ECS7024 Statistics for Artificial Intelligence and Data Science

Topic 8: Contingency Tables

William Marsh

Outline

- Aim: Understand how to use a Contingency Table to look at the relationship between categorical variables

- Probability recap: joint and conditional
- Contingency tables (cross tabulation)
- Comparing categorical variables
- Odds and odds ratios
- Comparing distribution (of a continuous variable) for different categories

Context

- Correlation: do two variables 'vary together'?
- Strength of correlation
  - Approach: average of $(x_i - \mu_X)(y_i - \mu_Y)$
  - ... also written $(x_i - \bar{x})(y_i - \bar{y})$ *for sample values*
  - Only applies to continuous variables (with a mean)

- What about categorical variables?

- ANS: look at proportions of one variable, conditional on the values of others

Probability Recap

Recap on Joint and Conditional Probability

What is P(A)?

- Survey of plants
- Plant height (H)
  - Categorical variable
  - Values: tall (t), short (s)
- P(H) is a <u>table</u>

Tall	7/12
Short	5/12

- How do we estimate probabilities from data?

Probabilities Come from Counts

- Counts in each category

Plant	Height
100	tall
101	tall
102	short
103	tall
104	short
105	short
106	short
107	tall
108	short
....	

data: records of 60 plants

counting by category →

tall	35
short	25

Total = 60

divide by total →

tall	7/12
short	5/12

Total = 1

table of probabilities (sample)

Joint Probability P(H, L)

- Counts in two Categories

Leaves (L)	Stem Height (H)	
	Short (s)	**Tall (t)**
Broad (b)	9	21
Long (g)	16	14

Still 60 plants in total

- Probabilities

P(H,L) is also a table

Leaves (L)	Stem Height (H)	
	Short (s)	**Tall (t)**
Broad (b)	9/60	21/60
Long (g)	16/60	14/60

Total is 1

Conditional Probability

- Look at some entries in the table
  - The short ones

Leaves (L)	Stem Height (H)	
	Short (s)	Tall (t)
Broad (b)	9/60	21/60
Long (g)	16/60	14/60

25 short plants

- P(L | H = short)

broad	9/25
long	16/25

probability of leaf types given the plant is short

Conditional Probability

- Look at some entries in the table
  - The short ones or the tall ones

Leaves (L)	Stem Height (H)	
	Short (s)	Tall (t)
Broad (b)	9/60	21/60
Long (g)	16/60	14/60

25 short plants

35 tall plants

- P(L | H = short)

broad	9/25
long	16/25

*probability of leaf types given the plant is **short***

P(L | H = tall)

broad	21/35
long	14/35

*probability of leaf types given the plant is **tall***

Conditional Probabilities Table

- P(L|H) – a table of tables
  - Each column is a probability distribution

Leaves (L)	Stem Height (H)	
	Short (s)	Tall (t)
Broad (b)	9/25	21/35
Long (g)	16/25	14/35

P(L | H = short)

P(L | H = tall)

total is 1

total is 1

Probabilities and Tables: Summary

- Assume variable with two categories (binary)
  - Generalises to more categories

Probability	Table Entries	Total
P(H)	2: short & tall	1
P(L)	2: broad & long	1
P(H, L)	4: (s, b) & (s, g) & (t, b) & (t, g)	1
P(L, H)	Same table as P(H, L)	
P(H | L)	Broad column: s & t Long column: s & t	1 1
P(L | H)	Short column: b & g Tall column: b & g	1 1

Relationship between Joint and Conditional Probabilities

Contingency Tables

Recall Rule for P(A,B)

$$P(A,B) = P(A).P(B|A)$$

$$P(A,B) = P(B).P(A|B)$$

- *The joint probability of A & B is the probability of A multiplied by the probability of B given A*
- *(also the other way around)*

P(H).P(L|H) Using Tables

p(H)

tall	7/12
short	5/12

p(L | H = tall)

broad	21/35
long	14/35

p(L | H = short)

broad	9/25
long	16/25

p(L , H)

tall	broad	7/12 x 21/35 = 21/60
tall	long	7/12 x 14/35 = 14/60
short	broad	5/12 x 9/25 = 9/60
short	long	5/12 x 16/25 = 16/60

Marginalisation

- Going from p(H, L) to p(H) or p(L)

Leaves (L)	Stem Height (H)	
	Short (s)	Tall (t)
Broad (b)	9/60	21/60
Long (g)	16/60	14/60

Add up value with same height

	25/60	35/60

$$p(H) = \sum_{l \in \{\mathrm{b,g}\}} p(H, L = l)$$

$$p(L) = \sum_{h \in \{\mathrm{s,t}\}} p(H = h, L)$$

Independence

If two variables A, B are independent then:

