

ECS7024 Statistics for Artificial Intelligence and Data Science

Topic 3: Continuous Distributions

William Marsh

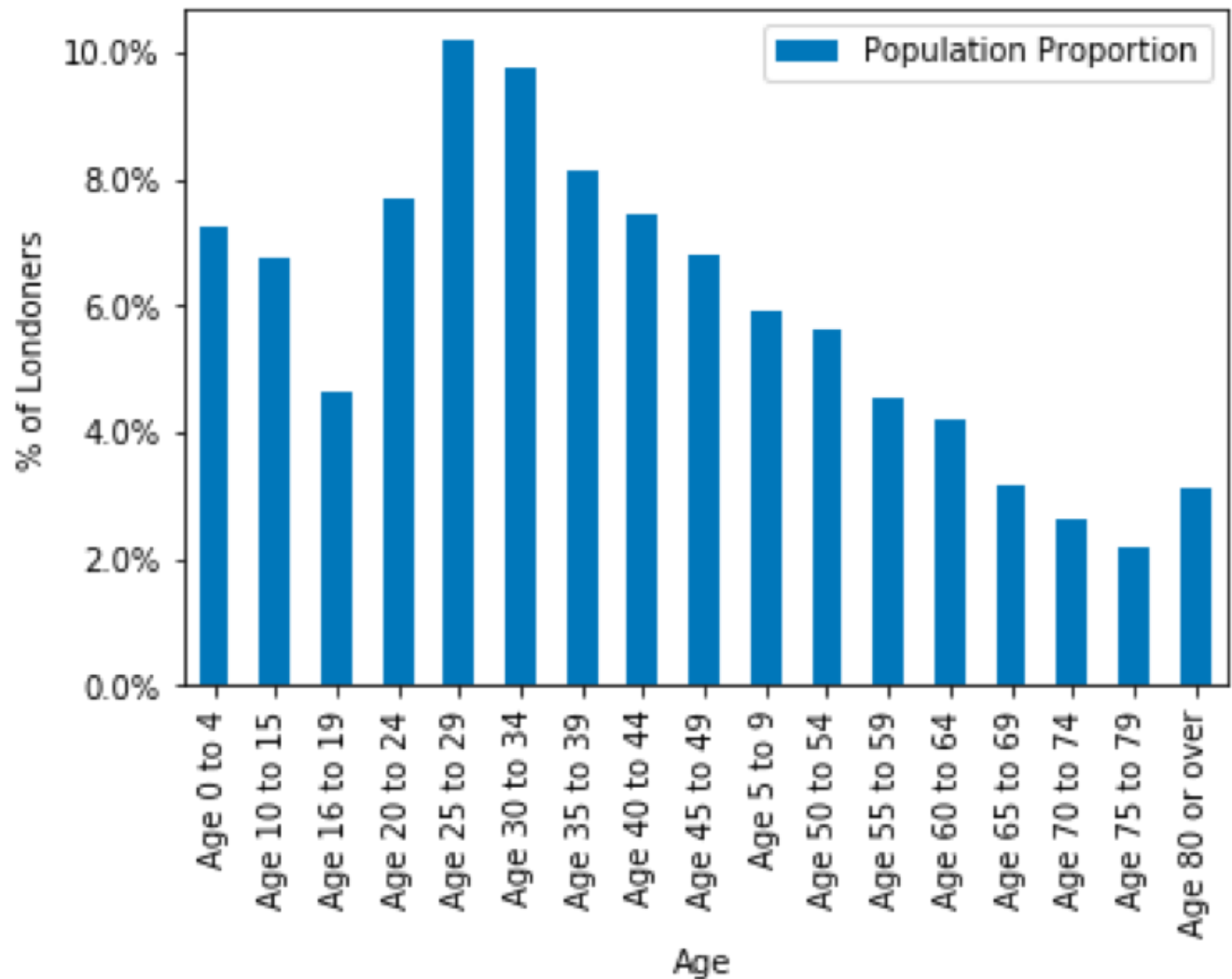
Outline

- Aim: Understand how to look at the distribution of a continuous variable
- Continuous distribution
 - Problem: how it differs from the discrete case
 - Solutions: histograms and density plot
- Relationship to probability
- Mean, median and mode
- Quantiles
- Skew

Continuous Distributions

Recap: Categorical Distribution

- Randomly select a Londoner (in 2011):
 - 10% are in age range 25-29
 - Probability of 10% of this age range
- Distribution shows probability, proportion or frequency



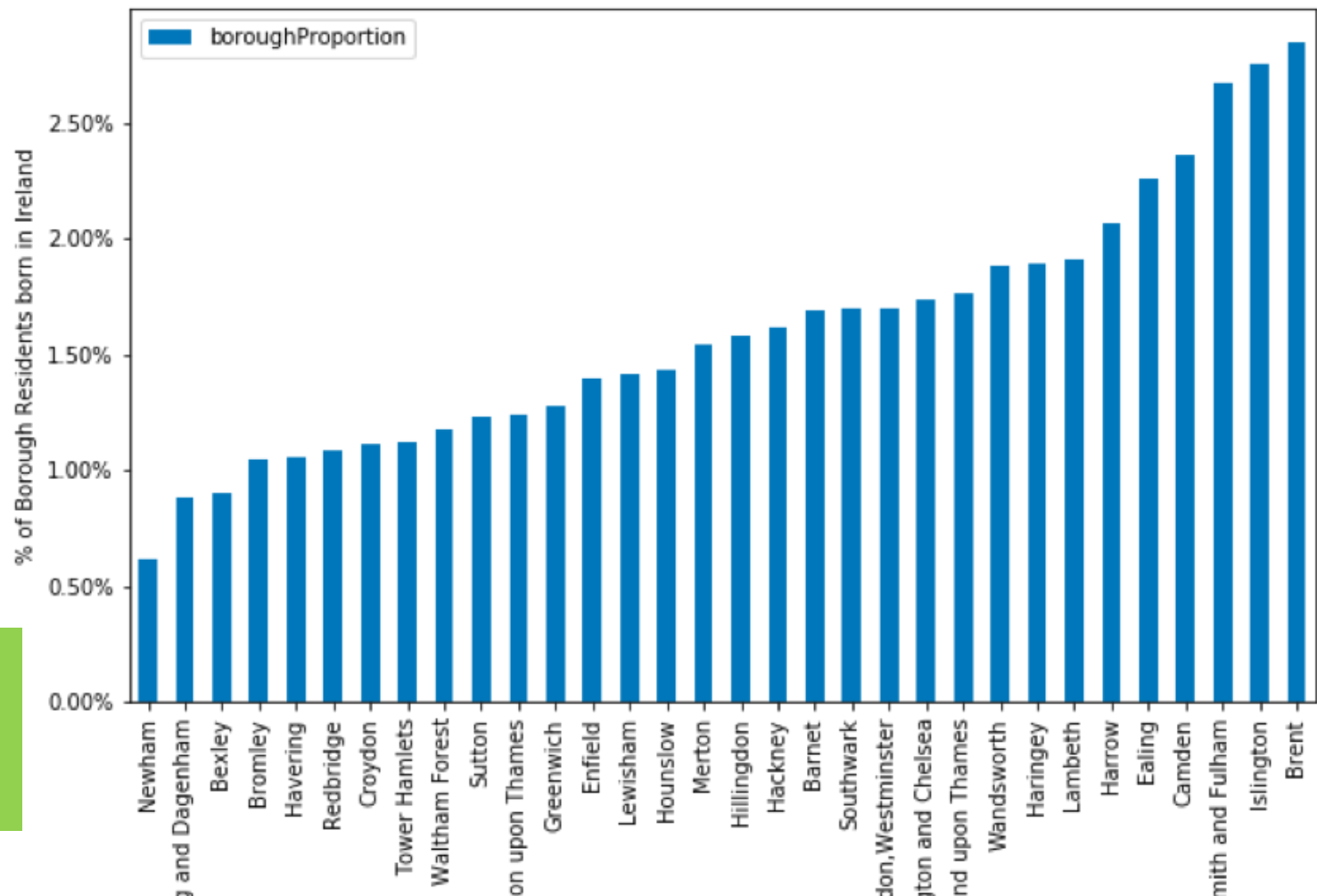
Problem of Continuous Proportion

- Problem
 - Infinite number of values → infinitesimal proportions
 - Age: 25 years, 4 months, 12 days, 8 hours, 3 minutes
 - Not many same age
- Solution 1: Histogram
 - Proportion in a range
 - ... in the data, Age already represented as ranges
- Solution 2: Density plot

Histograms

Bar Chart: % of Borough Residents born in Ireland

- Percent is a continuous value
- What percentages are more common?
 - New variable: '% Percent of Borough Residents born in Ireland'



Bar chart order:

- Increasing
- Not significant

Histogram of 'Ireland %'

