Module 5A : A Non-Parametric Test

Odds Ratio, GWAS, and PRS

Agenda:

- Odds ratio
- Genome-Wide Association Studies

Odds Ratio:

Another type of "Contingency analysis" that **measures the magnitude of association between two categorical variables that each only have two categories:**

- Explanatory and response variables
  - the response variable has usually adopts "success" and "failure" as the labels for its two categories
  - Used in **case-control** groups
  - **Proportion** of success/failure between two groups
  - Step 1: Usually testing **Ho: OR=1**

Step 2 (the test statistic)

Odds:

Probability of success divided by the probability of failure

$$O = \frac{p}{1-p}$$

As per usual, we will be using estimates:

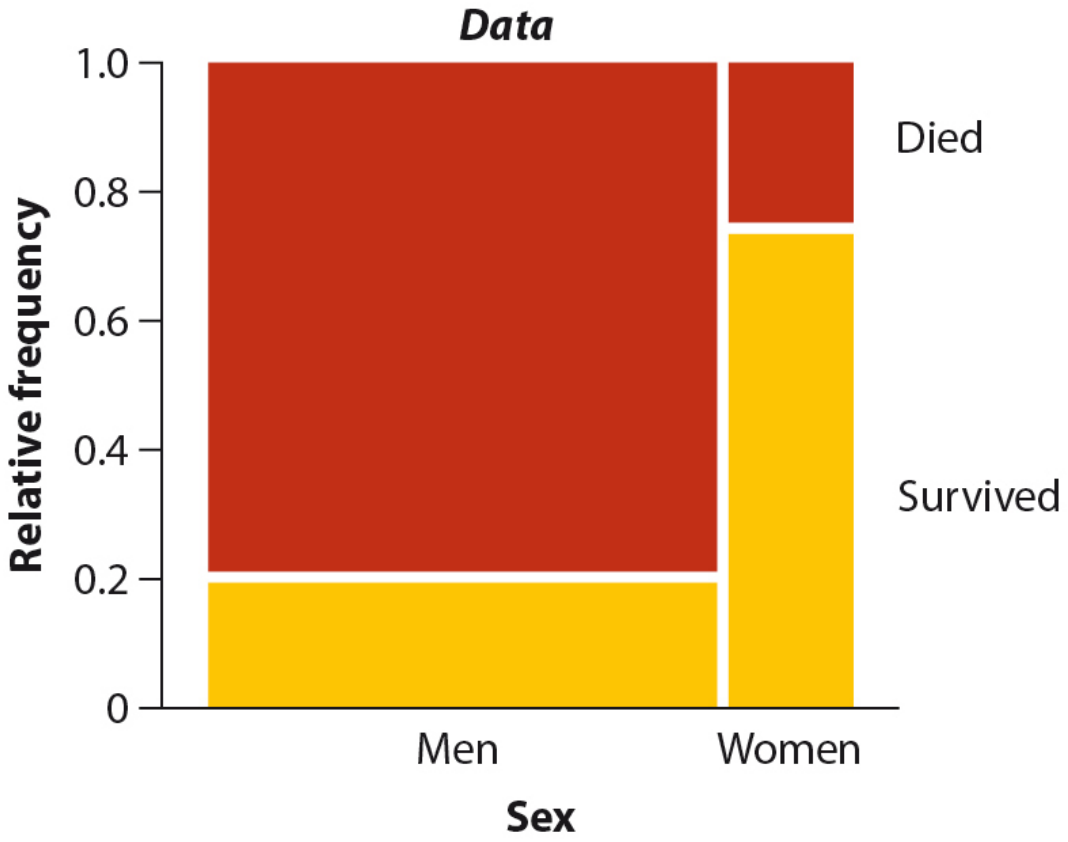
$$\hat{O} = \frac{\hat{p}}{1-\hat{p}}$$

Mosaic plots!