- p(A | B) = p(A)
  - *knowing B makes not difference to what A to expect*
- p(B | A) = p(B)
  - *Same the other way around*
- p(A, B) = p(A).p(B)
  - *The joint probability is given by the product*

- Dependent?
  - If p(A | B=b1) differs from p(A | B=b2)

Quiz 1

Python Nugget

Keyword parameters

Keyword Parameters

- Function parameters can be given by
  - Position
  - Keyword
- Keywords have a default value
- Also allows 'extra' parameters

```
def addxy(x, y=1):
  return x + y

addxy(2, 3) # result 5
addxy(2, y=3) # result 5
addxy(2)      # result 3 - default
addxy(y = 3) # error
```

Examples of Keywork Arguments

pandas.DataFrame

class `pandas.DataFrame`(*data=None, index=None, columns=None, dtype=None, copy=False*) [source]

Two-dimensional, size-mutable, potentially heterogeneous tabular data.

Data structure also contains labeled
thought of as a dict-like container fo

Parameters: **data** : *ndarray (struc*

Dict can contain
insertion-order.
Changed in versio

index : *Index or arra*

Index to use for
no index provide

columns : *Index or a*

Column labels to
provided.

dtype : *dtype, defau*

Data type to for

copy : *bool, default*

Copy data from

pandas.DataFrame.plot

`DataFrame.plot`(**args, **kwargs*)

Make plots of Series or DataFrame.

Uses the backend specified by the option `plotting.backend`. By default, matplotlib is used.

Parameters: **data** : *Series or DataFrame*

The object for which the method is called.

x : *label or position, default None*

Only used if data is a DataFrame.

y : *label, position or list of label, positions, default None*

Allows plotting of one column versus another. Only used if data is a DataFrame.

kind : *str*

The kind of plot to produce:

- 'line' : line plot
- 'bar' : vertical b

****kwargs**

Options to pass to matplotlib plotting method.

Cross Tabulation

Contingency Tables

Generalising to More Variables

- Contingency tables can show counts over many categories

Leaves (L)	(Seed) Size (Z)	Stem Height (H)	
		Short (s)	Tall
Broad (b)	Large (l)	4	7
	Medium (m)	3	10
	Small (s)	2	4
Long (l)	Large (l)	7	3
	Medium (m)	5	4
	Small (s)	5	6

$$p(L=b, H=t, Z=m) = 10/60$$

Probability tables given by ‘normalising’

Contingency Table Summary

- Also called ‘cross tabulation’
  - Pandas has ‘crosstab’
- Table shape not fixed
  - Can have multiple variables on each axis
- Table contains
  - Counts

Comparing Categories

Heart Data Again

Heart Disease: Categorical

Variable	Meaning	Type
Sex	1 = male, 0 = female	Categorical
ChestPain	The chest pain experienced (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic)	Categorical
Bsugar	The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)	Binary
Angina	Exercise induced angina (1 = yes; 0 = no)	Binary
RestECG	Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable e left ventricular hypertrophy)	Ordinal (?)
ECG_ST_slope	The slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: downsloping)	Categorical
Vessels	The number of major vessels (0-3) coloured by fluoroscopy	Ordinal
Thallium	Thallium update test (0 = normal; 1 = fixed defect; 2 = reversible defect)	Categorical
Disease	Heart disease (0 = no, 1 = yes)	Binary

Questions: Do They Vary Together?

- Two categorical variables
  – Does knowing one change value the distribution of the other

- ‘Dependent’ or ‘response’ variable: Disease
  – Often one of the variables

Cross Tabulation Example

- 'Crosstab' counts occurrences in categories
- Can 'normalise'
- Results are joint or conditional probability distributions

p(Sex, Disease)

Sex	M	F
Disease		
False	0.30	0.24
True	0.38	0.08

'Normalise all'

p(Disease | Sex)

