

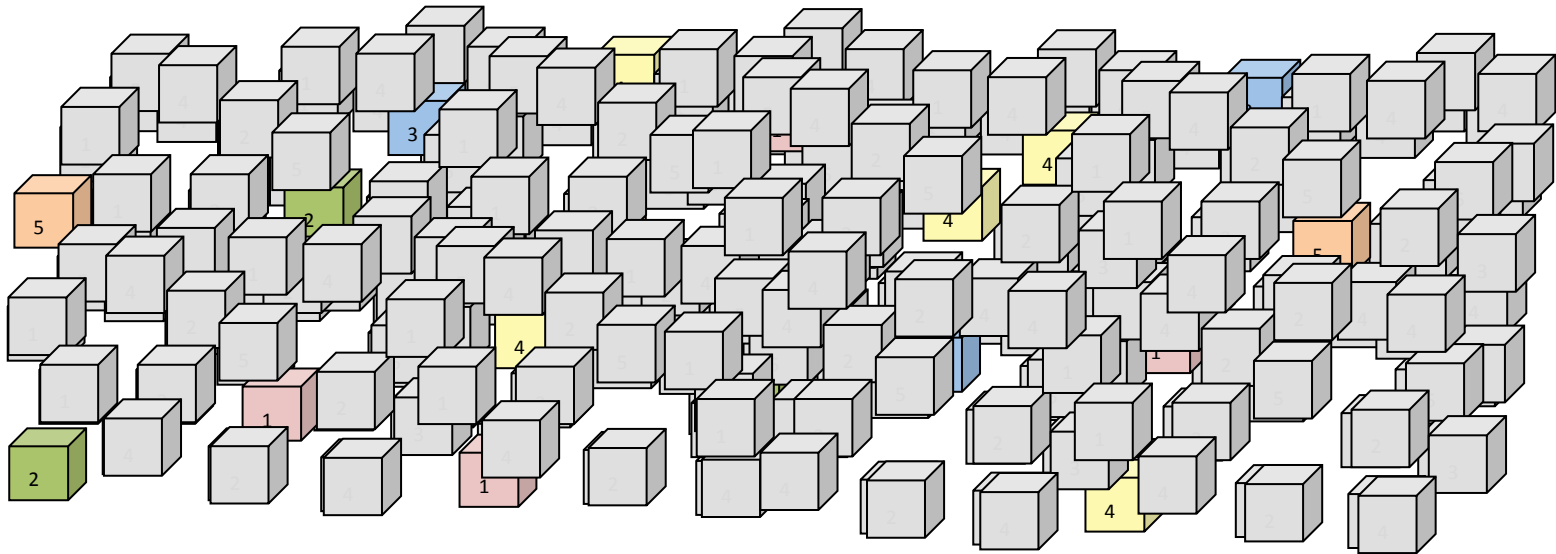
# Stats 101 for GIS

Spring 2009

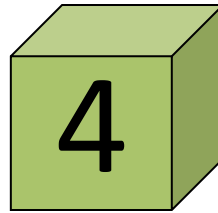
A **POPULATION** is a collection of entities



A **sample** is an observed subset of a population



# Multiple properties can be measured



For example, here we might have color and weight.

# Or several census variables...

STFID	POP2000	WHITE	MALES	MED_AGE	MED_AGE_M	MED_AGE_F
60014069001	691	495	327	43.3	41.7	43.7
60014069002	1095	637	500	45.1	44	46
60014069003	1909	399	881	32.1	30.4	34.4
60014076001	1144	283	497	35.6	33.1	37.7

**Datum** means one observation about one entity...

**Data** refers to more than one observation or measurement...

"How is the data distributed?"

What values show up how often?