Male versus female passengers on the Titanic

Odds:

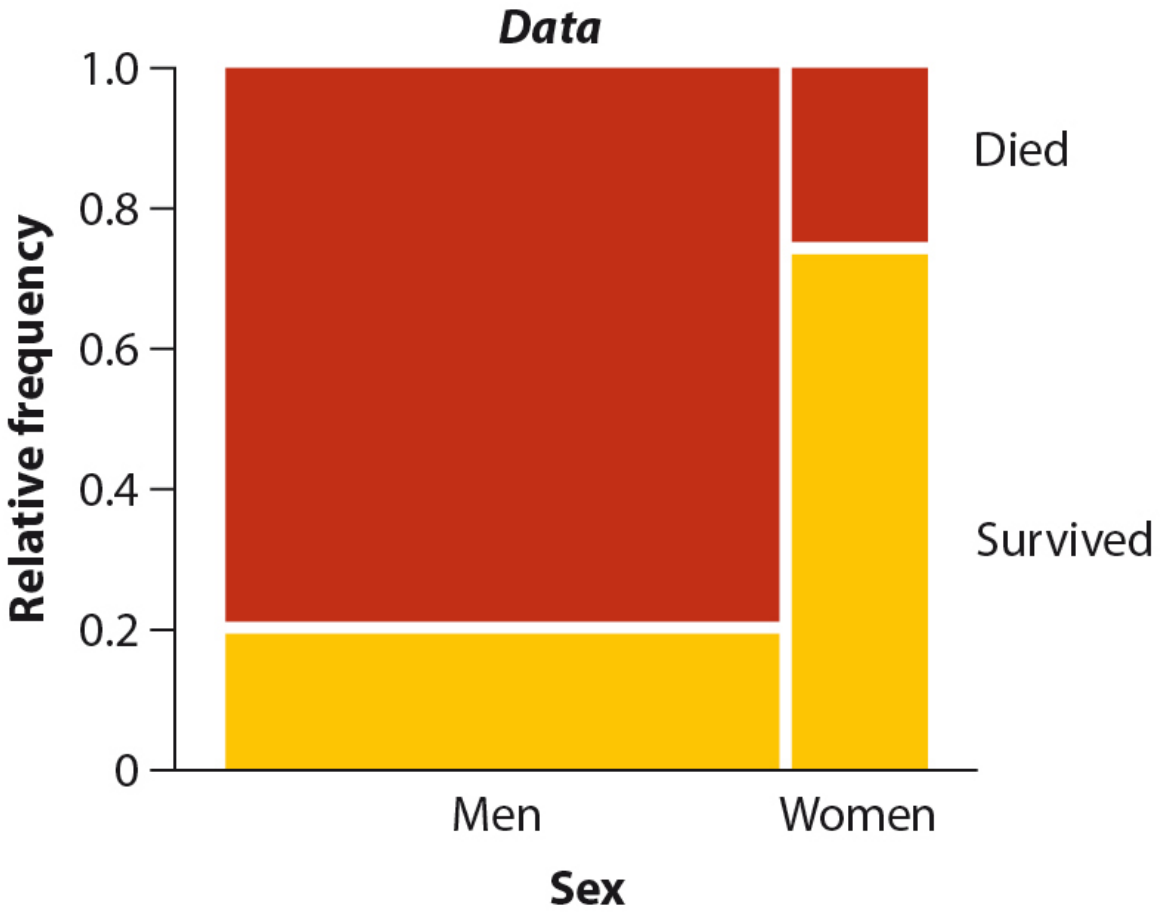
Probability of success divided by the probability of failure



$$O = \frac{p}{1-p}$$

Odds:

Probability of success divided by the probability of failure



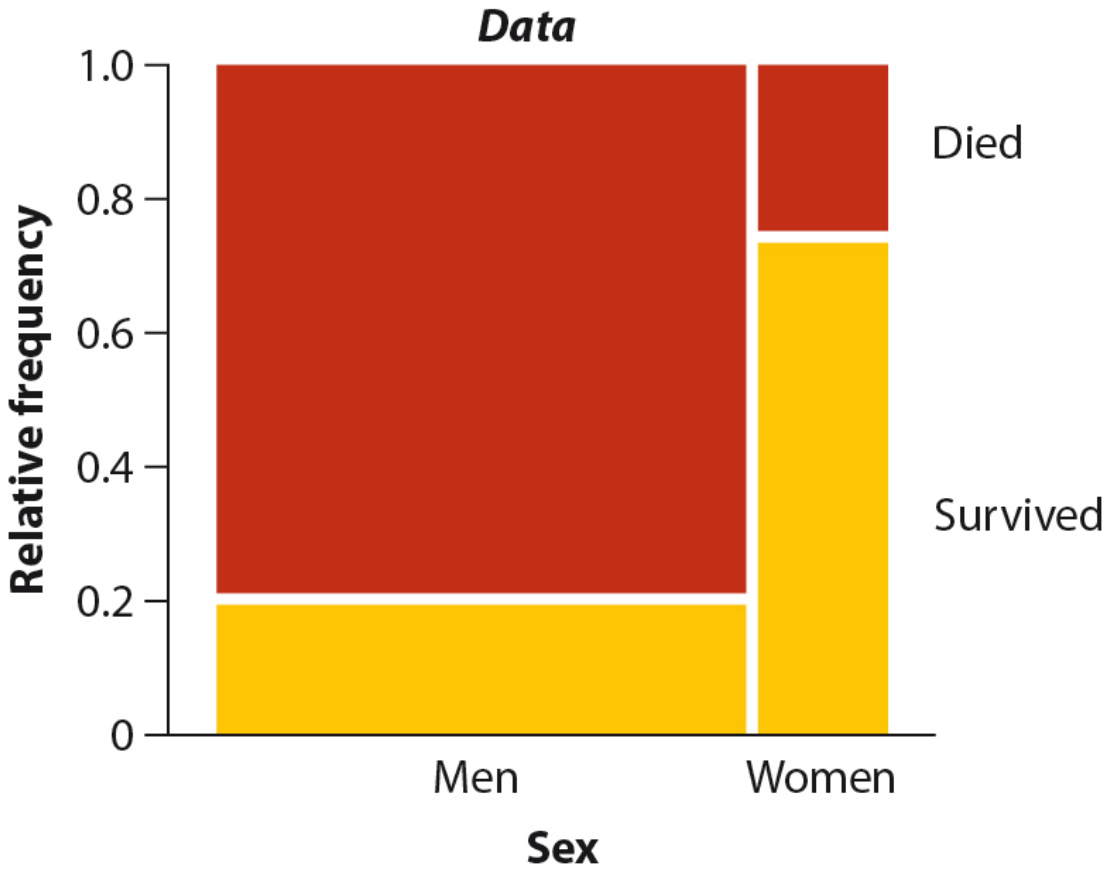
$$O = \frac{p}{1-p}$$

$$O_{men} = \frac{0.20}{1-0.20} = 0.25$$

$$O_{women} = \frac{0.74}{1-0.74} = 2.85$$

Odds:

Probability of success divided by the probability of failure



$$O = \frac{p}{1-p}$$

1 to 4

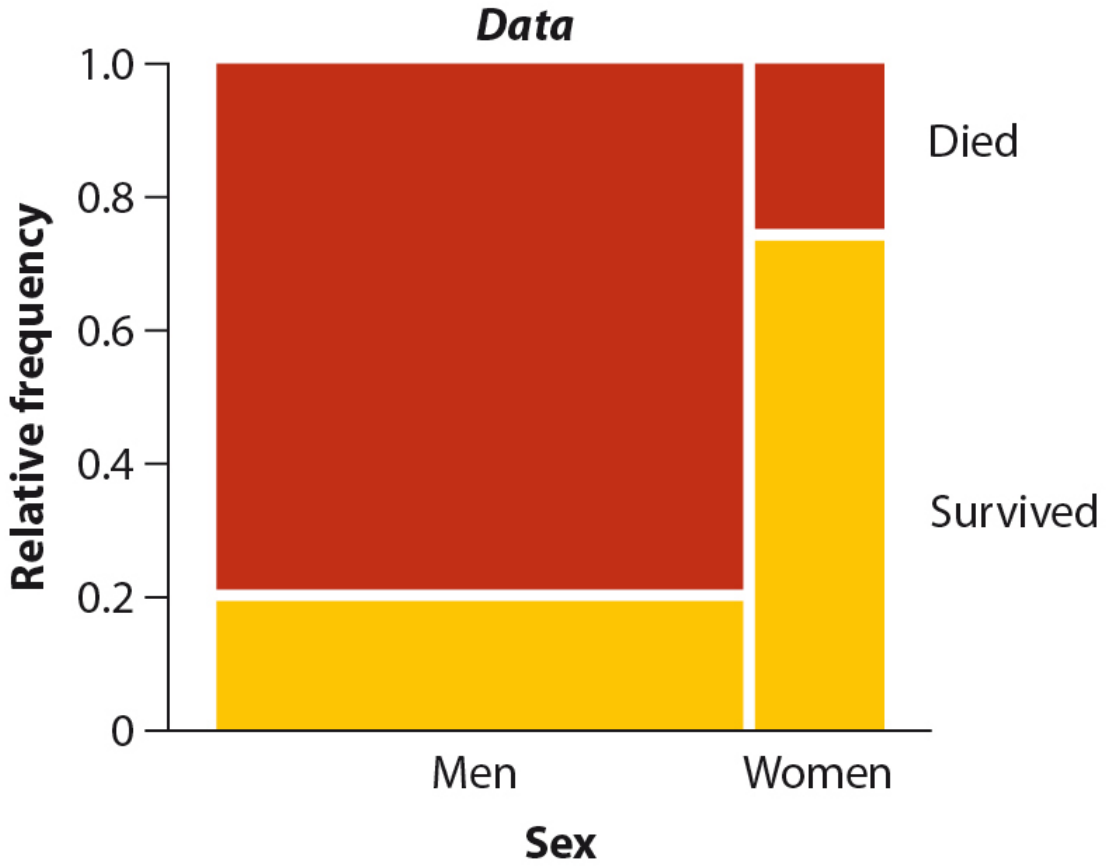
$$O_{men} = \frac{0.20}{1-0.20} = 0.25$$

$$O_{women} = \frac{0.74}{1-0.74} = 2.85$$

3 to 1

Odds Ratio:

The odds of success in one group divided by the odds of success in another group



$$OR = \frac{O_1}{O_2}$$

Odds ratio of female to male survival:

$$OR = \frac{2.85}{0.25} = 11.4$$

Odds Ratio:

The odds of success in one group divided by the odds of success in another group

- *usually asking "Does the treatment/intervention" change the outcome (compared to control)?*

$$\hat{O}R = \frac{\hat{O}_1}{\hat{O}_2} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

	Treatment	Control
Success	a	b
Failure	c	d

Odds Ratio:

Measures the magnitude (or strength) of association between two categorical variables that each only have two categories:

- Explanatory and response variables
  - the response variable usually adopts "success" and "failure" as the labels for its two categories
  - Used in **case-control** groups
  - **Proportion** of success/failure between two groups
  - Step 1: Usually testing **Ho: OR=1**

The most challenging parts of an odds-ratio:

1. *Keep track of which one is a success, and which one is a failure*
2. *The TRANSFORMATION that is **necessary** for step 3*

Confidence Interval Odds Ratio:

- Confidence interval is used to determine whether O.R. >>1 or << 1 is statistically significant (Ho: OR =1)

- Same basic idea as confidence intervals:

Point Estimate ± Z*Standard Error

For example, 95% Confidence Interval: $\bar{X} \pm 1.96 * SE_{\bar{x}}$
This corresponds to an interval: $\bar{X} - 1.96 * SE_{\bar{x}} < \mu < \bar{X} + 1.96 * SE_{\bar{x}}$

but… the OR sampling distribution is right skewed not Normally distributed!

What do we do?