Sex	M	F
Disease		
False	0.44	0.74
True	0.56	0.26

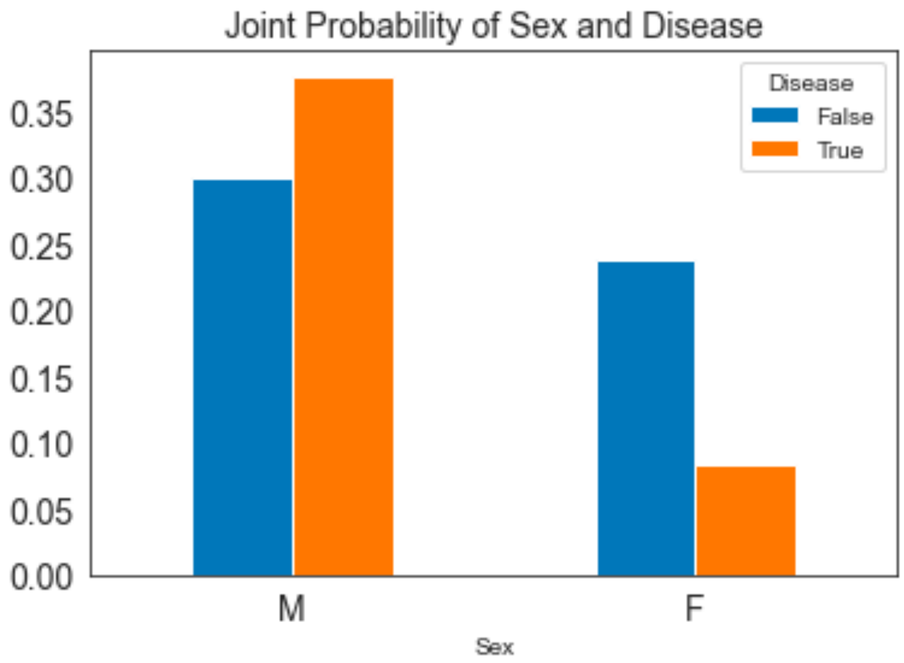
'Normalise columns'

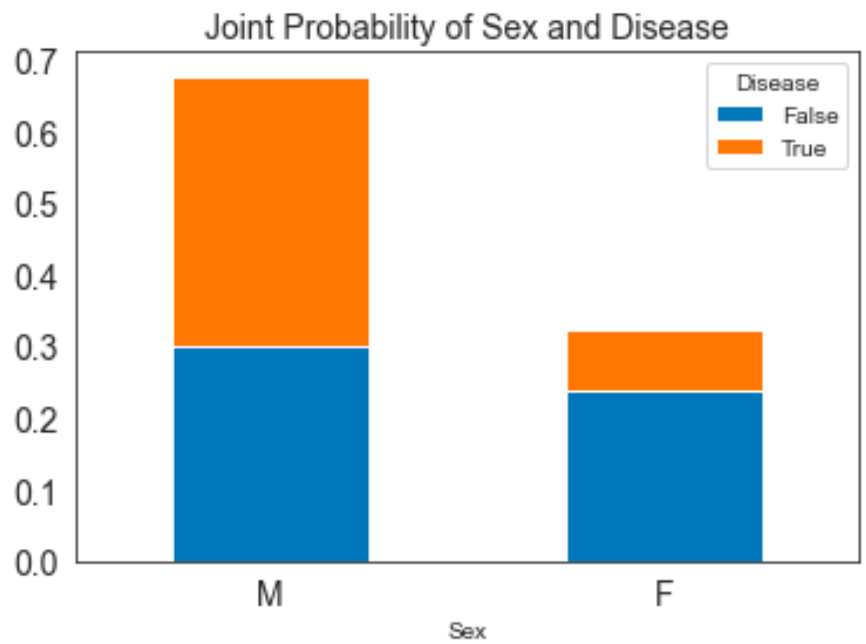
Cross Tabulation Example

- Joint probability

  p(Sex, Disease)