- 10 bins

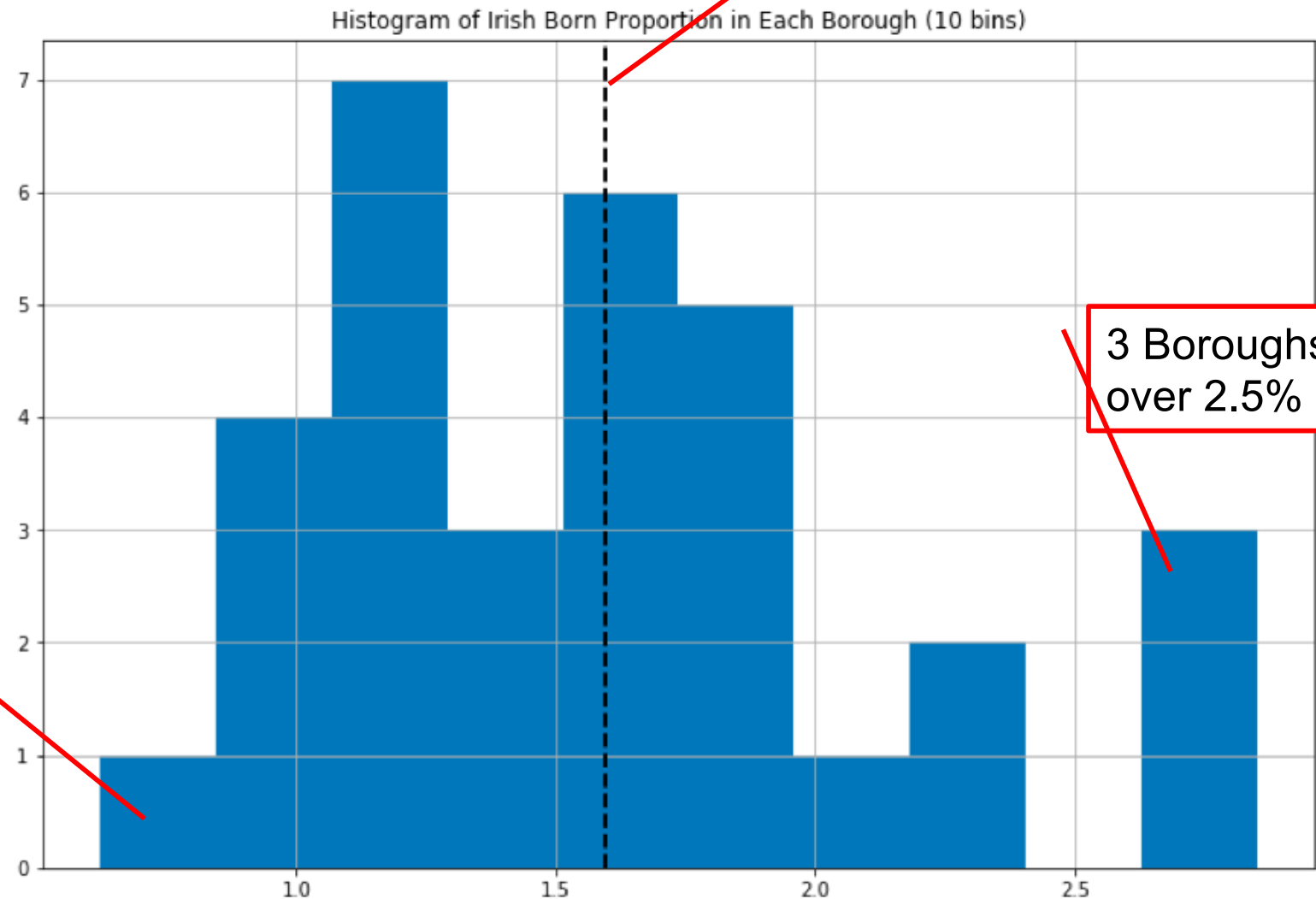
Axis is
frequency or
count

- Divide by
number of
boroughs
to get
proportion

One borough
(Newham)

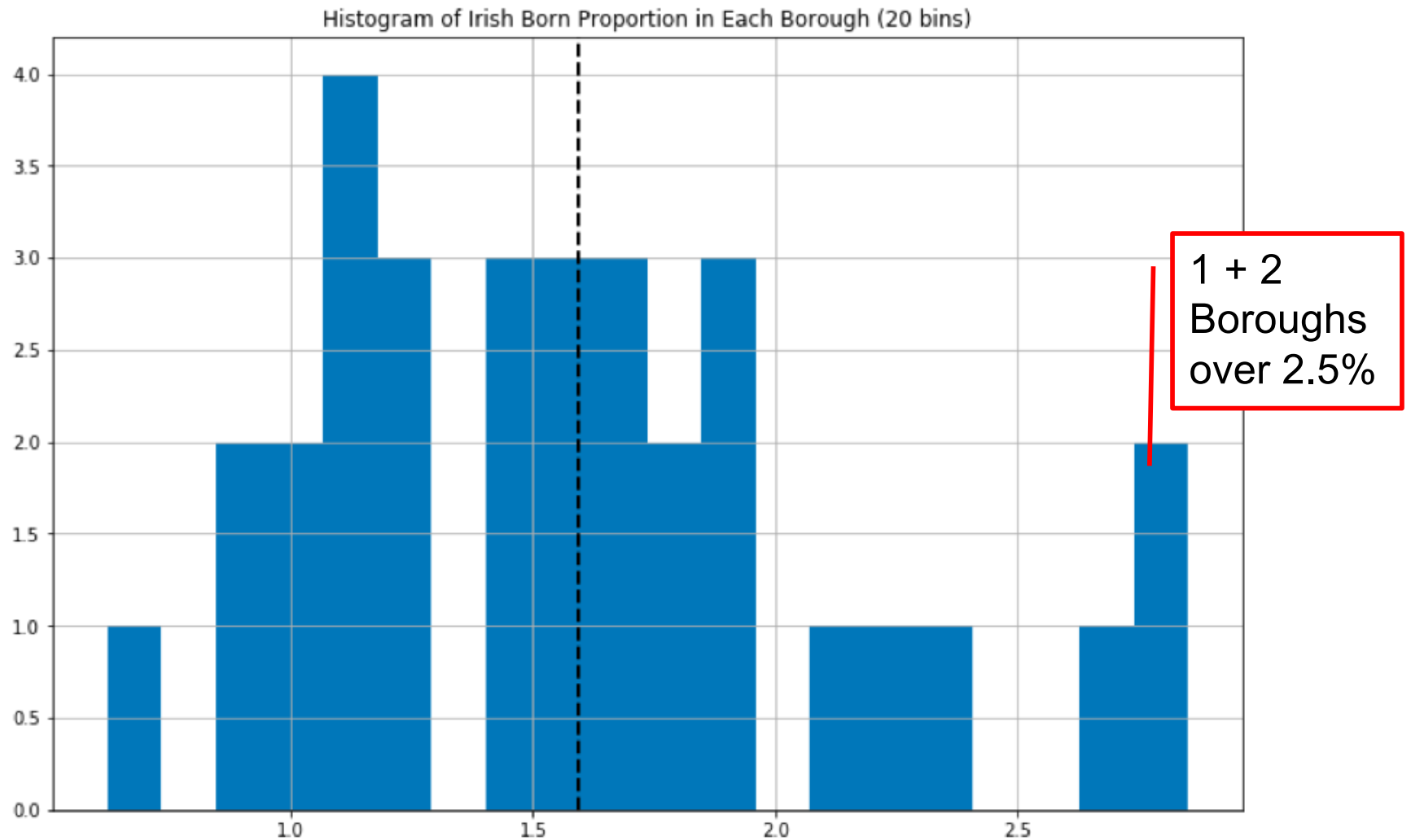
Average:
approx. 1.6%

3 Boroughs
over 2.5%



Histogram of 'Ireland %'

- 20 bins; shape varies with number of bins



Bar Chart versus Histogram

- Histogram appears similar to a bar chart

| | Bar Chart | Histogram |
|-------------------------------|--|---------------------|
| Horizontal axis | Categorical variable | Continuous variable |
| Gaps between columns | Yes | No |
| Order of (x-axis) significant | No | Yes |
| Vertical axis | Proportion (probability), count or frequency | |

Quiz 1

Every lecture will have a 'learning reflection' slide

Learning Programming

How does learning programming
compare with your previous study?

Learning Programming and Pandas

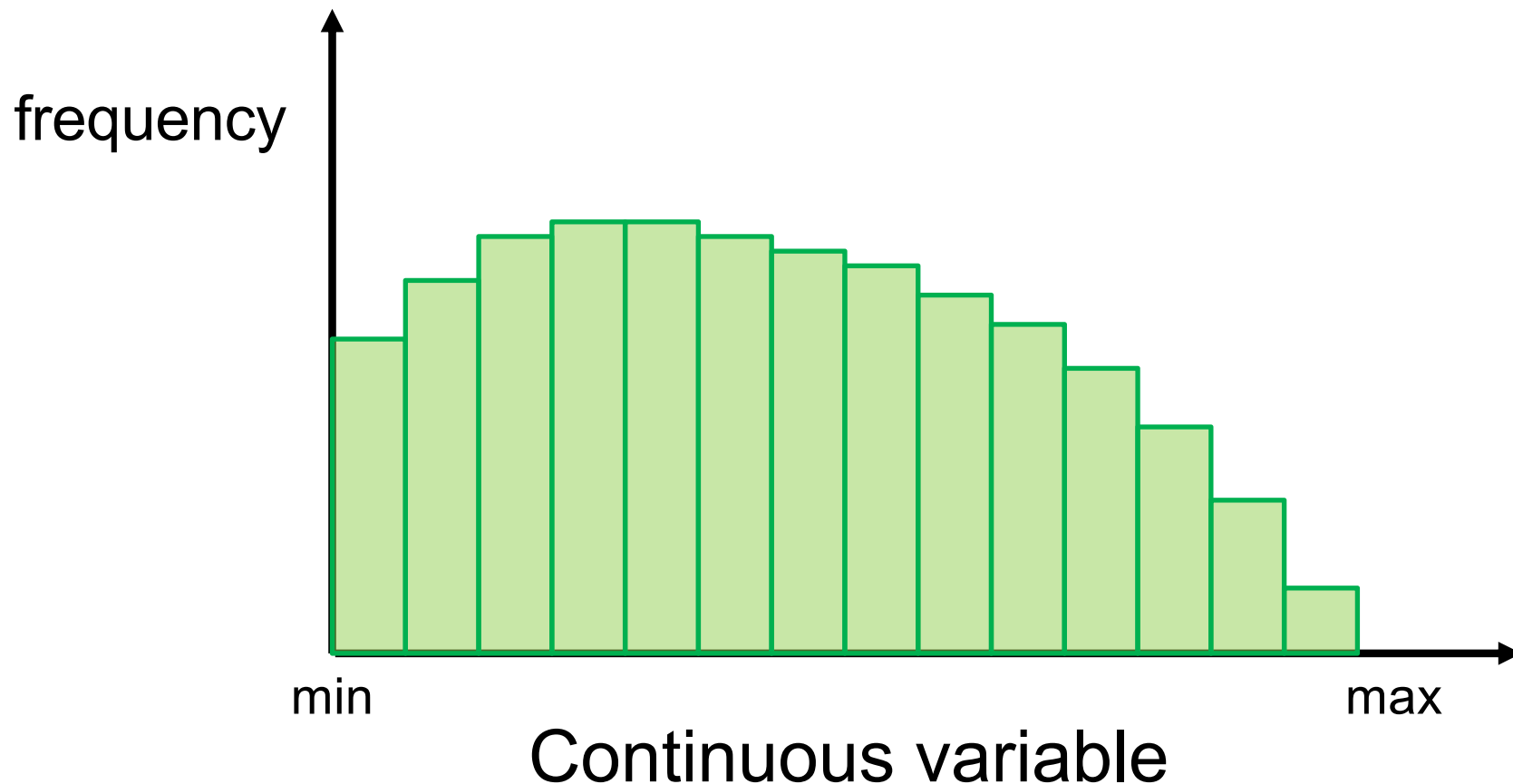
- Mental model
 - Misconceptions
 - Barrier concepts
- Working with example
 - Complex: 'top-down'
 - Read it – what does it do?
 - Change it (a bit)
 - Apply it to a new problem
- Learn to read the error messages
 - Extract some information!
- Large libraries
 - Not expected to memorise
 - Too many alternatives
 - Learn to read documentation
- Two tasks
 - Data analysis: what are the steps? [Decomposition]
 - Writing and planning
 - Programming: how to achieve it?
- Perfect code problem