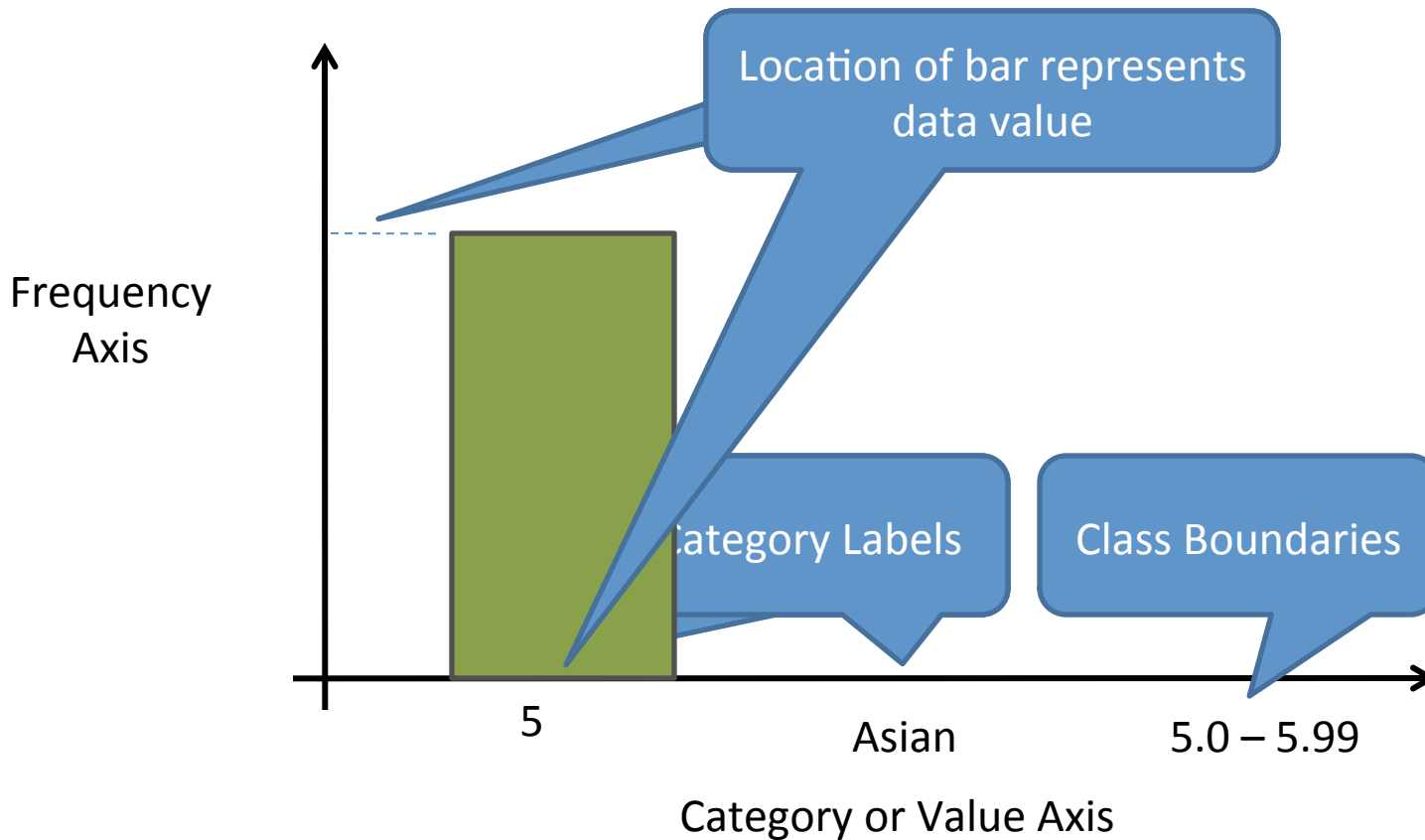
# How to characterize a distribution I

- Central Tendency – mean, median, mode
- Dispersion – how spread out?
- Skewness – symmetric? Leans right or left?