Sex	M	F
Disease		
False	0.30	0.24
True	0.38	0.08



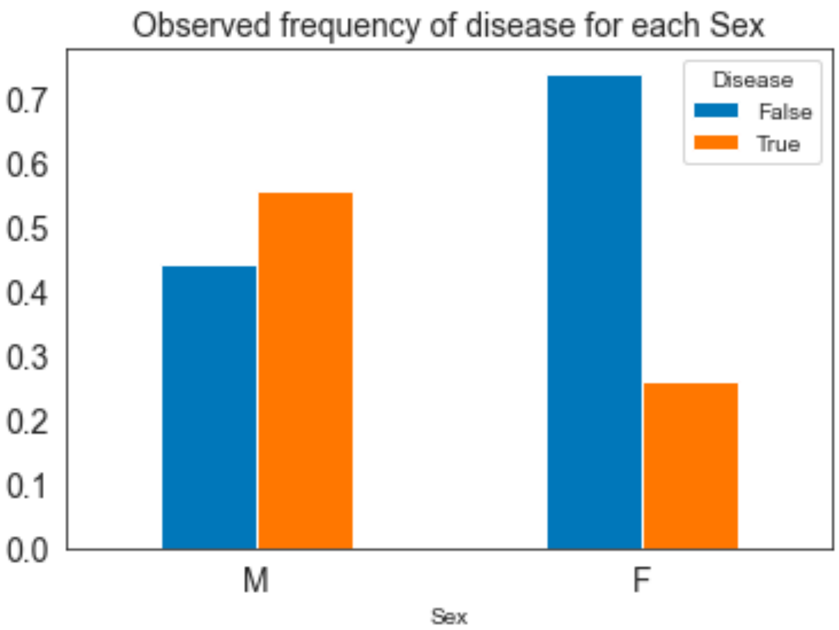


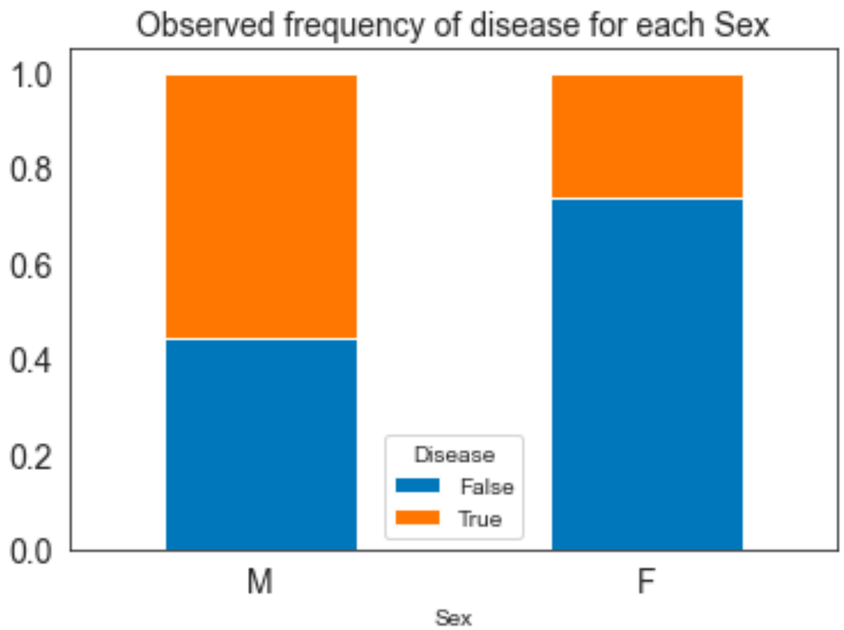
Cross Tabulation Example

- Conditional probability

$$p(\text{Disease} \mid \text{Sex})$$

Sex	M	F
Disease		
False	0.44	0.74