Density Plot

Histogram can be thought of as an approximation to a density plot

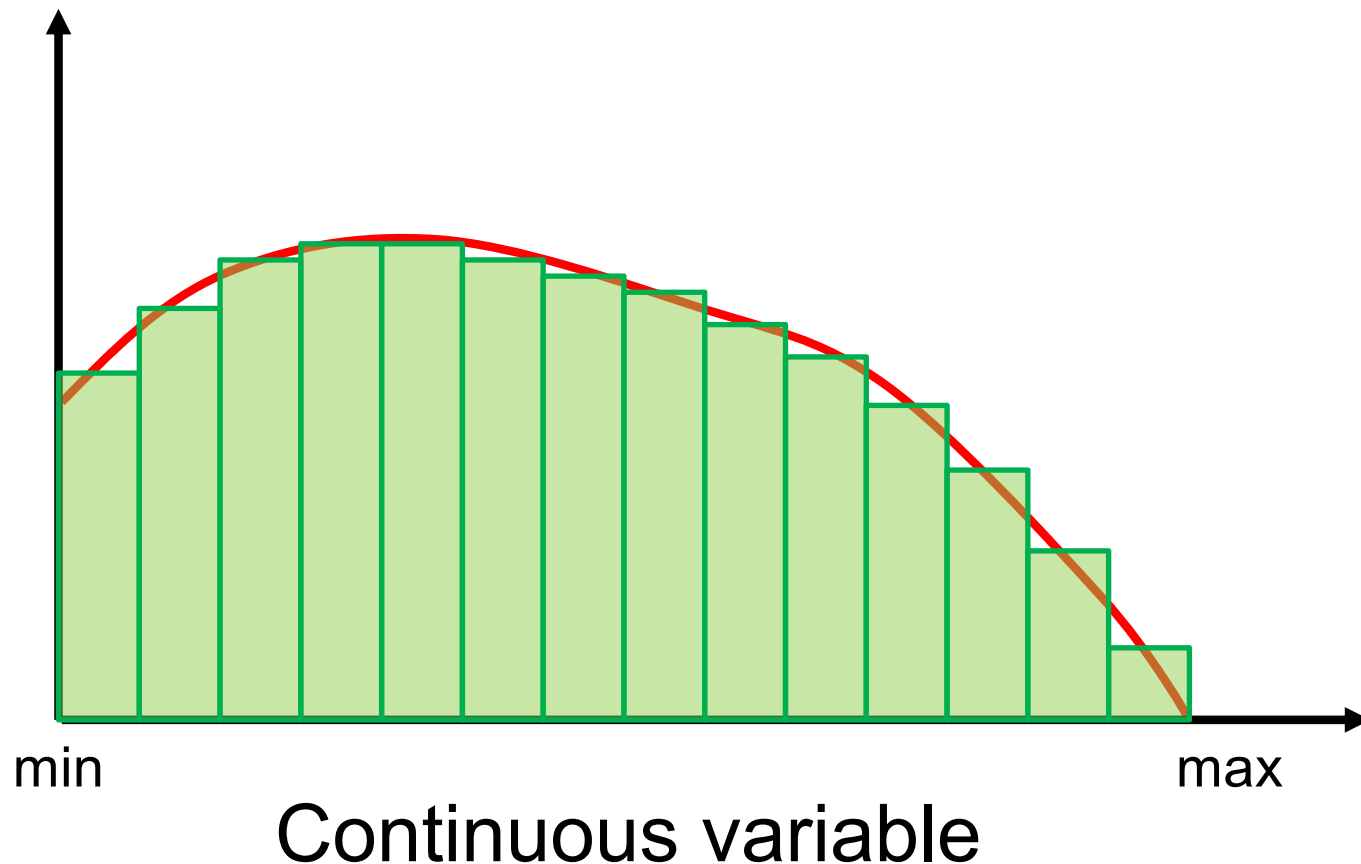
Histogram to Probability Density Plot

- Start with a histogram



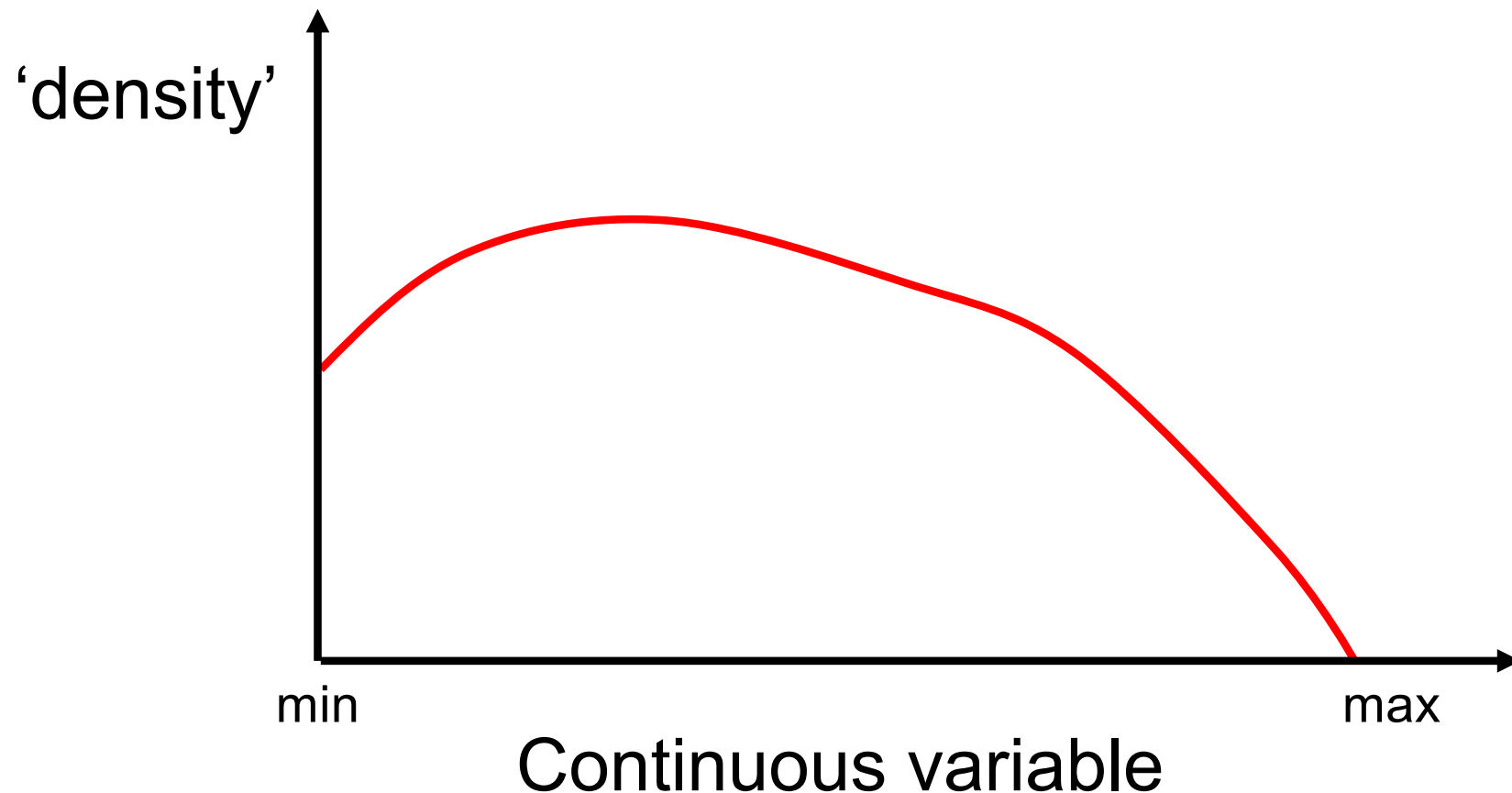
Histogram to Probability Density Plot

- Add a curve



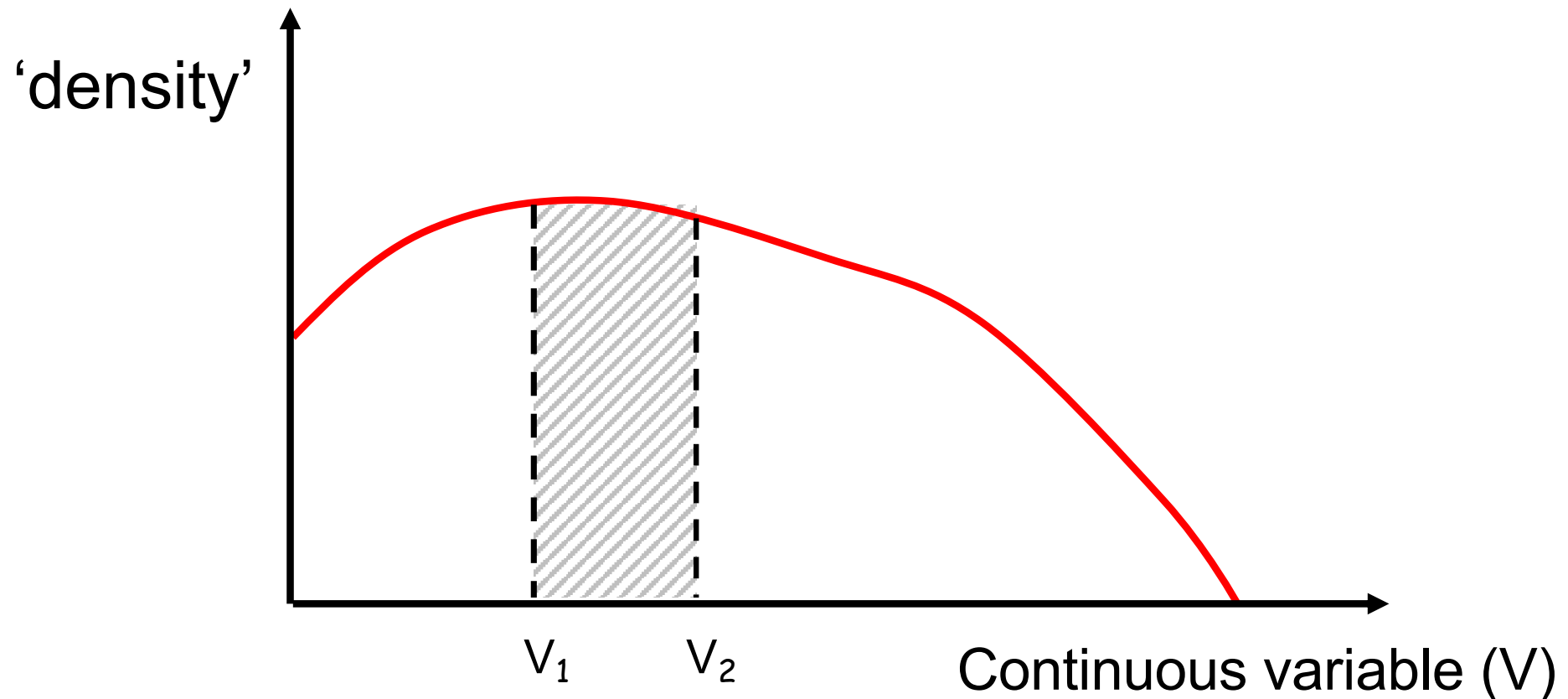
Histogram to Probability Density Plot

- Just look at the curve