# Visualizing a Distribution

- Frequency Histogram

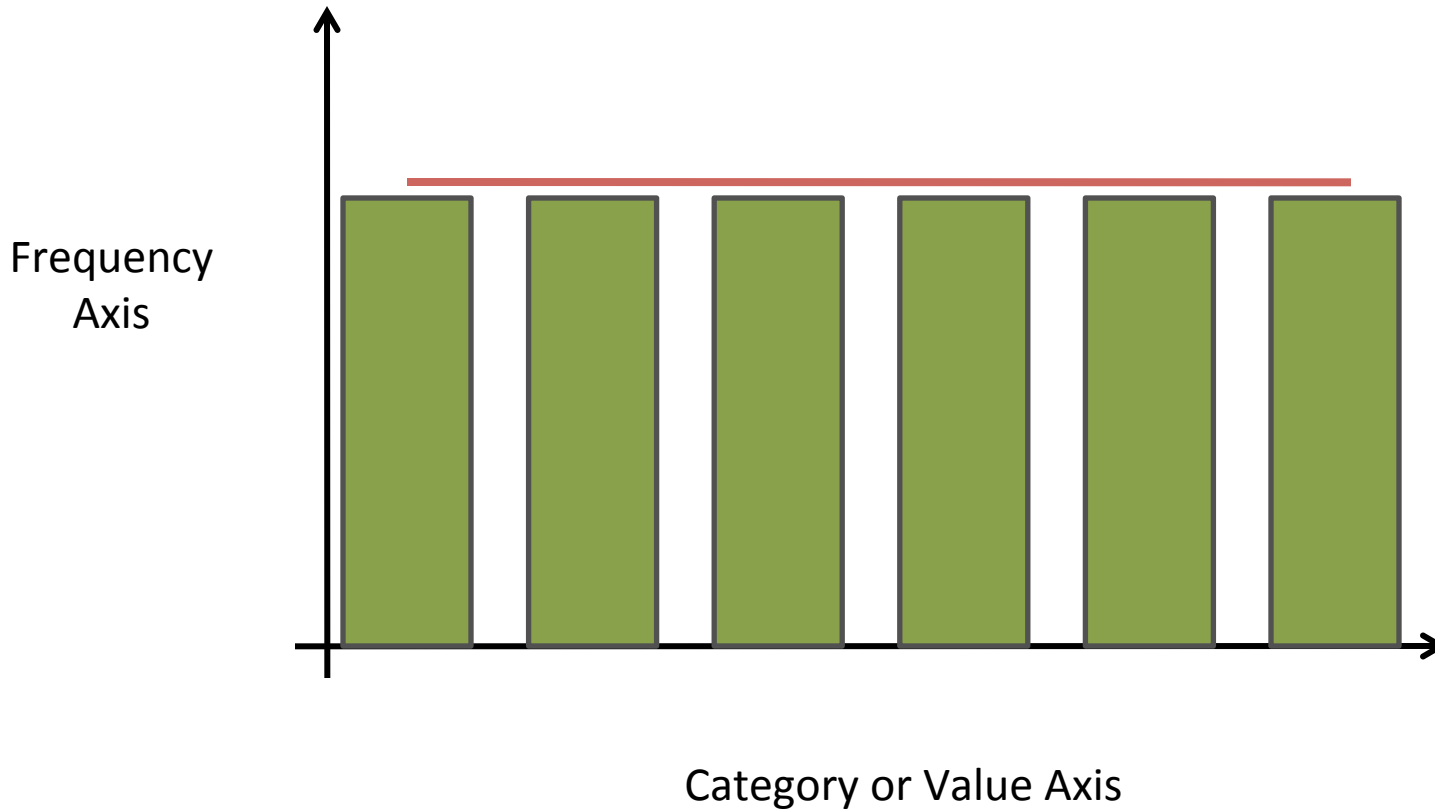


# How to characterize a distribution II

- General shape