True	0.56	0.26



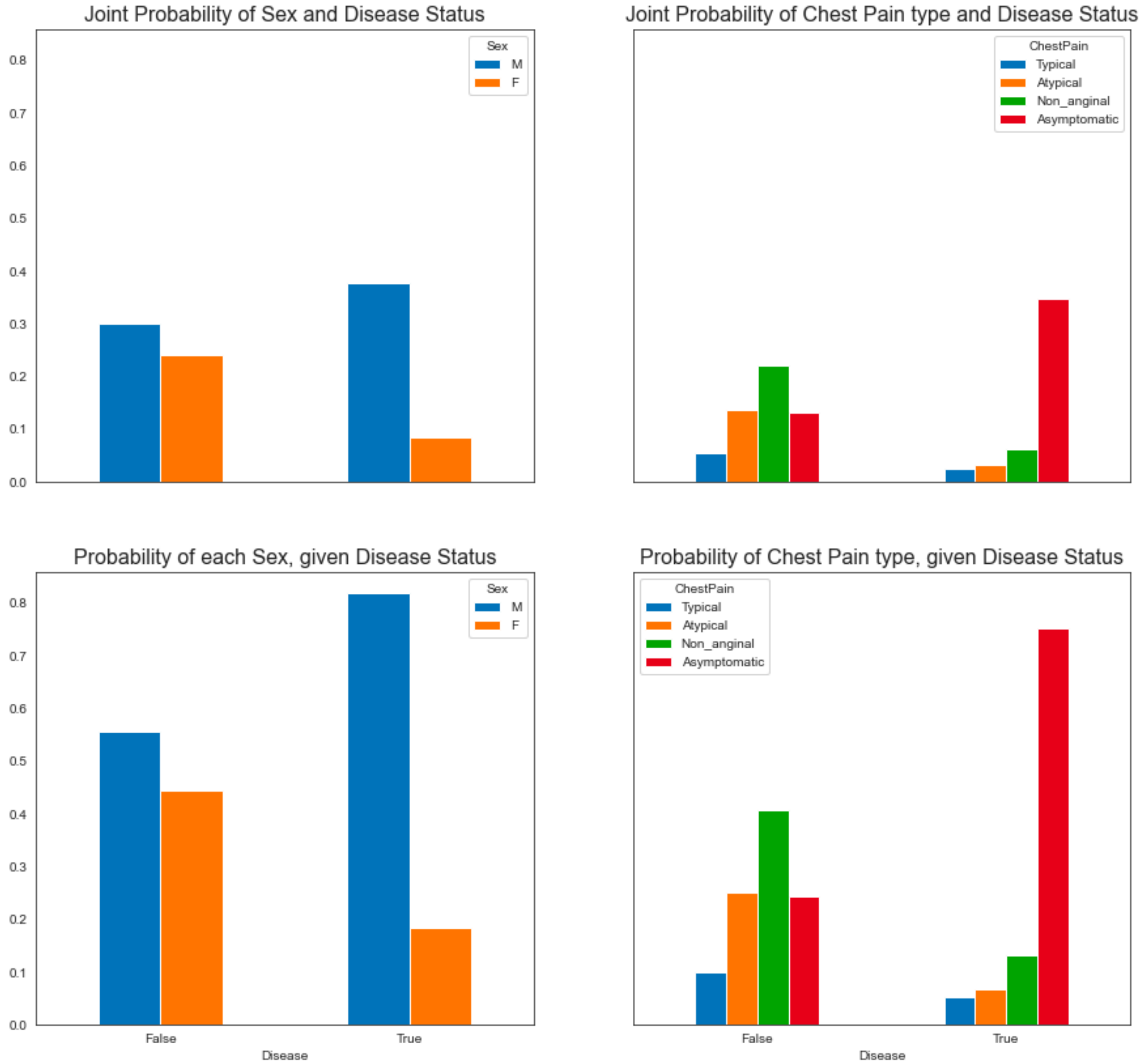


Which Conditional Probability?

Joint distribution: easy to understand, but harder to see ‘association’

p(? | Disease)

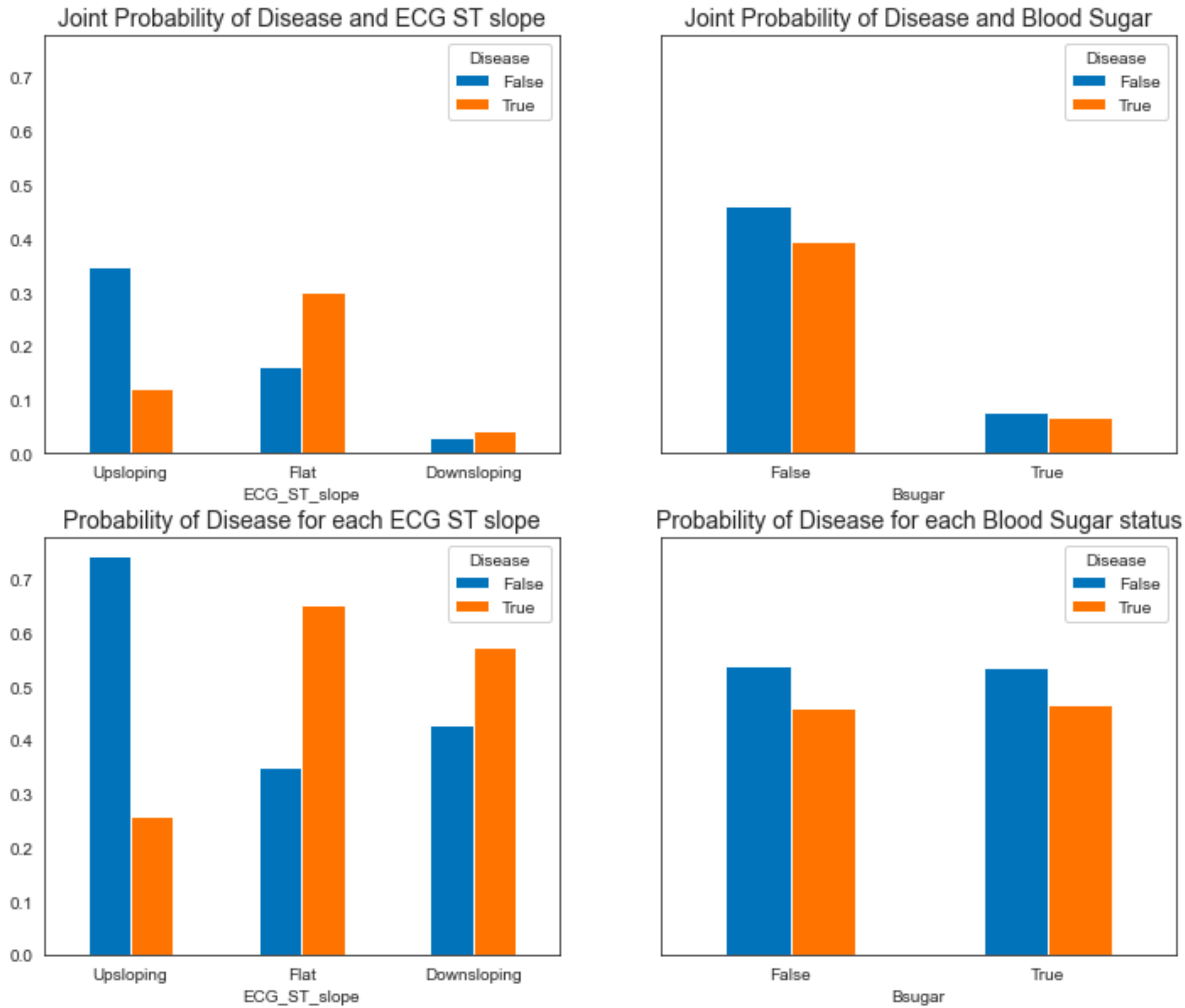
Is this the best way around?



