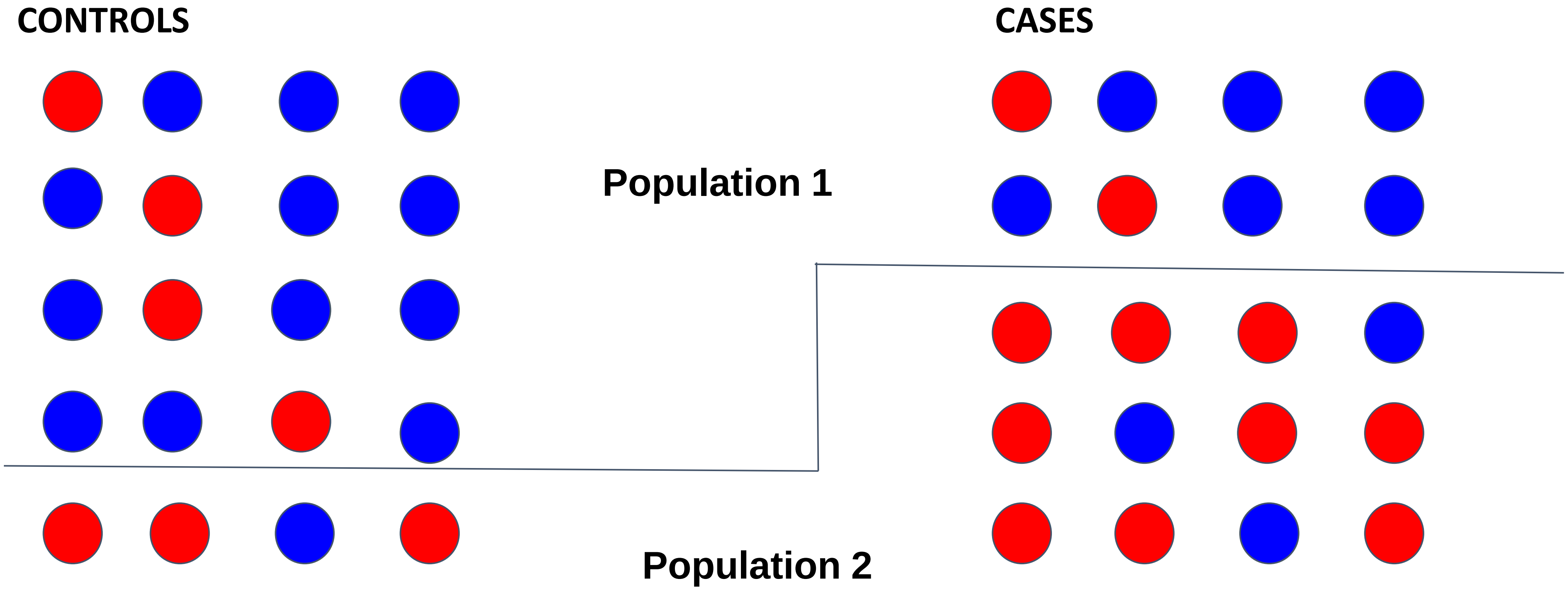
CONTROLS

CASES

Population 1

Population 2

Case-control study - is red associated?