- Known distributions

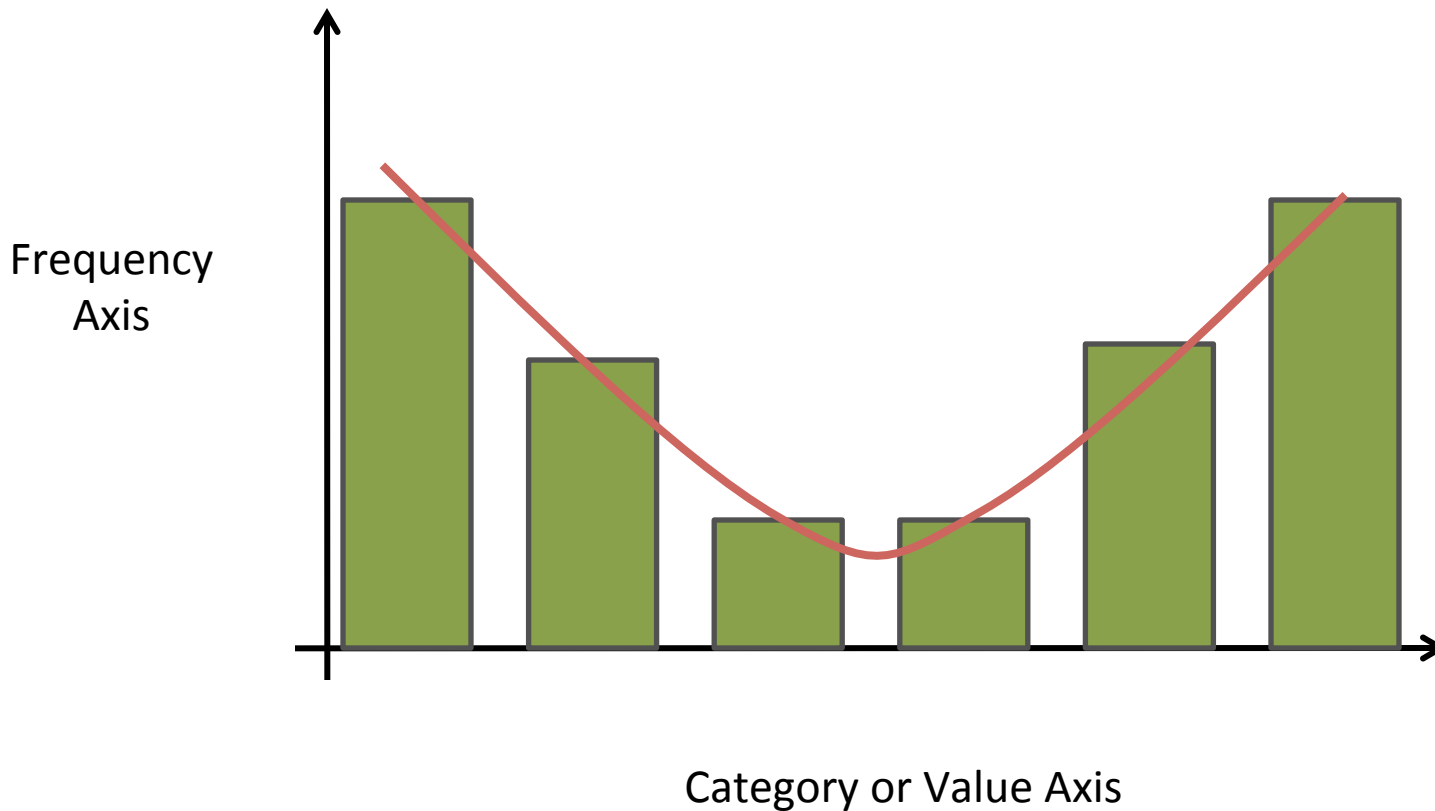
# Visualizing a Distribution

- **Uniform** Distribution – every