Categorical data analysis

MF9130E – Introductory Course in Statistics
28.04.2023

Chi Zhang
chi.zhang@medisin.uio.no
Oslo Center for Biostatistics and Epidemiology
Department of Biostatistics, UiO

Outline

8:30-9:00	Review: proportions, exposure vs outcome, risk ratio, odds ratio, chi-squared test
9:15-9:30	Demonstration in R
Practice	**Practice (exercise 1, 2)**
11:10–11:30	Summary and wrap up

Lab notes for today:
(under *R Lab and Code* tab)

Categorical data analysis

Link to *R Lab and Code*

https://ocbe-uio.github.io/teaching_mf9130e/lab/lab_categorical.html

Categorical data analysis

So far: we have compared 2 groups, continuous measurements (t-test)

What if the data is **in categories**: smoker or not, low birth weight or not

Test whether a **proportion equals a certain value** (z-test)

Different measures of proportions, exposure and outcome (**risk ratio, odds ratio**)

Strength of association between exposure and outcome (chi-squared test)

Proportion (one group)

Example 15.3 KS

In September 2001 a survey of smoking habits was conducted in **a sample of 1000** teenagers aged 15-16, selected at random from all 15-16 year-olds living in Birmingham, UK.

A total of **123** reported that they were smokers.

What is the **proportion** of smokers? What is the 95% confidence interval?

Sample proportion $$p = \frac{123}{1000} = 0.123 = 12.3\%$$

Confidence interval for (sample) proportion

$$CI = \left(p - z' \times \sqrt{\frac{p(1-p)}{n}}, p + z' \times \sqrt{\frac{p(1-p)}{n}}\right)$$

Can also do hypothesis test:

H0: p = 0.5
H1: p != 0.5 (not equal)

(Similar to t.test(), doing a z-test in R returns confidence interval)

```
# in R:
prop.test(x = 123, n = 1000, p = 0.5)
```

Proportion (two groups)

Outcome: getting a disease or not, whether a drug is effective or not

Exposure: how we define the two groups:
exposed / unexposed to X

(changing outcome should NOT change exposure!)

X: ‘**risk factor**’
- sex (men, women)
- drug (treatment, placebo)
- age groups (below 65, above 65)

Risk factor can be continuous too; today we focus on cateogorical (2 categories)

	Experienced outcome?		Total
	Yes	No	
Exposure	D (disease)	H (healthy)	
Group 1 (exposed)	**d1**	**h1**	n1
Group 0 (unexposed)	**d0**	**h0**	n0
Total	d	h	n

Proportion (two groups)

(Example 2 in categorical lab notes)
Lung data (PEFH98-english)

High value of pefmean versus gender
We want to investigate the association between having a high value of pefmean (in 2 categories), with gender

Note: for this variable, we have the continuous (numeric) measurements, so we do not have to use categorical analysis.

The purpose of this example is to show you how to split a continuous variable in 2 categories.