Confidence Interval Odds Ratio:

Step 3 (determining if it is statistically significant or not):
General approach involves **Transformation (let R handle it!)**:

- ln(OR) ~ Normally distributed
- Confidence Interval boundaries are found
  - Calculate S.E.
- Converted back using **exponential distribution**

Confidence Interval Odds Ratio:

Step 4:

- **Conclude:**

  OR = 1, If the 95% (or 99%) Confidence interval contains 1, the indicates that there is no association at the (5% or 1%) significant level.

  If the 95% (or 99%) Confidence interval does not contain 1 then we can conclude that there is statistically significant (at the 5% or 1% level) association between the variables (i.e.. lack of disease and treatment etc.)

Revisit this Example:

The influence of SES on preterm delivery rates

Socio-Economic status	Cases	Controls
Upper	11	40
Upper-middle	14	45
Middle	33	64
Lower-middle	59	91
Lower	53	58
Unknown	5	5

This time, focus on the two extreme categories, Upper and Lower, so you can use Odds Ratio instead and compare your answer to the X^2 Contingency Analysis result.

Confidence Interval Odds Ratio:

Example: The influence of SES on preterm delivery rates

Step 1: Odds ratio = (53/58)/(11/40) = 3.32

Step 2: Calculate ln(OR):

ln(3.32) = 1.20

Step 3: The confidence interval for the **ln (OR)** is a normally distributed sampling distribution, we can use **Z***. So, for a 95% confidence interval ($\alpha = 0.05$), we can use 1.96.

Relative Risk

- Another commonly used measure of association between two categorical variables (when both have two categories)
- Especially appropriate for comparing risk of rare and undesirable outcome i.e. SID syndrome in children who sleep facing upward and those who sleep on their stomach (as long as babies are randomly sampled)
- Like odds ratio but perhaps more intuitive
- Should give similar answer as OR for rare events

$$\hat{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{'treatment'}{'control'}$$

Relative Risk

Example:

Given, event B = diagnosis of breast cancer w/i 2 years

event A = positive mammogram at present

event A^C = negative mammogram at present

P(diagnosis of breast cancer w/i 2 years|negative mammogram) = 20/100,000

P(diagnosis of breast cancer w/i 2 years|positive mammogram)=1/10

$$R\hat{R} = \frac{P(B \mid A)}{P(B \mid A^C)} = \frac{1/10}{2/10{,}000} = 500$$

Individuals with a positive mammogram at present have a relative risk of developing breast cancer within two years that is 500 times those of individuals with a negative mammogram.

Genome-**W**ide **A**ssociation **S**tudies (GWAS)

- Each SNP (Single Nucleotide Polymorphism) is an independent test
- Associations are tested by comparing the frequency of each allele in cases and controls
  - You can extend this analysis. For instance, the frequency of each of 3 possible genotypes can also be compare. We'll see an example of 2 (allele counting) and 3 (genotype counting).

  - *Always remember: GWAS doesn't assume that the SNP itself is causing the disease, but that it is located close enough to the causative allele that they have tight linkage (they are not usually separated by recombination)*.

  - *Here is a short YouTube video from an 'old' series called "Useful Genetics" that has a fantastic introduction to GWAS*

https://www.youtube.com/watch?v=5sgPkRXR6pE

GWAS is odds ratio

Odds ratio = odds(event | exposure)
odds(event | lack of exposure)

Example:

P (D | genotype “AT”) = 0.75

P (D | genotype “TT”) = 0.25

GWAS Question:

OR for getting the disease with genotype AT compared to TT?

OR = (0.75 / 0.25) / (0.25 / 0.75) = 9

What’s the OR for AT individuals relative to an average population risk of 10%?

OR = (0.75 / 0.25) / (0.10/ 0.90) = 27

GWAS and Odds Ratio

Association of rs1234567 with some type of cancer

	CC	CT	TT
Cases	250	375	150
Controls	460	940	500

* C is risk allele

$\mathbf{OR_{TT}}$ = odds(disease|TT)/odds(disease|TT) = 1

$\mathbf{OR_{CT}}$ = odds(disease|CT)/odds(disease|TT) =375*500/150*940 = 1.33

$\mathbf{OR_{CC}}$= odds(disease|CC)/odds(disease|TT) =250*500/460*150 = 1.81

(These are all taken with respect to the lowest risk genotype, TT; they cannot be applied to an individual. To convert this into risk estimate, the prevalence of the disease and the genotypic frequencies must be accounted).

