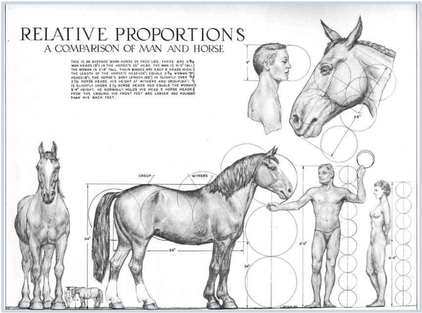
Institute for Veterinary Epidemiology and Biostatistics (IVEB)
Department of Veterinary Medicine

DRS Spring School 2018 – week 2

Working with proportions



Horse Anatomy by Lewis S. Brown

PD Dr. Lorenz Gygax
(HU Berlin)

Objectives

- Estimate difference between 2 proportions
- Measure the association between an exposure and a disease

Plan

1. **Terminology**
2. **Confidence Interval (CI) for proportion**
3. **Comparing proportions**
4. **Measures of association**
    1. **Risk ratio**
    2. **Odds ratio**
    3. **Incidence rate ratio**

1. Terminology

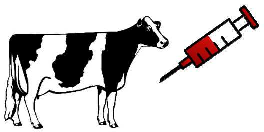
- In statistics, proportions are parameters that summarise the observation of a binary variable.
- A binary variable is a categorical variable with only 2 categories of response often termed <u>success</u> and <u>failure.</u>

1. Terminology

Examples

- When tossing a coin, we can define
  HEAD= **SUCCESS** and *TAIL*= **FAILURE**

- When testing cows for *Leptospira* antibodies, we can define
  +*TEST*= **SUCCESS** and –*TEST*= **FAILURE**



2. Confidence Interval (CI) for a proportion

- Sampling distribution is approximately Normal if the sample size (n) is large
- Sample proportion (p) is an unbiased estimate of population proportion (π)
- Standard deviation (SD) = $\sqrt{\frac{p(1-p)}{n}}$
- 95% CI = p ± 1.96*SD

3. Comparing proportions

⇒ **Is there an association between an exposure and a disease?**

Example: Proportion in 2*2 table

Effect of early castration on male mice diabetes?

3. Comparing proportions

Effect of early castration on male mice diabetes?

Building frequency or contingency table

Observed frequencies

	Groups	
Outcome	**Group 1 (castrated)**	**Group 2 (control)**
Success (Diabetic)	a	b
Failure (non Diabetic)	c	d

3. Comparing proportions

Effect of early castration on male mice diabetes?

Building frequency or contingency table

Observed frequencies

	Groups		
Outcome	**Group 1** (castrated)	**Group 2** (control)	Row total
Success (Diabetic)	a	b	a+b
Failure (non Diabetic)	c	d	c+d
Column total	n1=a+c	n2=b+d	Overall total n=a+b+c+d
Observed proportion of successes	p1=a/n1	p2=b/n2	p=(a+b)/n

3. Comparing proportions

Chi-squared test

- Compare Observed (O) vs. Expected (E) frequency:

$$Test = \sum \frac{(|O - E| - 0.5)^2}{E}$$

Yates' continuity correction (has impact only for low sample sizes)

- Assumptions & alternatives
  - Each individual only represented once (one group, one outcome)
  - Each individual was randomly allocated to group (independence!)
  - Expected frequency is >5 in all cells
    - If not, use **Fisher's exact test**

3. Comparing proportions

Expected frequencies if H0 were true...

⇒ Reminder: H0 = No difference between Groups 1 and 2

	Groups	
Outcome	**Group 1**	**Group 2**
Success	$\frac{(a+c)(a+b)}{n}$	$\frac{(b+d)(a+b)}{n}$
Failure	$\frac{(a+c)(c+d)}{n}$	$\frac{(b+d)(c+d)}{n}$

3. Comparing proportions

Effect of early castration on male mice diabetes?

Example: Experience conducted by (Hawkins, 1993)

Observed frequencies

	Castrated mice	Control mice	Total
Diabetic	26	12	38
Non-diabetic	24	38	62
Total	50	50	100
Obs. Proportion of Diabetic	0.52	0.24	0.38

Expected frequencies

	Castrated mice	Control mice
Diabetic	(50*38)/100= 19	(50*38)/100=19
Non-diabetic	(50*62)/100= 31	(50*62)/100=31

Identical expected frequencies arise because the group sizes are equal

Chi-squared test: (|26-19|-05)^2/19 + (|12-19|-05)^2/19 + (|24-31|-05)^2/31 + (|38-31|-05)^2/31 = **7.17**

For interpretation, calculate your df and look at Chi-square distribution table...