CONTROLS

CASES

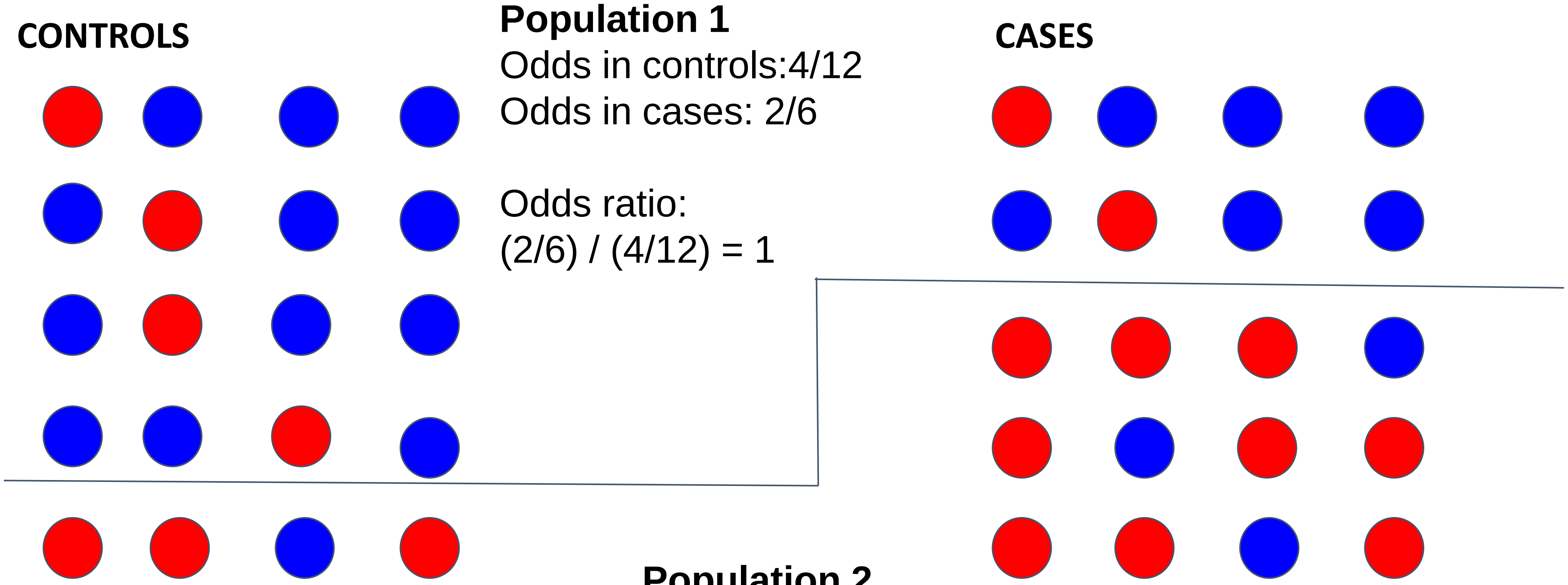
Population 1

Population 2

Now recalculate the odds ratios in population 1 and population 2 separately

Now recalculate the odds ratios in population 1 and population 2 separately

Case-control study - is red associated?