GWAS and Odds Ratio

Association of rs1234567 with some type of cancer

	C*	T
Cases	875 (56.5)	675 (43.5)
Controls	1860 (48.9)	1940 (51.1)

* C is risk allele

$$\mathbf{OR_C} = \frac{\text{odds(disease|C)}}{\text{odds(disease|T)}} = \frac{875*1940}{1860*675} = 1.35$$

Cases: C alleles = 2* 250 (CC) + 1*375 (CT) = 875

T alleles = 2*150 (TT) + 1*375 (CT) = 675

Controls: C alleles = 2* 460 (CC) + 1*940 (CT) = 1860

T alleles = 2*500 (TT) + 1*940 (CT) = 1940

(Still not useful for individual prognostications)

Odds Ratio to Probabilities

P(Disease) = prevalence = P(Dis|AA)P(AA) + P(Dis|Aa)P(Aa) + P(Dis|aa)P(aa) =0.212

Combining Bayes' with population genotype frequency (HapMap or BioBank) and disease prevalence for that population information!

P(Dis|AA) = 0.175

P(Dis|AG)= 0.210

P(Dis|GG) = 0.248

OR_{GG} = odds(disease | GG) / odds(disease | AA) = (0.248/ (1-0.248)) / (0.175/(1-0.175)) = 1.55

OR^*_{GG} = odds(disease | GG) / odds(disease in avg pop) = (0.248/ (1-0.248)) / (0.212/(1-0.212)) = 1.22

These OR*s are relative to the average population - can be directly applied to an individual

GWAS Another worked example

SNP	Risk Allele	Cases with Allele	Cases without	Controls with Allele	Controls without
rs101	A	40	60	20	80
rs202	G	70	30	50	50
rs303	T	25	75	25	75
rs404	C	10	90	20	80

GWAS

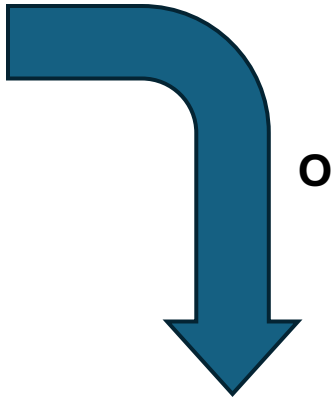
SNP	Risk Allele	Cases with Allele	Cases without	Controls with Allele	Controls without
rs101	A	40	60	20	80
rs202	G	70	30	50	50
rs303	T	25	75	25	75
rs404	C	10	90	20	80

OR=(a/b)/(c/d)=(axd)/(bxc)

where:

- a= cases with allele
- b= cases without
- c= controls with allele
- d= controls without

SNP	Risk Allele	Cases with Allele	Cases without	Controls with Allele	Controls without
rs101	A	40	60	20	80
rs202	G	70	30	50	50
rs303	T	25	75	25	75
rs404	C	10	90	20	80



OR=(a/b)/(c/d)=(axd)/(bxc)

SNP	Cases (a,b)	Controls (c,d)	OR	Interpretation
rs101	40, 60	20, 80	(40×80)/(60×20)=**2.67**	Risk allele A roughly doubles odds of disease
rs202	70, 30	50, 50	(70×50)/(30×50)=**2.33**	G allele modestly increases disease risk
rs303	25, 75	25, 75	(25×75)/(75×25)=**1.00**	No association, frequencies equal
rs404	10, 90	20, 80	(10×80)/(90×20)=**0.44**	C allele may be protective (less common in cases)

You could compute 95% Confidence Interval.....

As always: correlation ≠ causation

Upload to Canvas as part of your Certification of Completion

Module 5A Questions:

Using the following data, estimate the OR in favour of Myocardial Infarction (MI) over three years for an OC user compared with a non-OC user (i.e. the disease odds ratio):

MI incidence (3 yrs) / OC-use group	Yes	No
YES	13	4987
NO	7	9993