3. Comparing proportions

Effect of early castration on male mice diabetes?

Example: Experience conducted by (Hawkins, 1993)

Doing the same using R

`chisq.test()` OR `fisher.test( )`

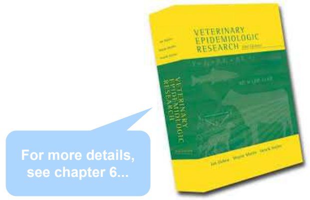
```
> Mice_diabete
     [,1] [,2]
[1,]   26   12
[2,]   24   38
> chisq.test(Mice_diabete)

        Pearson's Chi-squared test with Yates' continuity correction

data:  Mice_diabete
X-squared = 7.1732, df = 1, p-value = 0.0074
```

p-value = 0.007
It is rather unlikely that diabetic status in male mice is independent of early castration.

4. Measures of association

⇒ **What is the strenght of the association between an exposure and a disease?**

- The strength of an association usually expressed using Risk ratio (RR), Odds ratio (OR) and Incidence rate ratio (IR).
- Choice of appropriate measure depends on the study design and its corresponding measure of disease frequency:
  - Cohort studies: RR, IR
  - Cross sectional studies: RR, OR
  - Case-control studies: OR



4. Measures of association

Risk ratio (RR)

- RR is ratio of the risk of disease in the exposed group to the risk of disease in the non-exposed group
- RR ranges from 0 to ∞
  - RR<1 exposure is protective (e.g. Vaccines)
  - RR=1 exposure has no effect
  - RR>1 exposure is positively associated with disease

4. Measures of association

Risk ratio (RR)

e.g., Ocular carcinoma and eyelid pigmentation in cohort study of Hereford cattle

	Eyelids		
	Non-pigmented	**Pigmen-ted**	Row marginal total
Occular carcinoma +	38	2	40
Occular carcinoma -	4962	998	5960
Column marginal total	5000	1000	6000

$$RR = \frac{(\frac{38}{5000})}{(\frac{2}{1000})} = 3.8$$

Risk of cancer in cattle with white eyelids is 3.8 time higher than that of cattle with pigmented eyes

4. Measures of association

Odds ratio (OR)

- odd is the ratio of the probability that the event will happen to the probability that the event will not happen

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

- OR: ratio of two odds
- Same interpretation as RR

Examples odds:

- A pregnant woman has a 1 in 705,000 chance of giving birth to quadruplets 
- Someone eating an oyster has a 1 in 12,000 chance of finding a pearl inside of it 

4. Measures of association

Odds ratio (OR)

e.g., Ocular carcinoma and eyelid pigmentation in cohort study of Hereford cattle

Eyelids			
	Non-pigmented	Pigmented	Row marginal total
Occular carcinoma +	38	2	40
Occular carcinoma -	4962	998	5960
Column marginal total	5000	1000	6000

OR ≈ RR
Always the case when the disease is rare!

$$OR = \frac{\left(\frac{38}{4962}\right)}{\left(\frac{2}{998}\right)} = 3.82$$

Odds for cancer in cattle is 3.8 time higher in cattle with white eyes than in cattle with pigmented eyes

4. Measures of association

Incidence rate ratio (IR)

- IR is ratio of the incidence rate in an exposed group to the incidence rate in the non-exposed group
- Same interpretation as RR
- IR can only be computed in cohort studies

4. Measures of association

Incidence rate ratio (IR)

e.g., Mastitis and pre-dipping in a dairy herd

Teats			
	Not pre-dipped	Pre-dipped	Row marginal total
# cases	18	8	26
# cow months	250	236	486

$$IR = \frac{\left(\frac{18}{250}\right)}{\left(\frac{8}{236}\right)} = 2.12$$

Rate of mastitis is 2.12 time higher in cows whose teats are not pre-dipped than in pre-dipped cows.

Conclusion

Today we worked with 1 exposure and 1 disease both with binary categorical variables (2*2 table) and independant groups, But what if...

You have paired proportions?

You still have a binary outcome but
- your exposure variable has more than 2 levels?
- you have more than 1 exposure variable?
- you have continuous exposure variable(s)?

You have an outcome on a (near) continuous scale? (e.g., Milk production)

Generalized Linear(-mixed) models

Questions ?

http://advertisementfeature.cnn.com/think-brilliant/wrong-question-right-answer.html