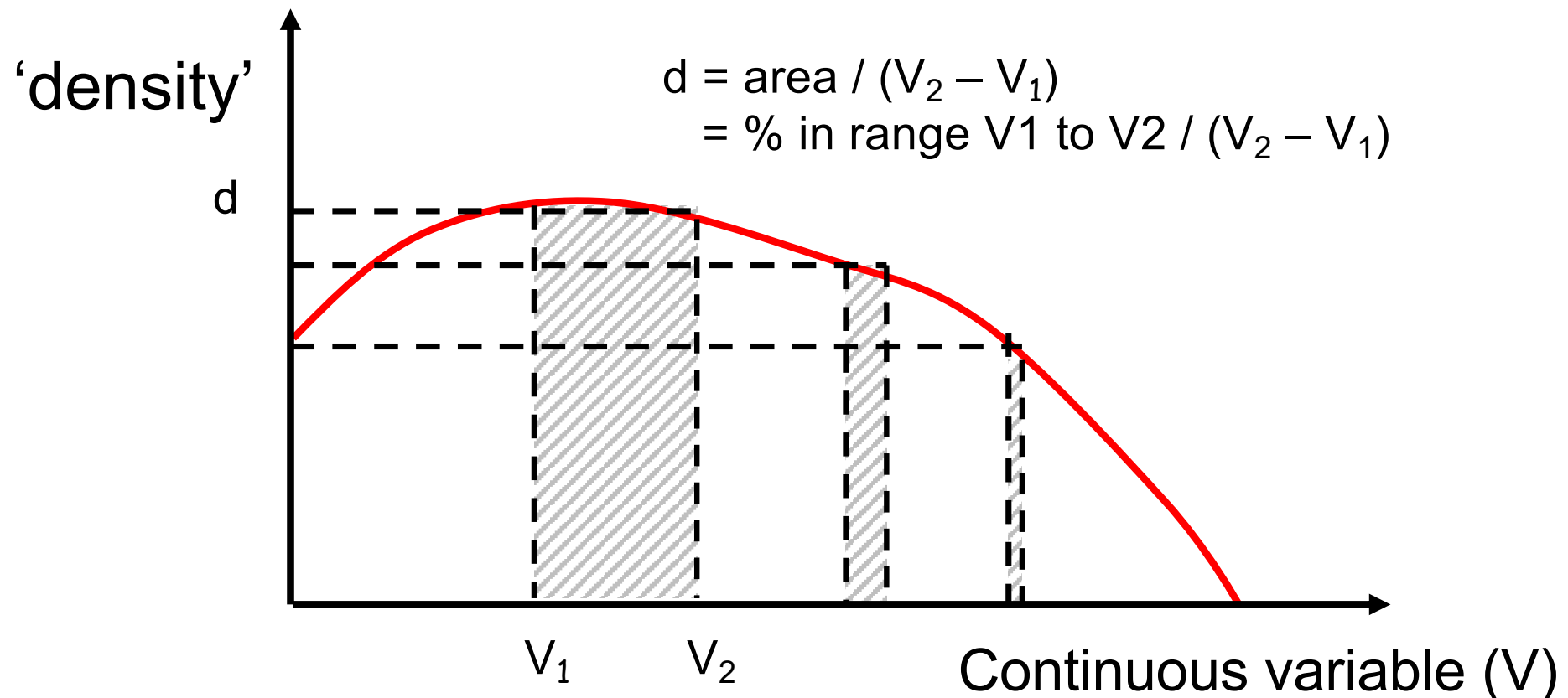
Histogram to Probability Density Plot

- Probability V in range V_1 to V_2
 - Given by shaded area
 - Total area is one (100%)

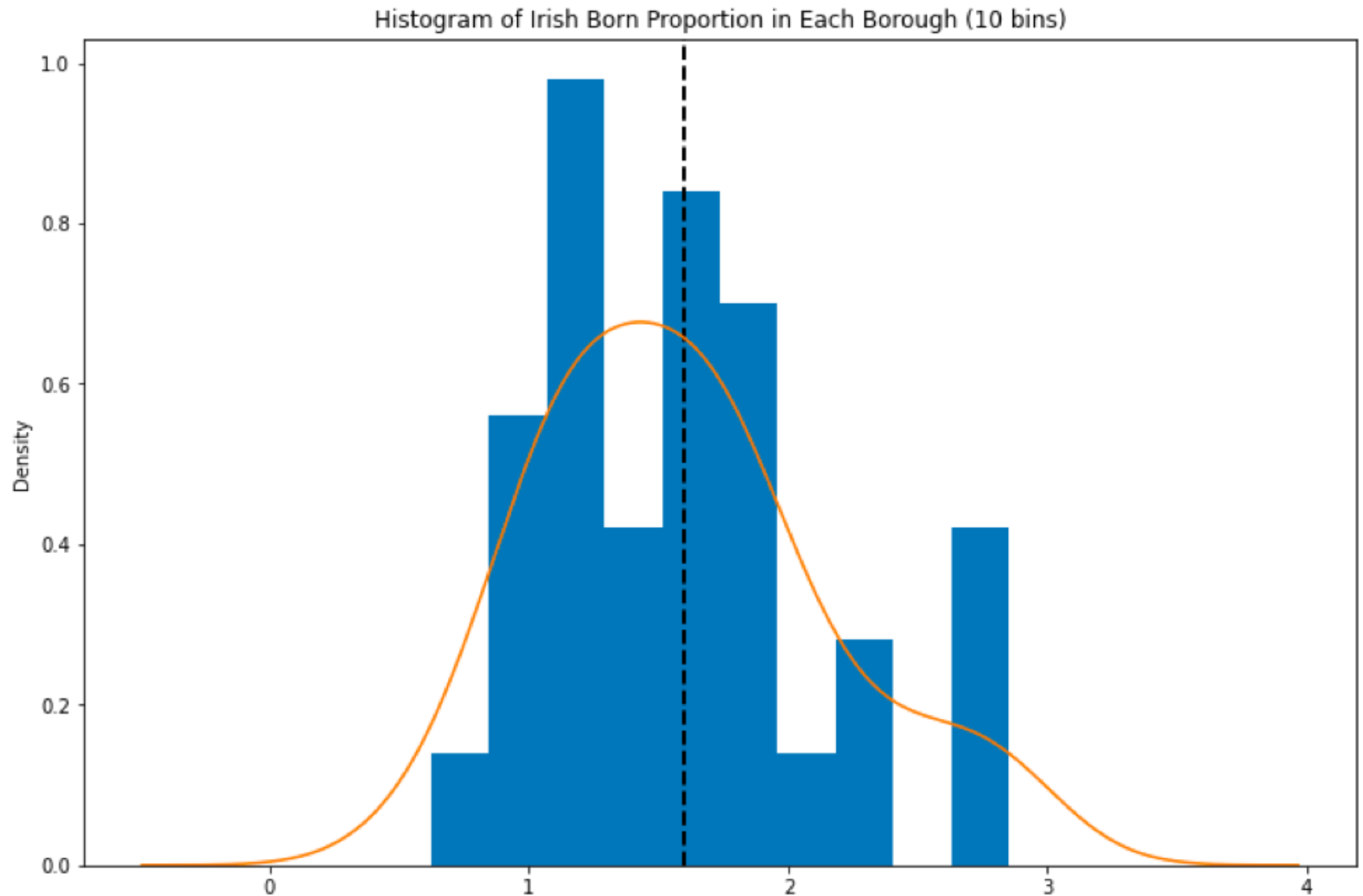


Histogram to Probability Density Plot

- Vertical axis is 'density'
- Curve is the 'Probability density function'



Estimating a Density



Quiz 2

Averages (Continuous Distributions)