Prefer: p(Disease | ?)

Joint

Conditional



Probability and Odds

Odds is Another Way to Write a Probability

- Two rules of probability
  - $0 \leq p(A) \leq 1$
  - $p(A) + p(\text{not } A) = 1$ (we write 'not A' as $\bar{A}$)
- Definition of odds: $o_A = {}^{p(A)}/_{p(\bar{A})}$
  - Odds ranges from zero upwards
  - $o_{\bar{A}} = {}^{1}/_{o_A}$ so that $o_A . o_{\bar{A}} = 1$
- Example: p(A) = 75% then odds_A = 75/25 = 3
  - Odds > 1 implies probability > 50%
  - Odds < 1 implies probability < 50%

Odds / Odds Ratio of Heart Disease

Sex	M	F
Disease		
False	0.44	0.74
True	0.56	0.26

- From sample probabilities
  - $\text{odds}_{\text{Disease} \mid M} = 56/44 = 1.27$
  - $\text{odds}_{\text{Disease} \mid F} = 26/74 = 0.35$
- Observation:
  - Men have increased chance of heart disease
  - Odds ratio: $\text{odds}_{\text{Disease} \mid M} / \text{odds}_{\text{Disease} \mid F} = 3.6$
- ‘Odds ratio’ can measure strength of one variable on another (for binary variables)

Quiz 2

Comparing Continuous Distributions for Categories

Heart Disease: Continuous

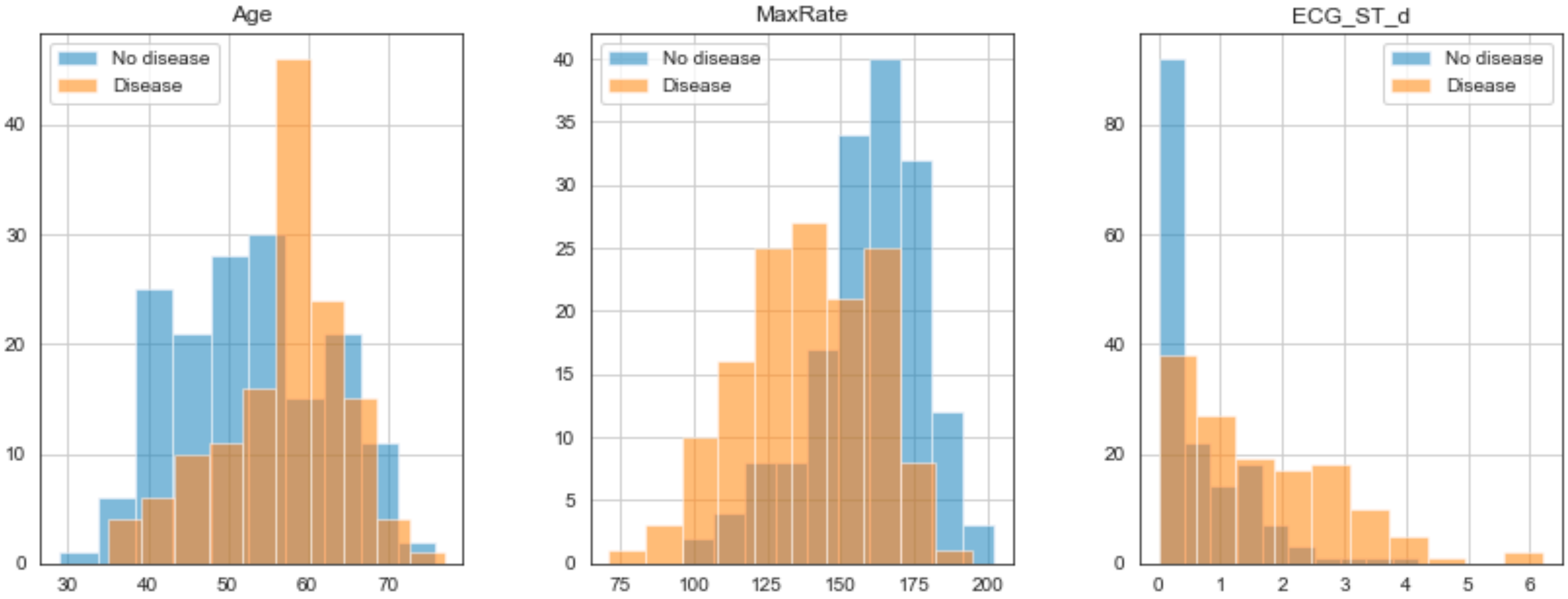
Variable	Meaning	Type
Age	The person's age in years	Continuous
RestBP	The person's resting blood pressure (mm Hg on admission to the hospital)	Continuous
Chol	The person's cholesterol measurement in mg/dl	Continuous
MaxRate	The person's maximum heart rate achieved	Continuous
ECG_ST_d	ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot)	Continuous