value has same frequency



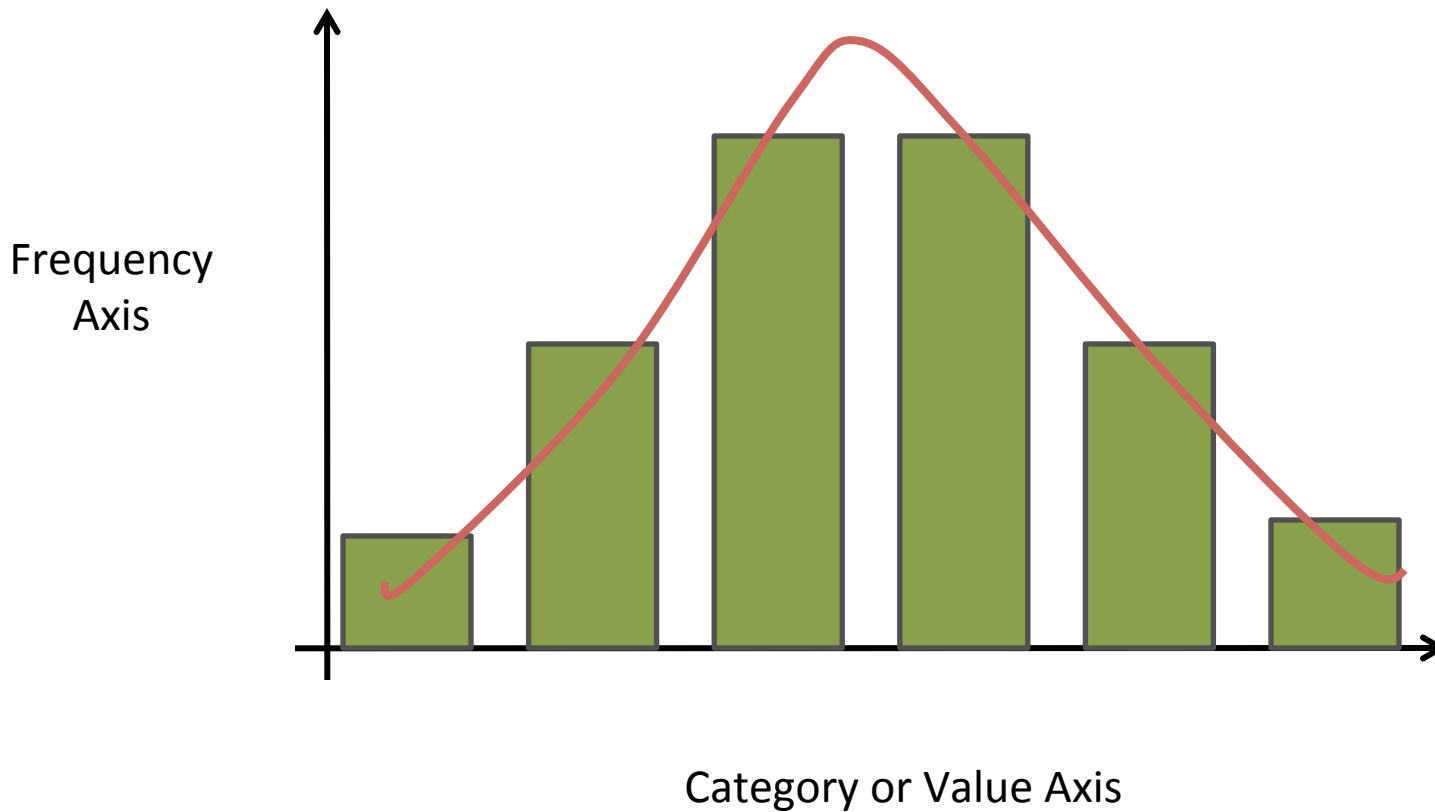
# Visualizing a Distribution

- **U-Shaped** Distribution – high