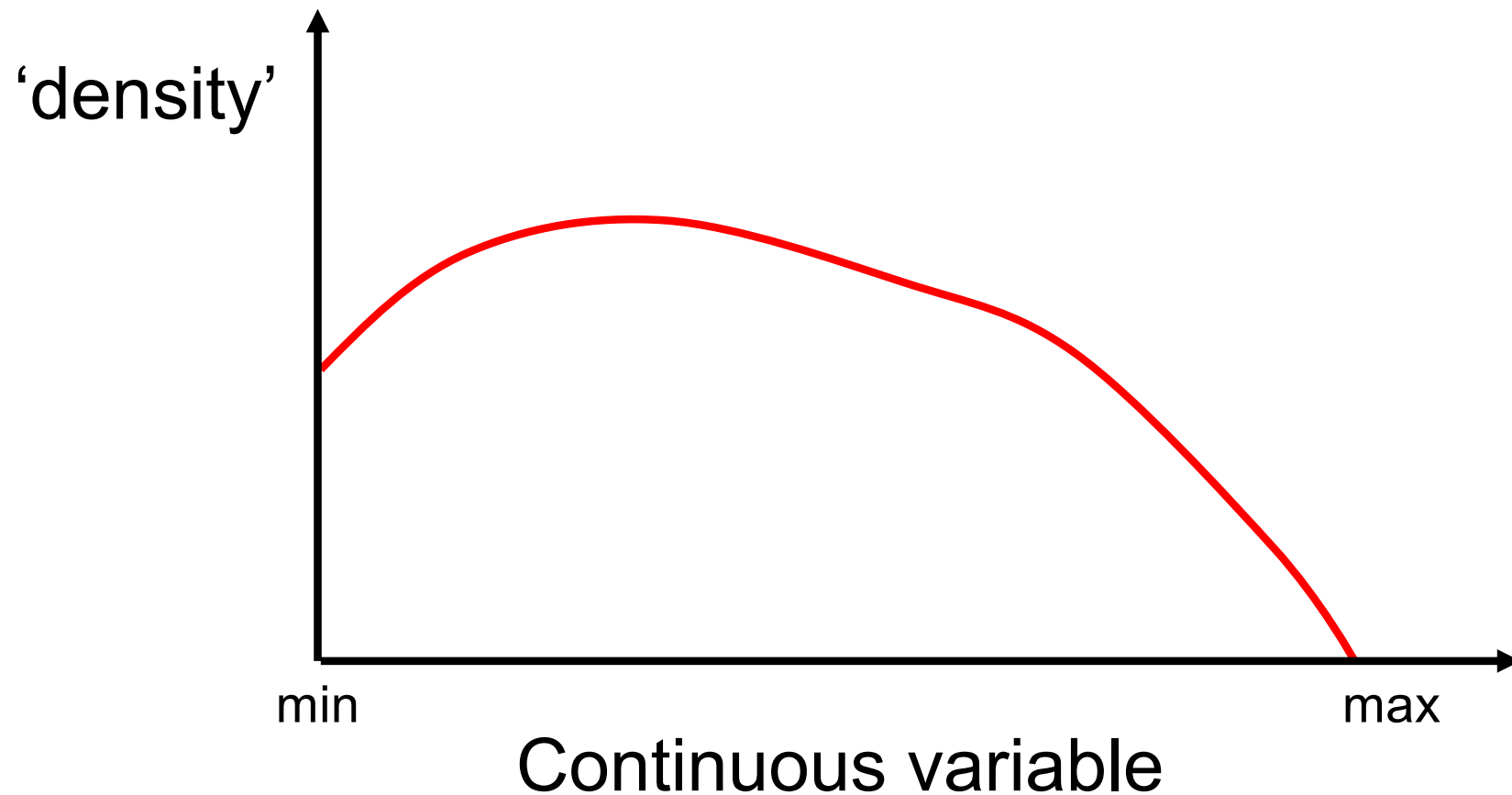
We can summarise a distribution

Location (or Average)

- What do we mean by 'average'?
 - Where is the 'middle' located?
- Types of average
 - Mean: $\text{sum of value} / \text{number of values}$
 - Median: divides population in half
 - Mode: most common value
- First way to summarize a distribution

