Applied Epistemology:

A Framework for how we know things scientifically.

A Refresher on Learning Path I: Hypothesis Testing

Agenda:

1. H_O/H_A: Our model of the test universe (the distribution of the variable)
2. Test & assumptions: are the assumptions met? Is the test valid?
3. Quantitative evidence: **p-value**, or critical value.
   - False positive =Type I (α), False Negative = Type II (β), Type III errors
   - Sensitivity, Specificity, Power → confusion matrix, ROC/AUC curve
   - Confusion Matrix
4. Conclusion & uncertainty/estimation
5. **Z-scores, χ^2 Goodness-of-fit test,** and **χ^2 Contingency test**

Perspectivism – why assumptions are important: https://hdsr.mitpress.mit.edu/pub/qasl4fza/release/3

Contingency Analysis

Contingency: *allows us to determine if two categorical variables are associated (some contingency tests will allow us to quantify the degree of association as well but not all do this).*

Major tests:

- **χ^2 Contingency Test** –> similar but not exactly as the same χ^2 Goodness of fit test. You can think of it as a subset of χ^2 Goodness of fit tests with some calculation differences. Basis of test is Multiplication rule with the assumption of independence. Degrees of freedom are calculated differently!

- **Odds ratio** –> used in case-control

Ho: OR=1. Challenge: transforming the sampling distribution of OR so that it is normally distributed.

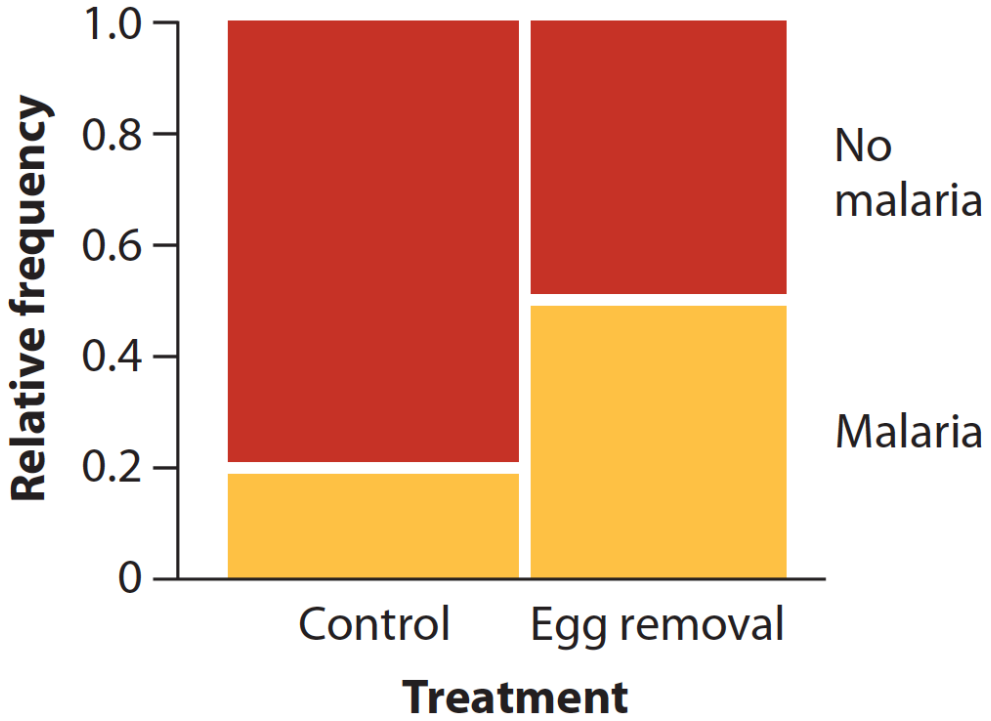
- **Relative Risk** -> like OR but used when you know the actual proportion of the focal character in the population (OR and RR are the same for rare events, definitely don't use RR for Case-Control)

- **Fisher's Exact test** -> exact calculation. You can think of it as the contingency version of calculating a p-value

Contingency Analysis:

- Associations between categorical variables
- **Test the independence of two or more categorical variables**

	Control Group	Egg-Removal Group	Row Total
Malaria	7	15	22
No Malaria	28	15	43
Column Total	36	30	65



Reminder: Multiplication Rule

Multiplication rule:

$$P[A \text{ and } B] = P[A|B]P[B] = P[B|A]P[A]$$

IFF INDEPENDENT, this collapses to:

$$P[A \text{ and } B] = P[A]P[B]$$

χ^2 Contingency Test:

- Tests goodness-of-fit to the data of the null hypothesis of independence of variables

- Two categorical variables but, unlike the Odds Ratio, each variable can have more than 2 categories

- Assumptions:
  - The value of the cell ***expected values*** should be 5 or more in at least 80% of the cells
  - No cell should have an **expected value** of less than one

- Description of χ^2 Contingency Test:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/

Example: You have a lean strain and an obese strain of mouse where previous QTL mapping has identified a locus on chromosome 6 associated with body weight. You cross them and genotype 120 male F_2 and classify them into to simple classes:

G: carries at least one "lean" allele (GG or Gg)

A: carries at least one "obesity" allele (AA or Aa)

The phenotype is classed as obese if its 20-week body weight exceeds 35g.