CONTROLS

Population 1
Odds in controls:4/12
Odds in cases: 2/6

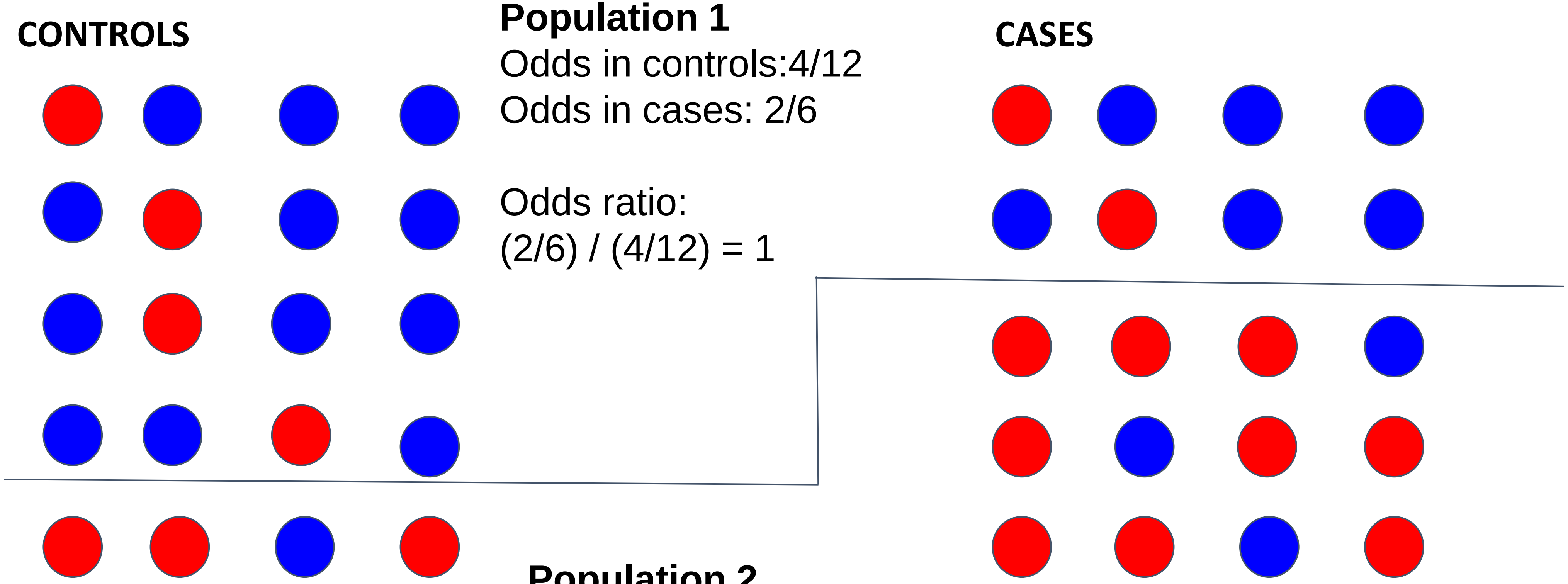
Odds ratio:
(2/6) / (4/12) = 1

CASES

Population 2

Now recalculate the odds ratios in population 1 and population 2 separately

Case-control study - is red associated?



CONTROLS

CASES

Population 1
Odds in controls:4/12
Odds in cases: 2/6

Odds ratio:
(2/6) / (4/12) = 1

Population 2
Odds in controls:3/1
Odds in cases: 9/3

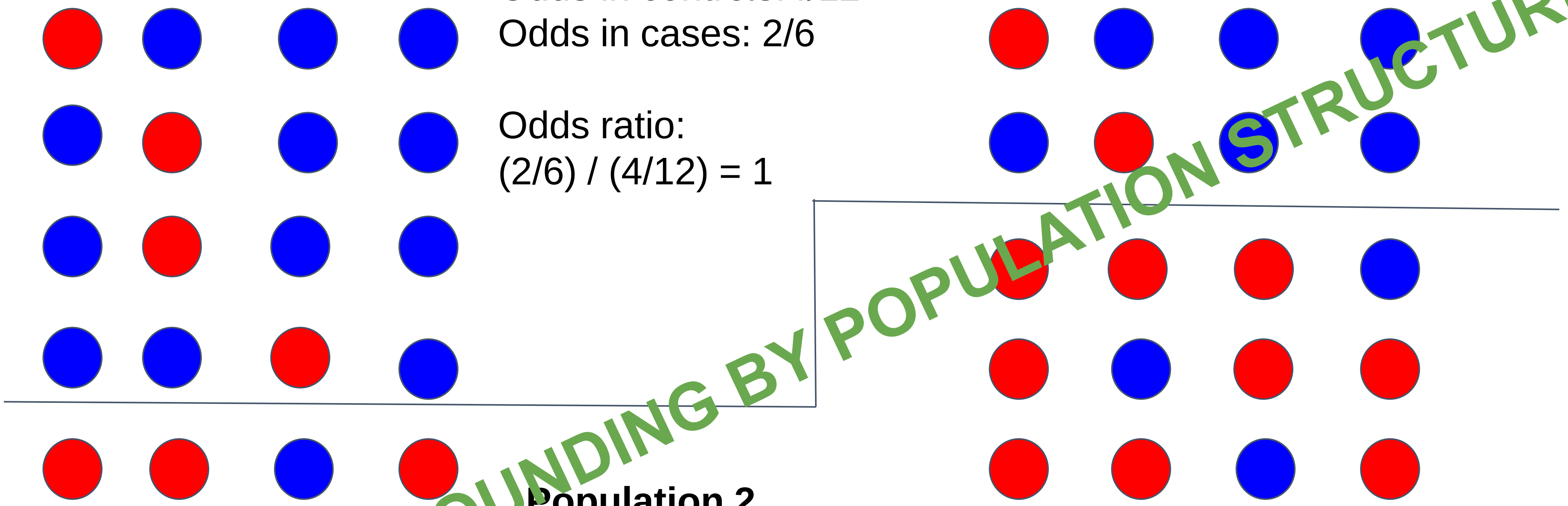
Odds ratio:
(9/3) / (3/1) = 1

Now recalculate the odds ratios in population 1 and population 2 separately

Case-control study - is red associated?

CONTROLS

CASES



Population 1

Odds in controls:4/12

Odds in cases: 2/6

Odds ratio:

(2/6) / (4/12) = 1

Population 2

Odds in controls:3/1

Odds in cases: 9/3

Odds ratio:

(9/3) / (3/1) = 1

CONFOUNDING BY POPULATION STRUCTURE