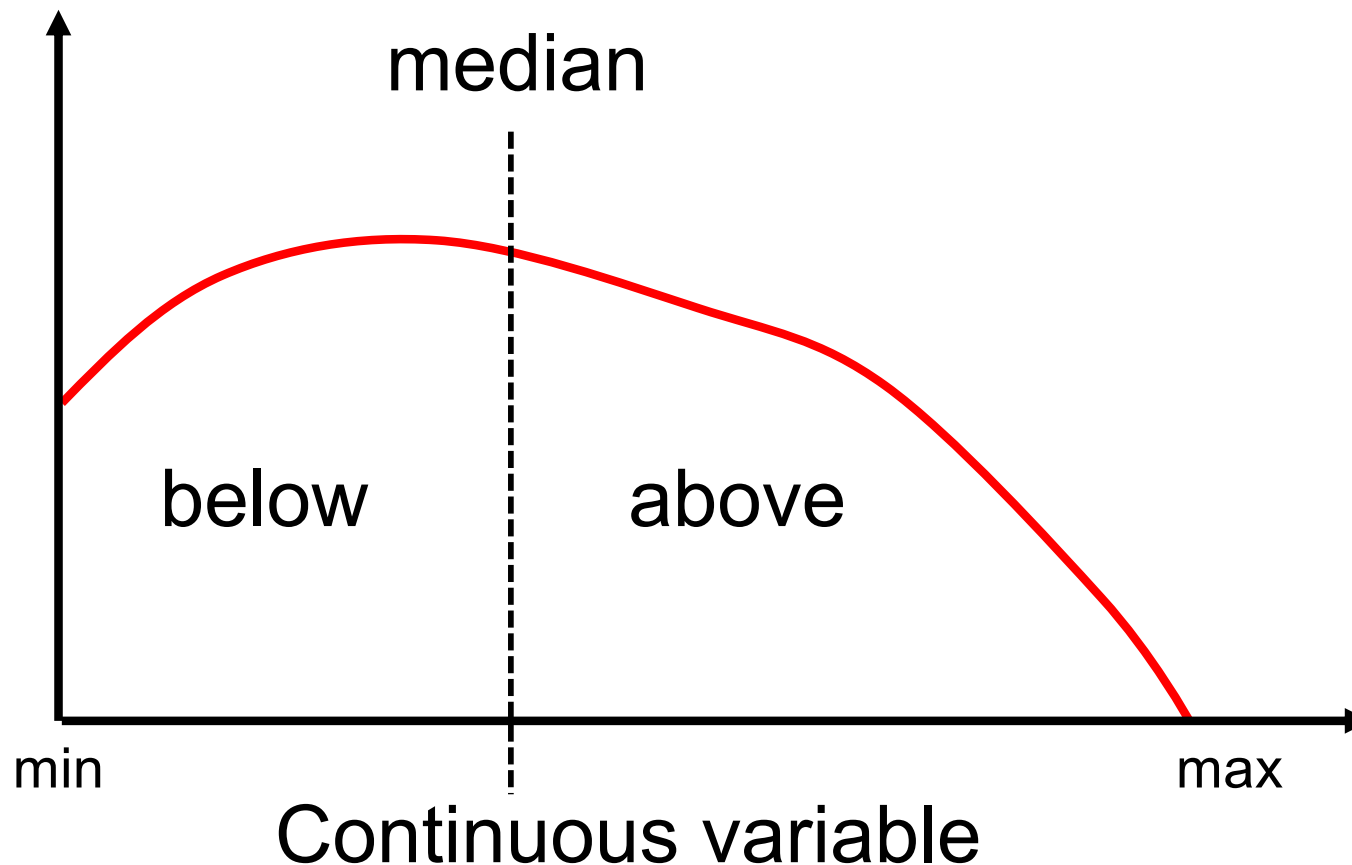
Median

- Median: divides population in half



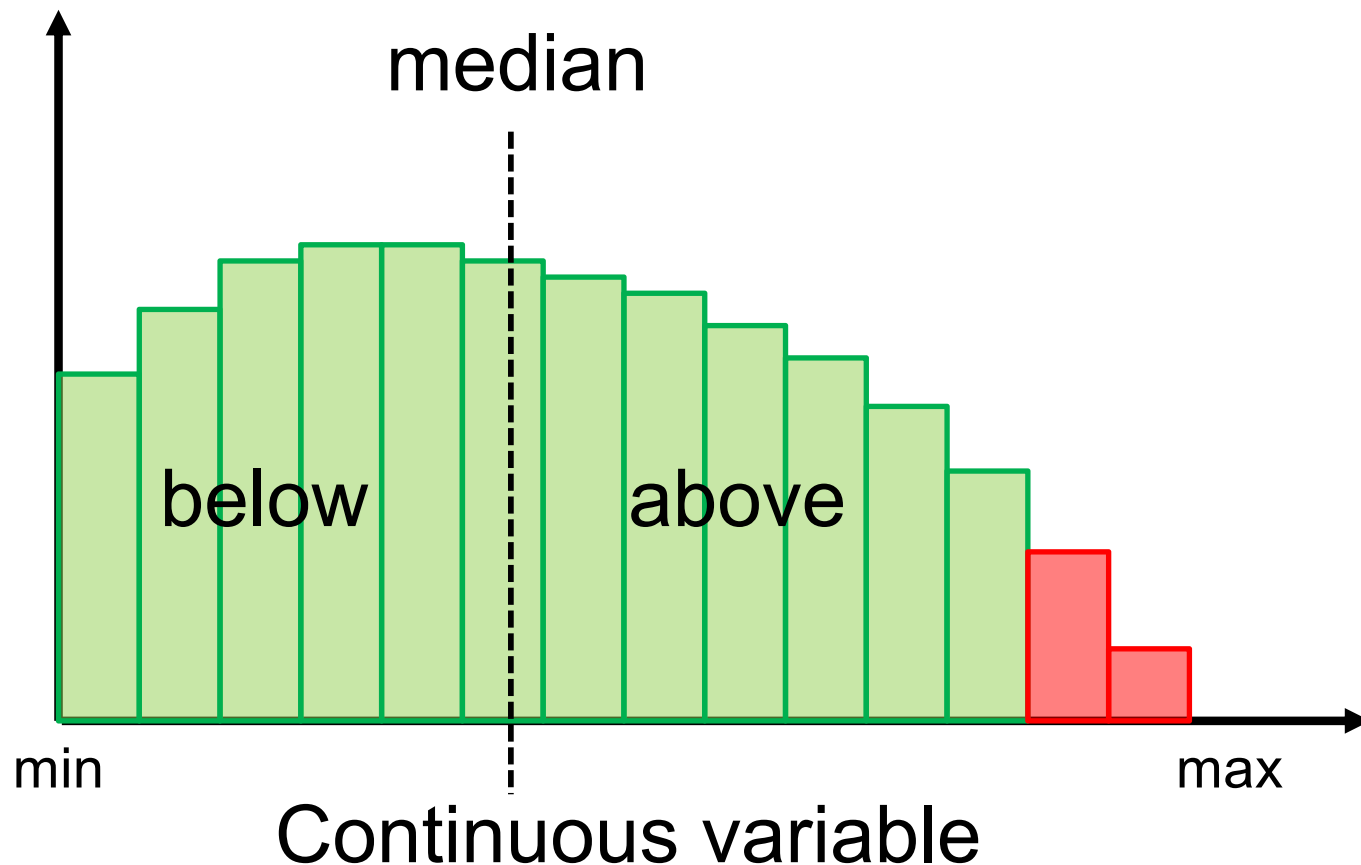
Median

- Median splits the area
 - Area (population) below = area above



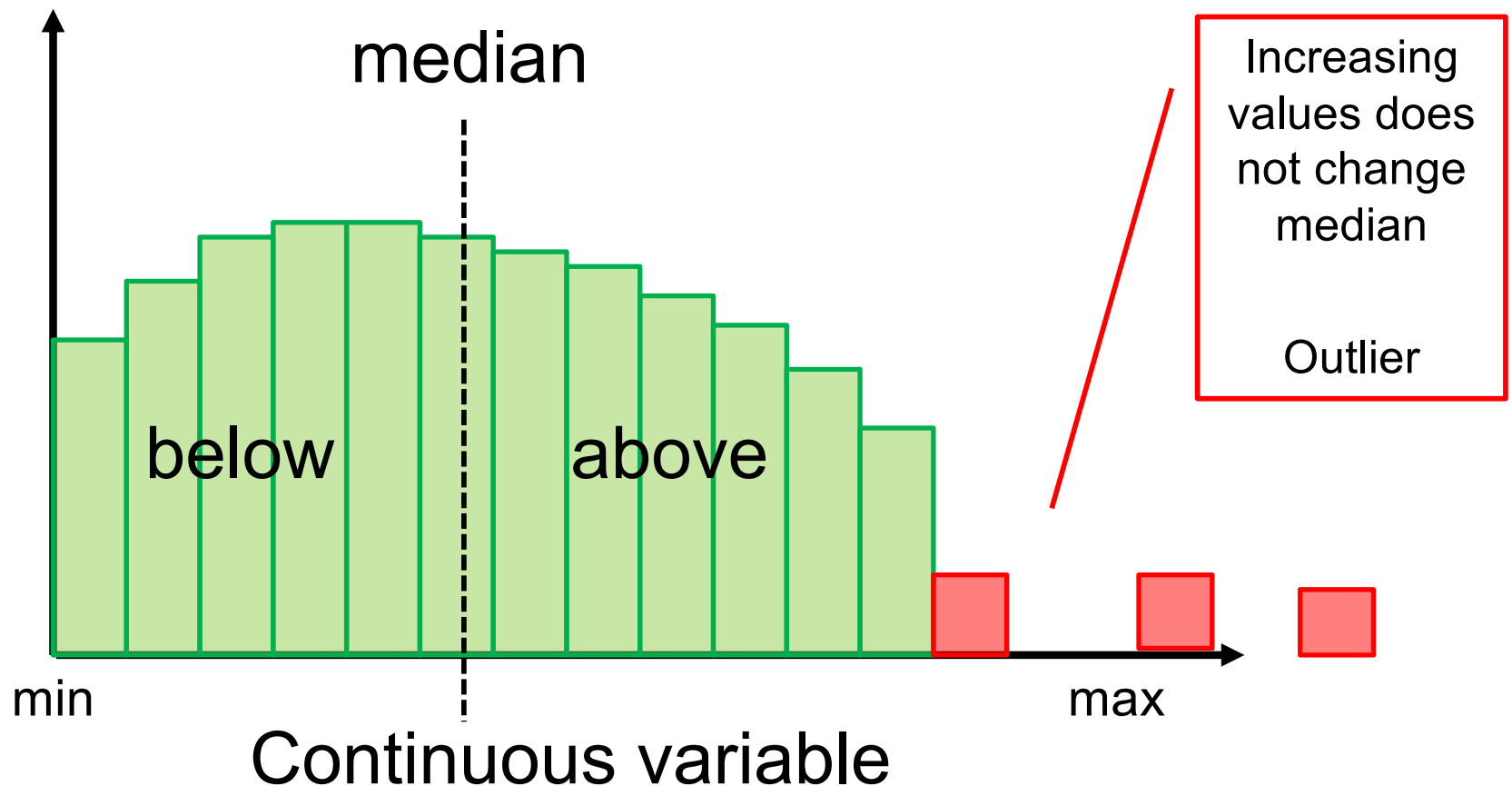
Median

- Median splits the area
 - Area (population) below = area above



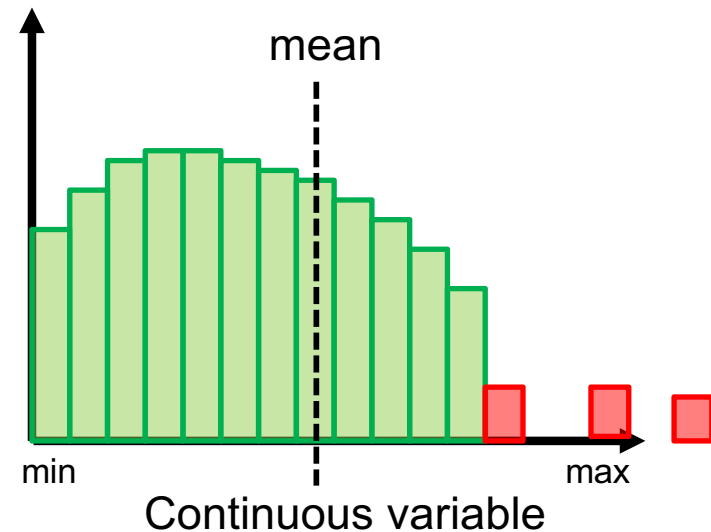
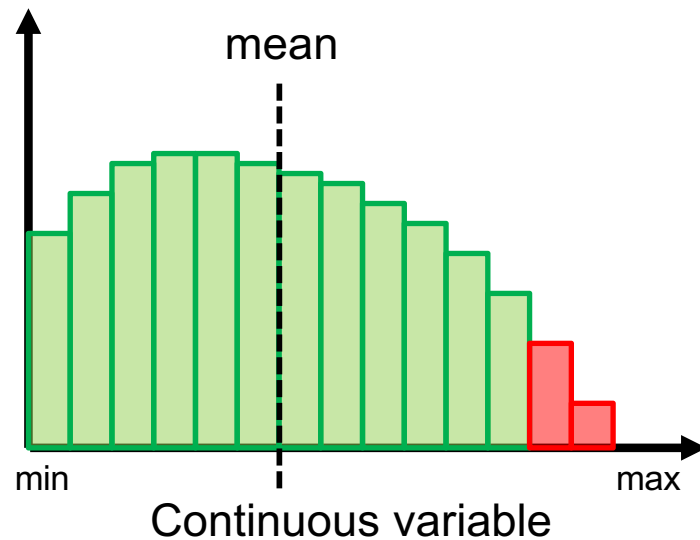
Median: No Change

- Median splits the area
 - Area (population) below = area above



Mean: Average Value

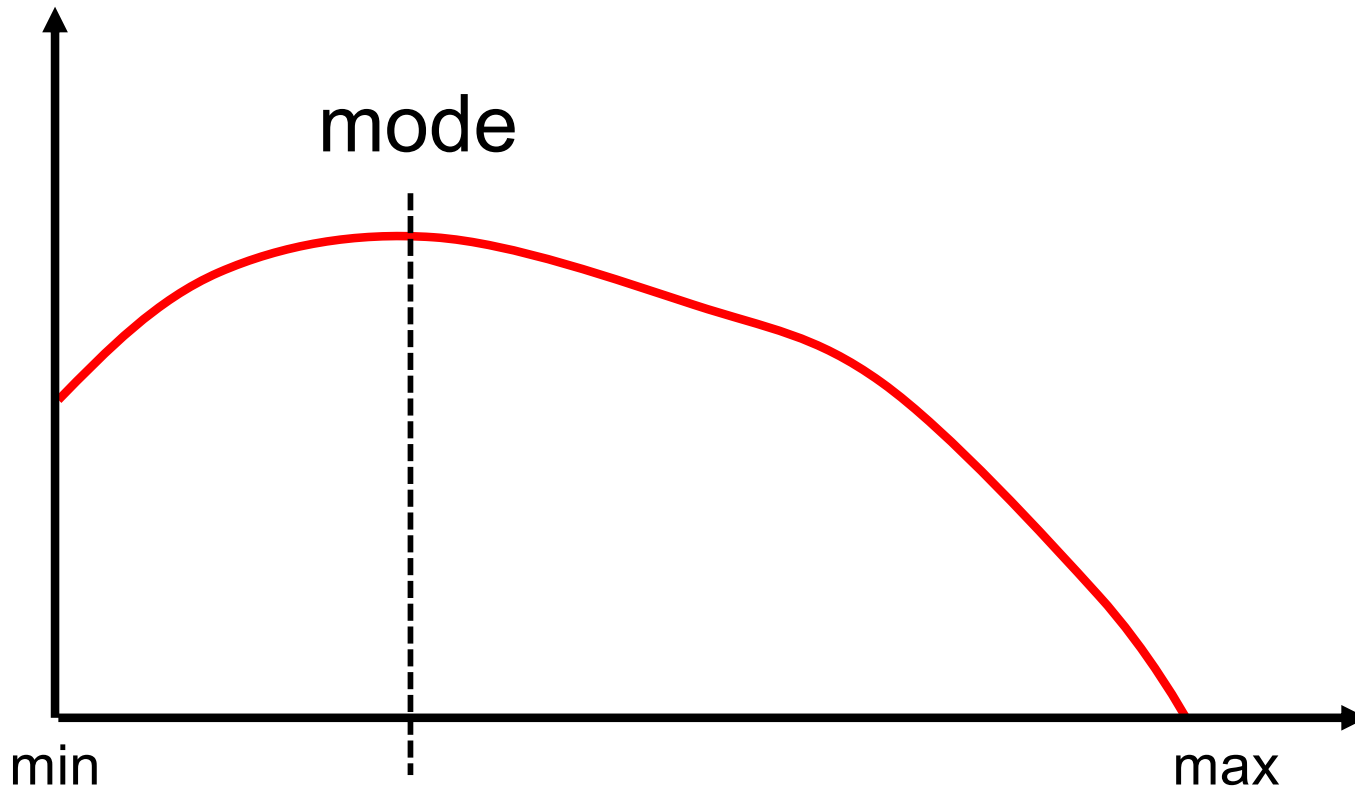
- Outliers do affect the mean
 - Sum values / number of values