	Obese	Not Obese
Genotype G (lean)	18	42
Genotype A (Obese)	32	28

Is obesity status independent of genotype?

A. Yes, we reject the null hypothesis
B. No, we fail to reject the null hypothesis
C. Yes, we fail to reject the null hypothesis
D. No, we reject the null hypothesis

Example: You have a lean strain and an obese strain of mouse where previous QTL mapping has identified a locus on chromosome 6 associated with body weight. You cross them and genotype 120 male F_2 and classify them into to simple classes:

G: carries at least one "lean" allele (GG or Gg)

A: carries at least one "obesity" allele (AA or Aa)

The phenotype is classed as obese if its 20-week body weight exceeds 35g.

	Obese	Not Obese
Genotype G (lean)	18	42
Genotype A (Obese)	32	28

Is obesity status independent of genotype?

Step 1: Formulate your null hypothesis

Ho: Obesity status is independent of genotype (No association)

Ha: Obesity status is associated with genotype

Example: You have a lean strain and an obese strain of mouse where previous QTL mapping has identified a locus on chromosome 6 associated with body weight. You cross them and genotype 120 male F_2 and classify them into to simple classes:

G: carries at least one "lean" allele (GG or Gg)

A: carries at least one "obesity" allele (AA or Aa)

The phenotype is classed as obese if its 20-week body weight exceeds 35g.

	Obese	Not Obese
Genotype G (lean)	18	42
Genotype A (Obese)	32	28

Step 1: Formulate your null hypothesis

Ho: Obesity status is independent of genotype (No association)

Ha: Obesity status is associated with genotype

Step 2: Identify the test statistic

χ^2 expectation under independence.

Assumptions: no cells less than 5 so both assumptions are met.

With independence,

P[A AND Obese] = ?

Example: You have a lean strain and an obese strain of mouse where previous QTL mapping has identified a locus on chromosome 6 associated with body weight. You cross them and genotype 120 male F_2 and classify them into to simple classes:

G: carries at least one "lean" allele (GG or Gg)

A: carries at least one "obesity" allele (AA or Aa)

The phenotype is classed as obese if its 20-week body weight exceeds 35g.

	Obese	Not Obese
Genotype G (lean)	18 (25)	42 (35)
Genotype A (Obese)	32 (**25**)	28 (35)

Step 1: Formulate your null hypothesis

Ho: Obesity status is independent of genotype (No association)

Ha: Obesity status is associated with genotype

Step 2: Identify the test statistic

χ^2 expectation under independence.

Assumptions: no cells less than 5 so both assumptions are met.

With independence, P[A AND Obese] = P[A]*P[Obese] = 60/120 * 50/120 * 120 = 25

The other expected cells can be calculated from this **one** calculation since the rows and columns need to add up to the same observational counts.

Example: You have a lean strain and an obese strain of mouse where previous QTL mapping has identified a locus on chromosome 6 associated with body weight. You cross them and genotype 120 male F_2 and classify them into to simple classes:

G: carries at least one "lean" allele (GG or Gg)

A: carries at least one "obesity" allele (AA or Aa)

The phenotype is classed as obese if its 20-week body weight exceeds 35g.

	Obese	Not Obese
Genotype G (lean)	18 (25)	42 (35)
Genotype A (Obese)	32 (**25**)	28 (35)

Step 1: Formulate your null hypothesis

Ho: Obesity status is independent of genotype (No association)

Ha: Obesity status is associated with genotype

Step 2: Identify the test statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E} = (18\text{-}25)^2/25 + (32\text{-}25)^2/32 + (42\text{-}35)^2/35 + (28\text{-}35)^2/35 = 6.29$$

Step 3: quantifying evidence

df = 1

χ^2_1=3.84

Step 4: Conclusion

Our calculated χ^2 value of 6.29 > χ^2_1=3.84 so we **reject** the null hypothesis

Example: You have a lean strain and an obese strain of mouse where previous QTL mapping has identified a locus on chromosome 6 associated with body weight. You cross them and genotype 120 male F_2 and classify them into to simple classes:

G: carries at least one "lean" allele (GG or Gg)

A: carries at least one "obesity" allele (AA or Aa)

The phenotype is classed as obese if its 20-week body weight exceeds 35g.

	Obese	Not Obese
Genotype G (lean)	18 (25)	42 (35)
Genotype A (Obese)	32 (**25**)	28 (35)

We can follow this up with a quick odds ratio to answer the question:
Which genotype looks riskier and by how much?

Odds of obesity for genotype A: 32/28

Odds of obesity for genotype G: 18/42

$$\text{OR} = \frac{32/28}{18/42} = 2.67$$