Question: Do They Vary Together?

- A categorical variable and a continuous one
  - Is the continuous variable distribution different for the different categorical values

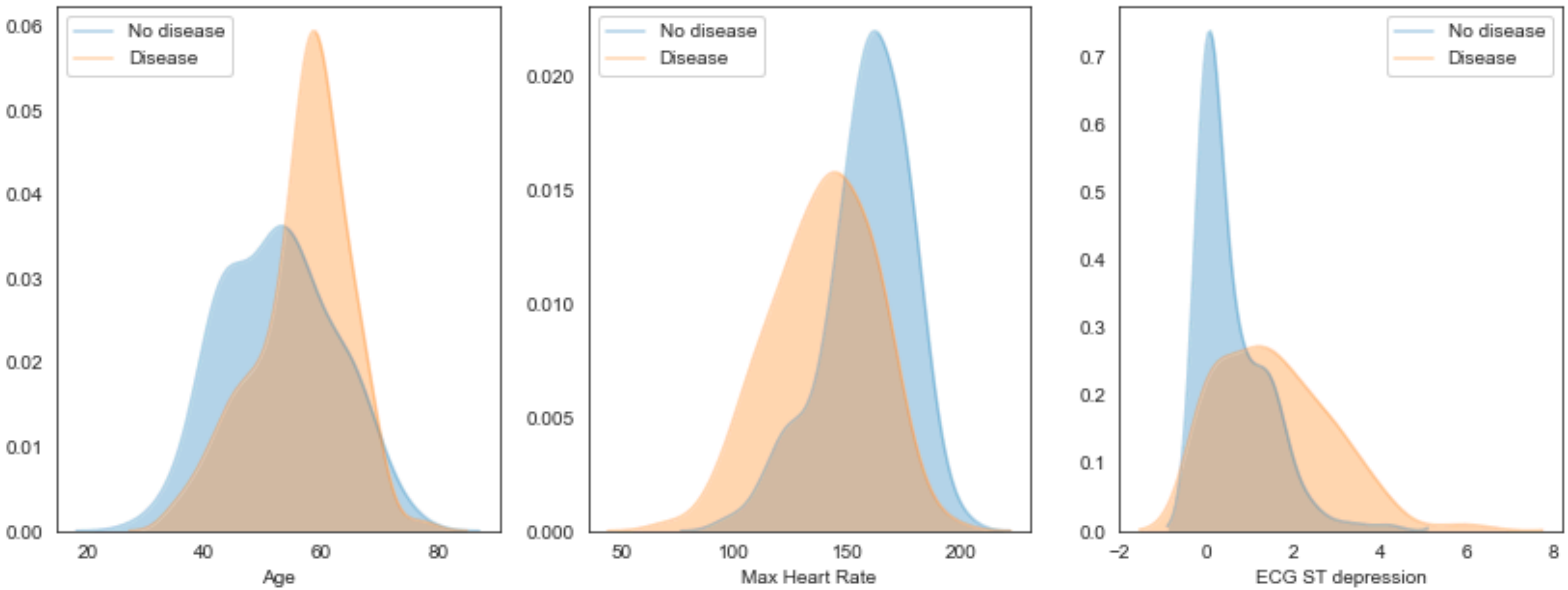
- Example: look at the different distribution of 3 continuous variables by disease status
  - Histograms
  - Kernel density
  - Boxplot