----->
mean increase

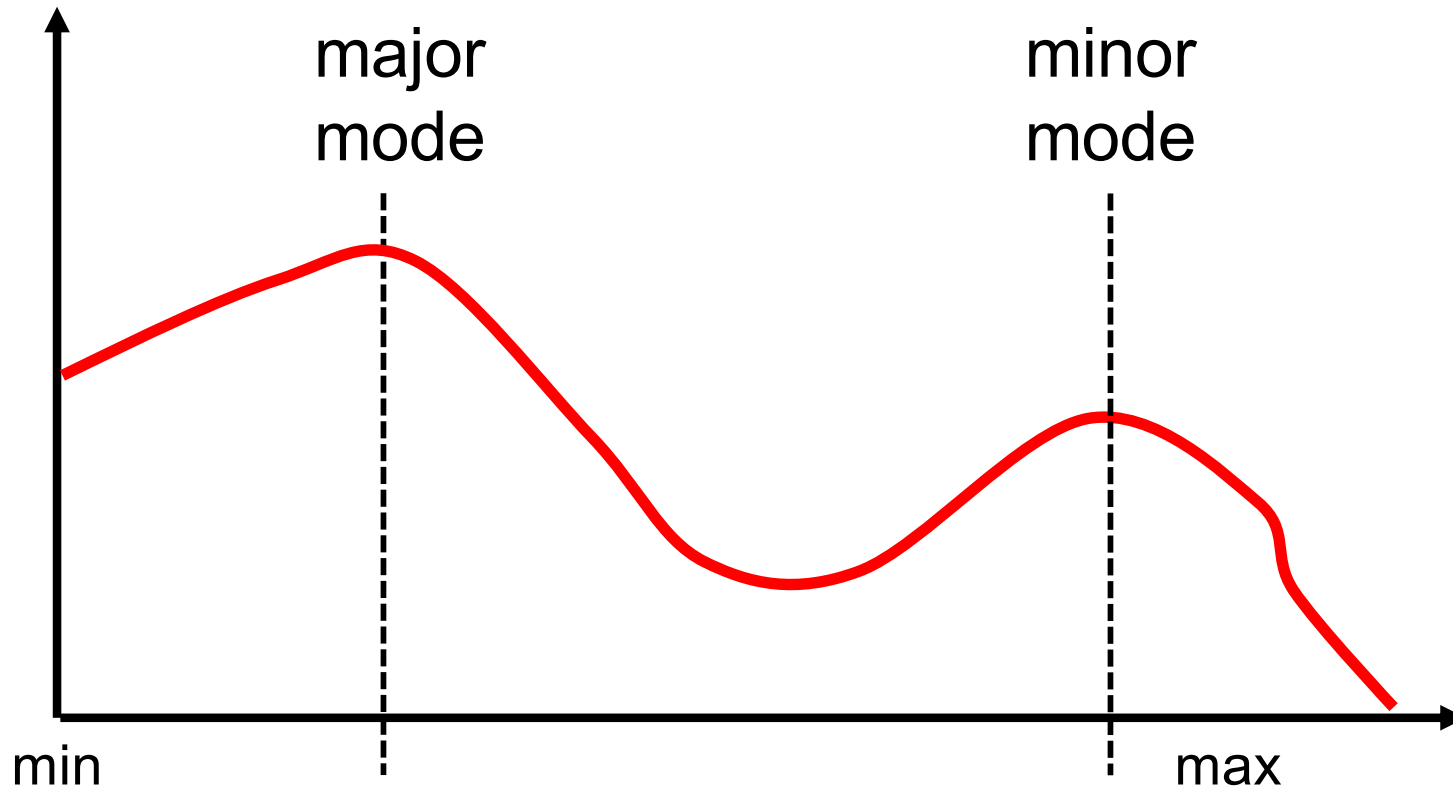
Mode

- Mode is 'most common value'
- Unimodal distribution



Multimodal Distribution

- Bimodal distribution
- Check for multimodality before reporting mean, median etc

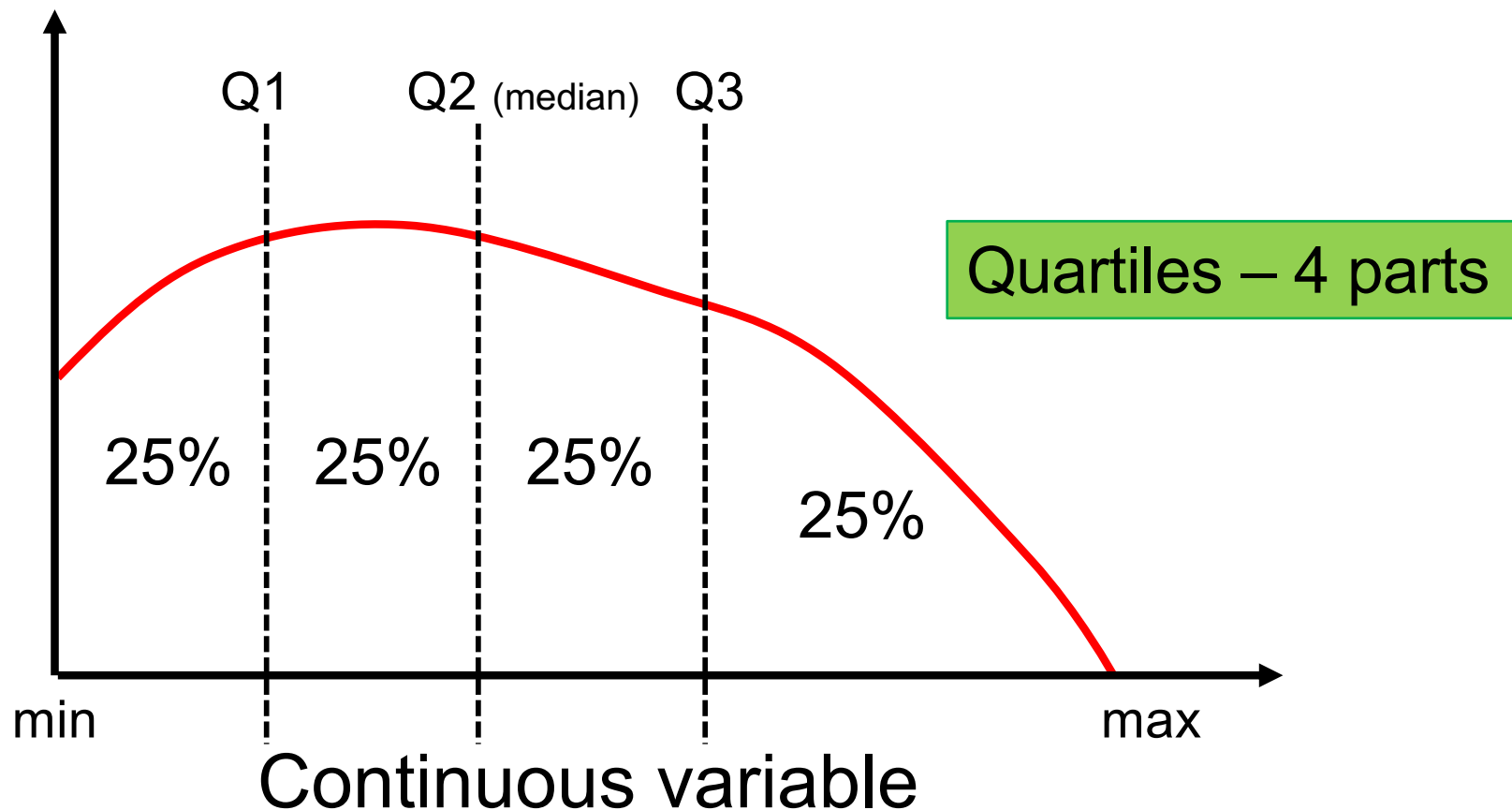


Quantiles

Generalise median

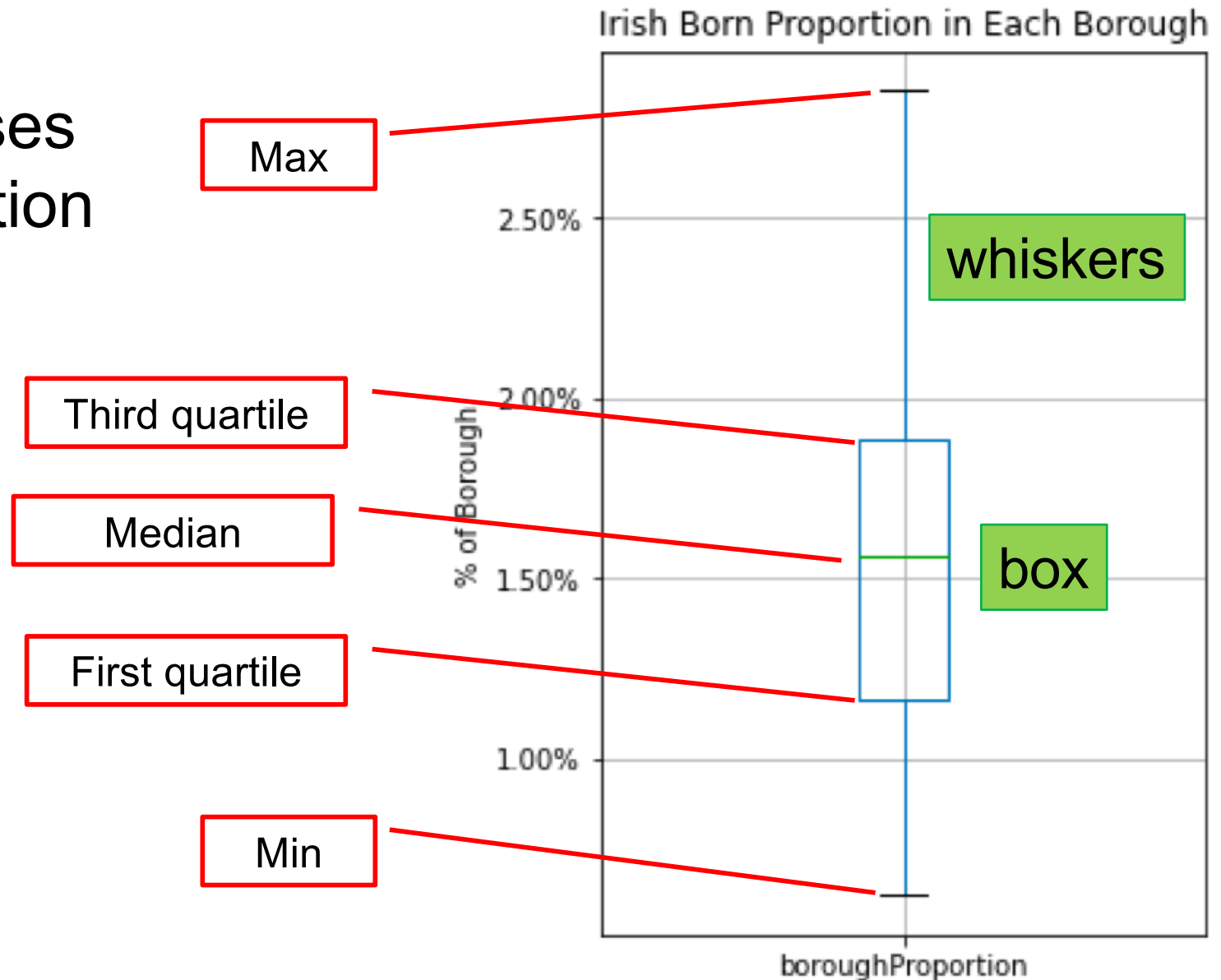
Quantiles of a Distribution

- Quantiles split the area into equal parts
 - Area (population) equal in each part
 - Quartile: 4 parts; deciles: 10 parts