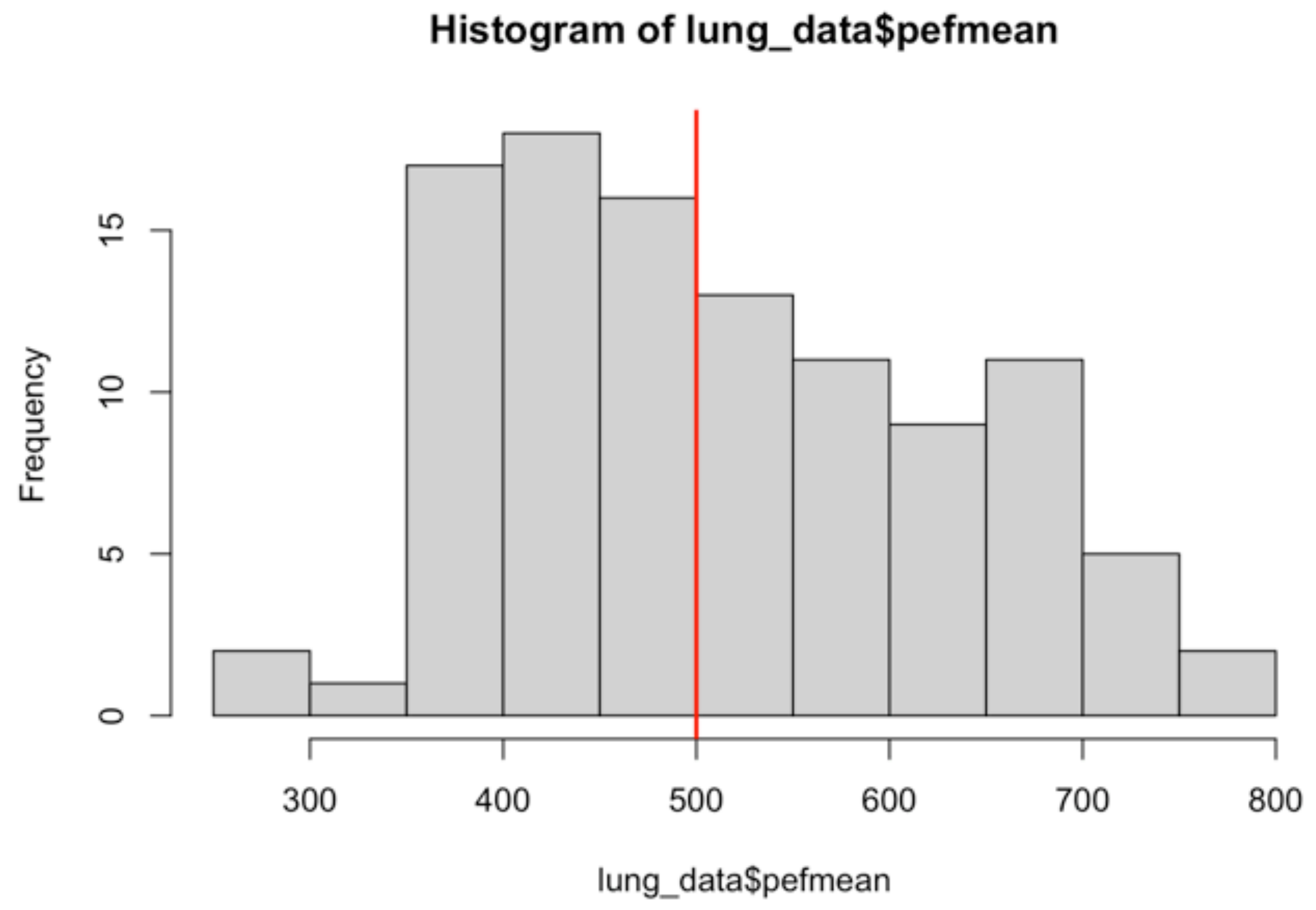
	age	gender	height	weight	pefsit1	pefsit2	pefsit3	pefsta1	pefsta2	pefsta3	pefsitm	pefstam	pefmean
1	20	female	165	50	400	400	410	410	410	400	403.3333	406.6667	405.0000
2	20	male	185	75	480	460	510	520	500	480	483.3333	500.0000	491.6667
3	21	male	178	70	490	540	560	470	500	470	530.0000	480.0000	505.0000
4	21	male	179	74	520	530	540	480	510	500	530.0000	496.6667	513.3333
5	20	male	196	95	740	750	750	700	710	700	746.6667	703.3333	725.0000
6	20	male	189	83	600	575	600	600	600	640	591.6667	613.3333	602.5000
7	32	male	173	65	740	760	720	705	690	680	740.0000	691.6667	715.8333
8	22	male	196	94	720	720	700	700	730	800	713.3333	743.3333	728.3333
9	21	female	173	66	480	530	540	520	520	530	516.6667	523.3333	520.0000
10	23	female	173	65	400	430	420	430	430	430	416.6667	430.0000	423.3333
11	22	female	169	65	500	510	540	520	580	530	516.6667	543.3333	530.0000
12	23	male	185	75	730	630	700	700	700	710	686.6667	703.3333	695.0000
13	21	male	194	84	630	690	670	680	700	690	663.3333	690.0000	676.6667
14	21	female	170	55	360	360	370	370	360	360	363.3333	363.3333	363.3333

Proportion (two groups)

(Example 2 in categorical lab notes)
Lung data (PEFH98-english)

High value of pefmean versus gender
We want to investigate the association between having a high value of pefmean (in 2 categories), with gender

We assume pefmean > 500 is high; otherwise not.



Step 1: understand your data

What is “high value of pefmean”? Where does the **threshold** (500) place in the data distribution?