frequency at extremes, low in the middle



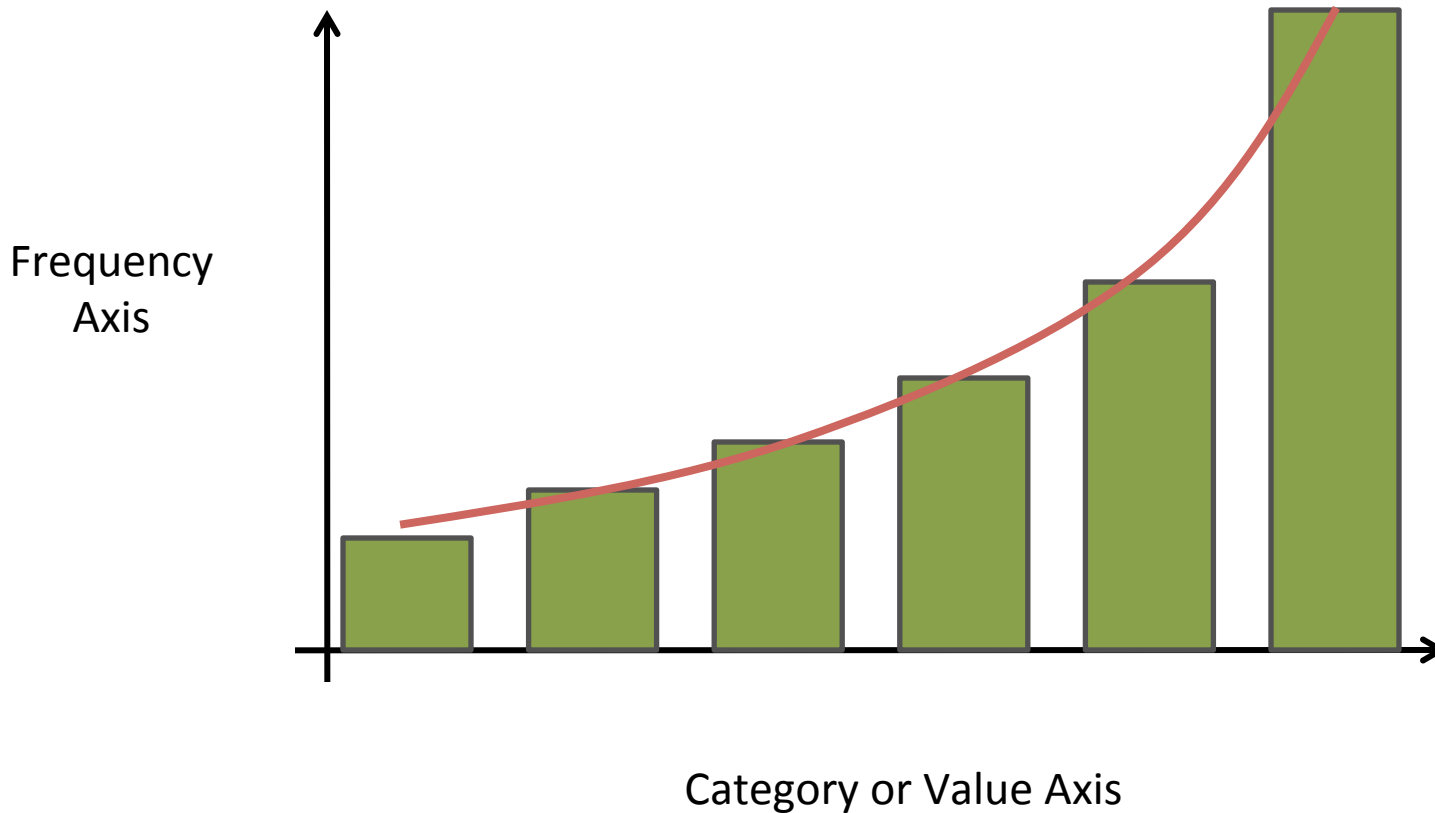
# Visualizing a Distribution

- **Bell/Normal** Distribution