Box Plot

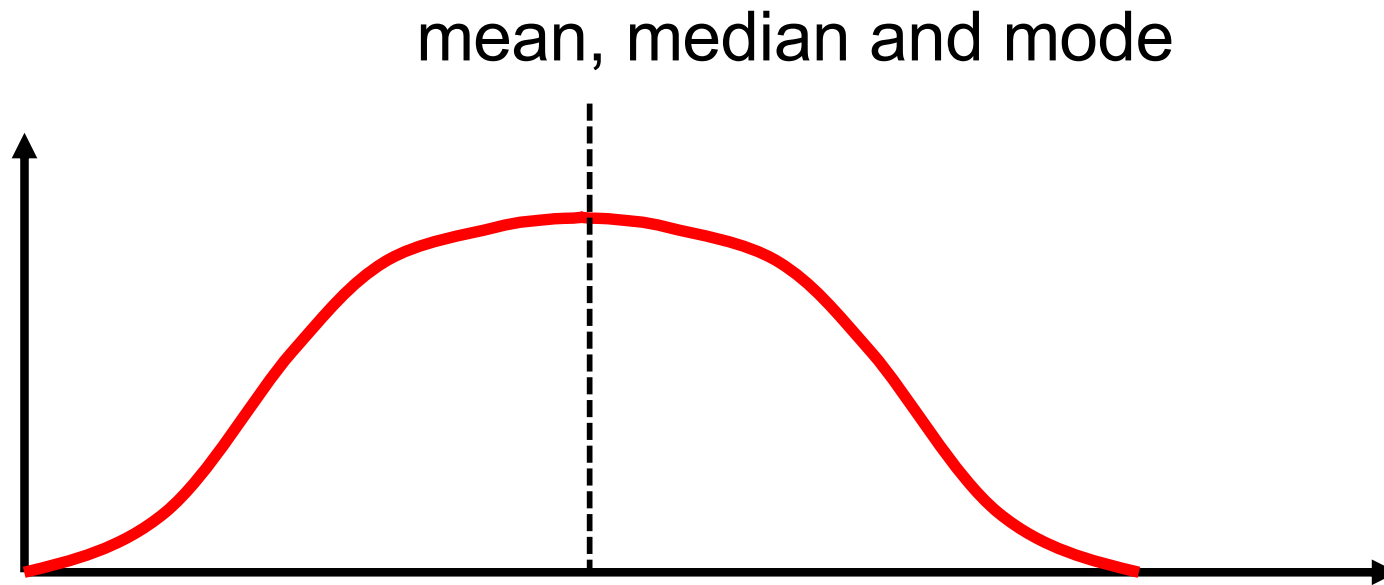
- Box plot summarises a distribution



Symmetry and Skew

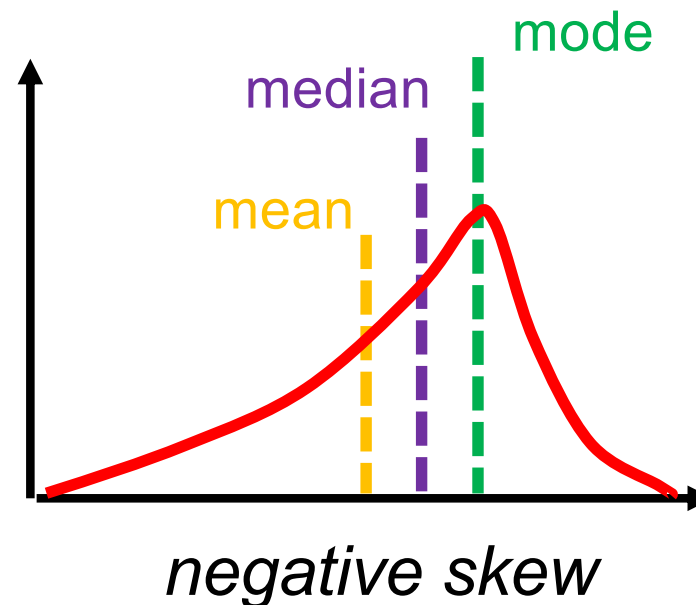
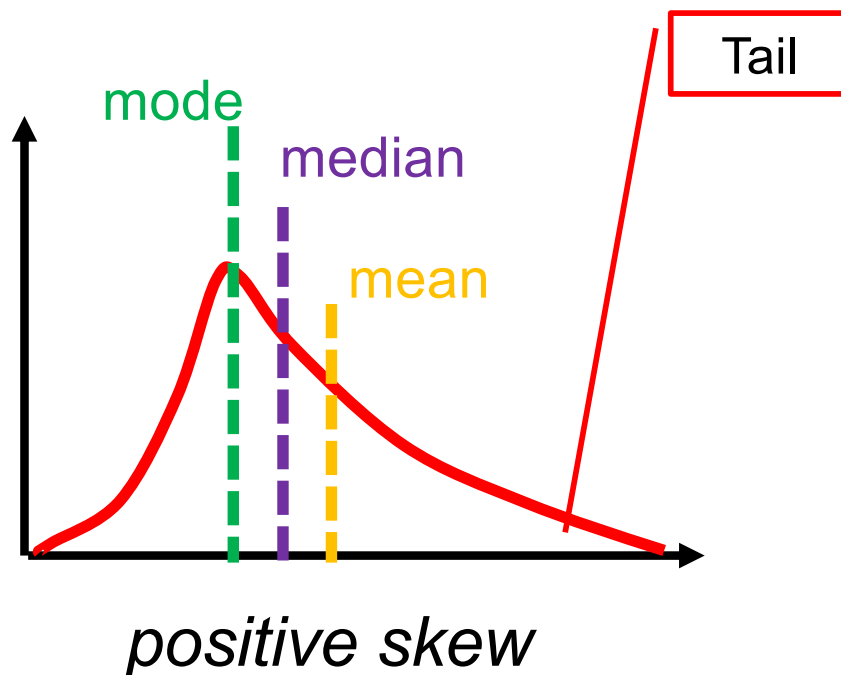
Symmetric Distribution

- A symmetric distribution is the same when reflected in the mean
 - The mean, mode median and mode are all the same



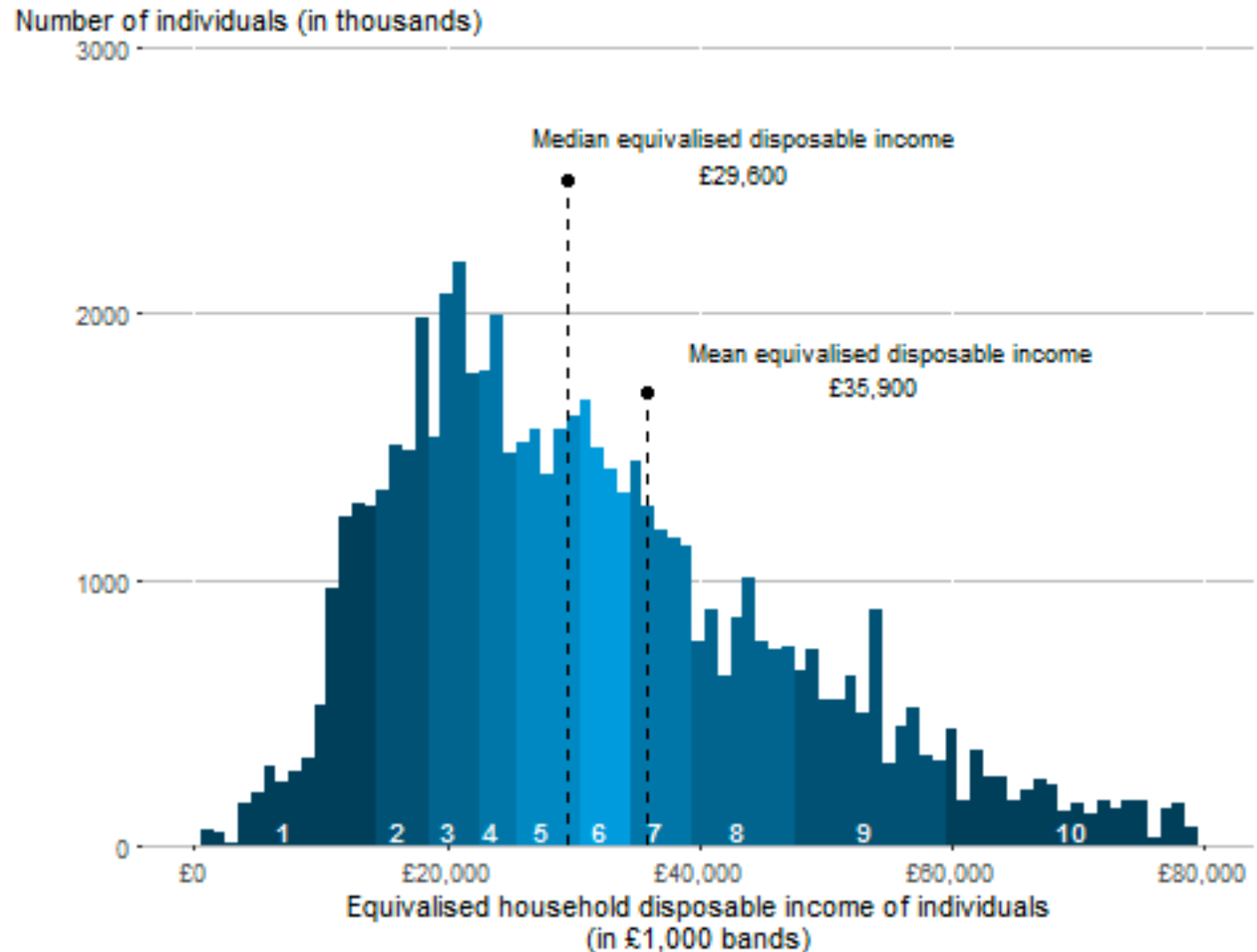
Skew: Measures Asymmetry

- 0 for a symmetric distribution
 - However, 0 skew does not imply symmetry
- Can be positive or negative
 - Positive skew: tail to right
 - Negative skew: tail to left



Example: (Household Disposable) Income


- Right skew:
mean
exceeds
median
- Lowest
income zero:
no highest
– Expect skew
- Mean not only
(or best)
summary
when skew



Skew and 'Location'

- Skew is also a summary (like location)
 - Different metrics: best to look at distribution
 - Can be defined on difference between mean and median

| Skew | Tail | Mean v Median | Median v Mode |
|----------|-------------------------|---------------|---------------|
| Positive | On the right (+ve) side | Mean bigger | Mode smaller |
| Zero | No tail | Same | Same |
| Negative | On the left (-ve) side | Mean smaller | Mode bigger |



Rules of thumb for unimodal distributions: not always true

Summary

- Continuous probability distribution
 - Show probability of any range of values
 - Cannot give a probability to a single value
- Histogram
 - Similar to a bar chart (but no gaps; order)
 - Divides continuous variable into intervals
- Density
 - Imagine histogram with very narrow intervals
- Summary 'statistic' for 'location'
 - Where is distribution centred ('located')?
 - Mean, median or mode
- Many real distributions asymmetric
 - Mean and median different