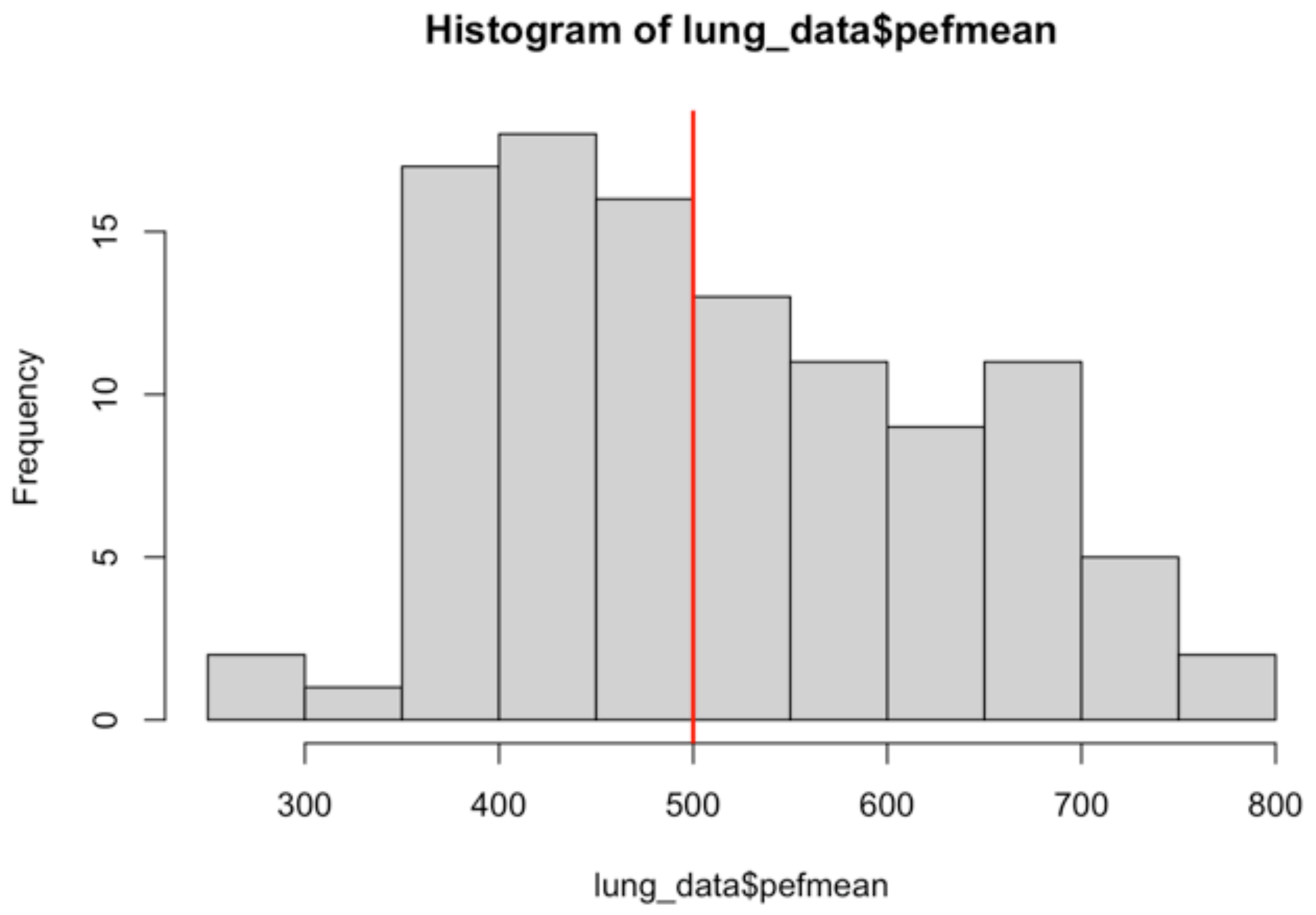
(This red line is the threshold to divide pefmean into 2 groups, NOT sample mean from yesterday!)

Proportion (two groups)

(Example 2 in categorical lab notes)
Lung data (PEFH98-english)

High value of pefmean versus gender
We want to investigate the association between having a high value of pefmean (in 2 categories), with gender

We assume pefmean > 500 is high; otherwise not.



Step 2: split pefmean into 2 groups, **higher than 500**; **not higher than 500**

If we have a new variable called “**high pef**”, the values would be **Yes** or **No**.

Visually, it looks like half people have high pef; half do not have high pef

We can count how many exactly from the data (Yes: 51; No: 54)

Proportion (two groups)

(Example 2 in categorical lab notes)
Lung data (PEFH98-english)

High value of pefmean versus gender
We want to investigate the association between having a high value of pefmean (in 2 categories), with gender

We assume pefmean > 500 is high; otherwise not.

Step 3: what is exposure, what is outcome?

In this case, we can consider **high pef** is the outcome, **gender** as exposure.

Why? (Would having high pef affect gender?)

Step 4: cross tabulation

Count: how many in each of the 4 cells

	High pef yes	High pef no
Male (exposed)	**46**	**6**
Female (unexposed)	**5**	**48**

Proportion (two groups)

Risk ratio

Risk (male) = 46/(46+6) = 0.885
Risk (female) = 5/(5+48) = 0.094

Risk ratio = 0.885/0.094 = 9.37
Males have 9.37 times the “risk” (or probability) of having high pef.

Odds ratio

Odds (male) = 46/6 = 7.667
Odds (female) = 5/48 = 0.104

Odds ratio = 7.667/0.104 = 73.6
The odds of having high pef among males is 73.6 times that of females

	High pef yes	High pef no
Male (exposed)	**46**	**6**
Female (unexposed)	**5**	**48**

Risk ratio is easier to interpret than odds ratio;

Odds ratio is used in logistic regression

RR, OR > 1 means association is positive: being exposed to the risk factor increases the risk of having the outcome (e.g. disease)

Strength of association

We carry out a chi-squared test to assess the strength of association.

It compares the **observed numbers**, and **expected numbers** (under the null hypothesis that there is **no association** between exposure and outcome)

Test statistic: 62.49

Compare test statistic with chi-squared distribution of degress of freedom 1, gives a p-value <0.001

Very strong evidence to reject the null hypothesis (of no association)

Conclude that there is strong association between gender and having high pef.

Observed / **expected**	High pef yes	High pef no
Male (exposed)	**46 (25.26)**	**6 (26.74)**
Female (unexposed)	**5 (26.74)**	**48 (27.25)**

Caution: chi-squared test does not account for what is exposure and what is outcome.
(Why? It computes the difference for all cells, no matter how you arrange it)

Report risk ratio and/or odds ratio, plus p-val from chi-squared test