Histograms

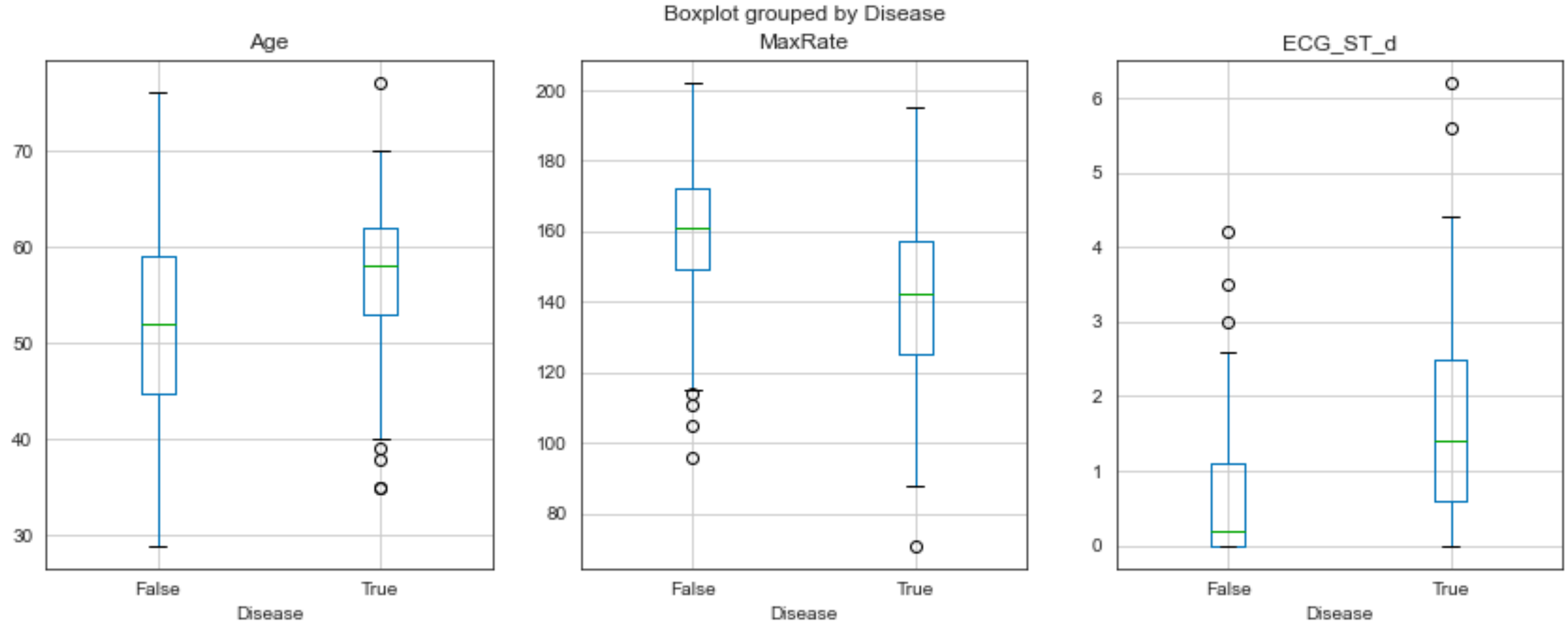

Age
No disease
Disease
0
10
20
30
40
30
40
50
60
70
MaxRate
0
5
10
15
20
25
30
35
40
75
100
125
150
175
200
ECG_ST_d
0
20
40
60
80
0
1
2
3
4
5
6

Using Kernel Density

- Kernel Density Estimator (KDE): smoothed histogram



BoxPlots



Boxplot grouped by Disease
Age
MaxRate
ECG_ST_d
70
60
50
40
30
200
180
160
140
120
100
80
6
5
4
3
2
1
0
False
True
Disease

Quiz 3

Summary

- Comparing categorical variables
  - Look at conditional probability or odds ratio
- Contingency tables
  - Counting, by category
  - Normalised to give probabilities (joint or conditional)
- Which conditional?
  - Very easy to get confused – label clearly
- Continuous by category
  - use boxplot, KDE or histogram
- *Future: are the differences we see 'reliable'?*