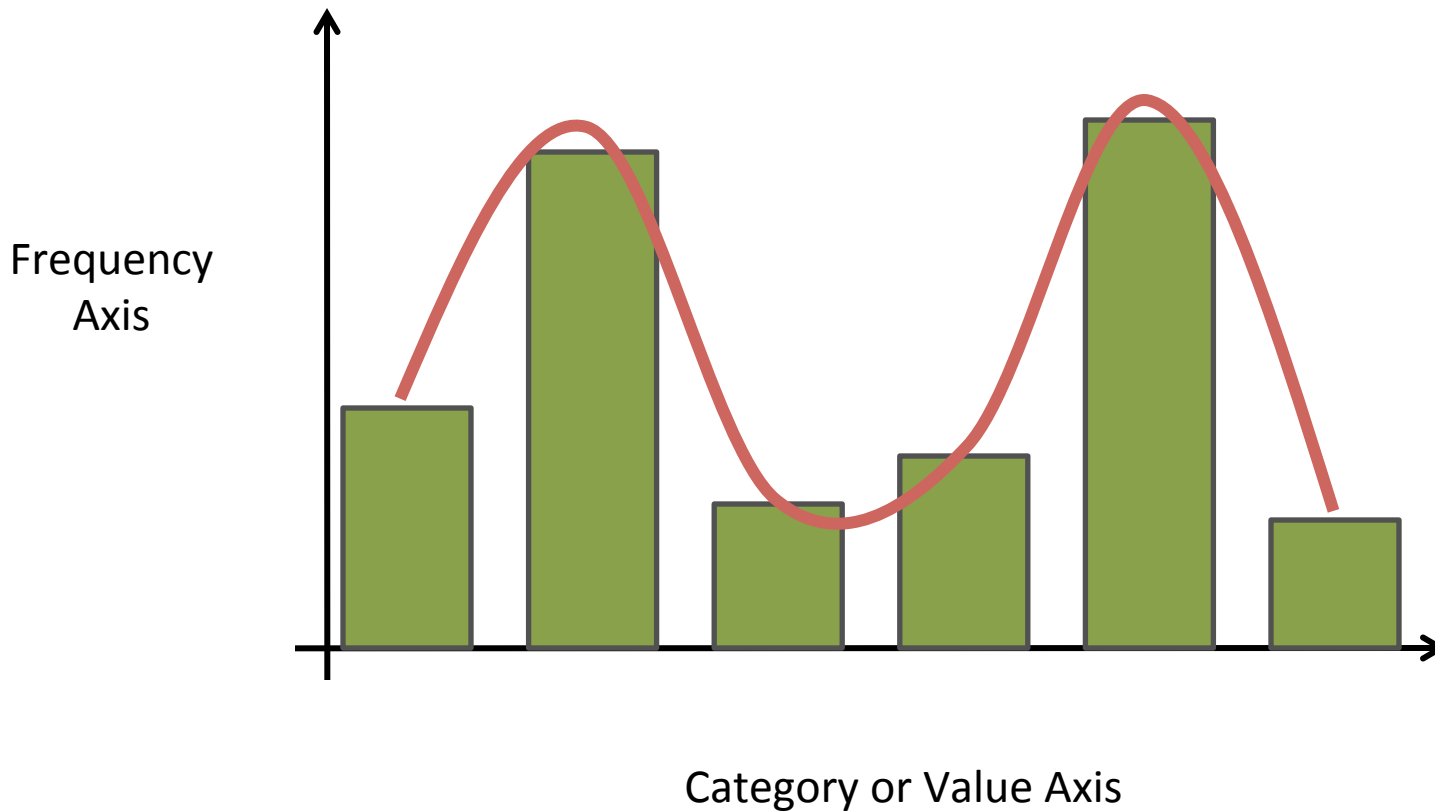
# Visualizing a Distribution

- **Skewed** Distribution – low at one end, high at the other

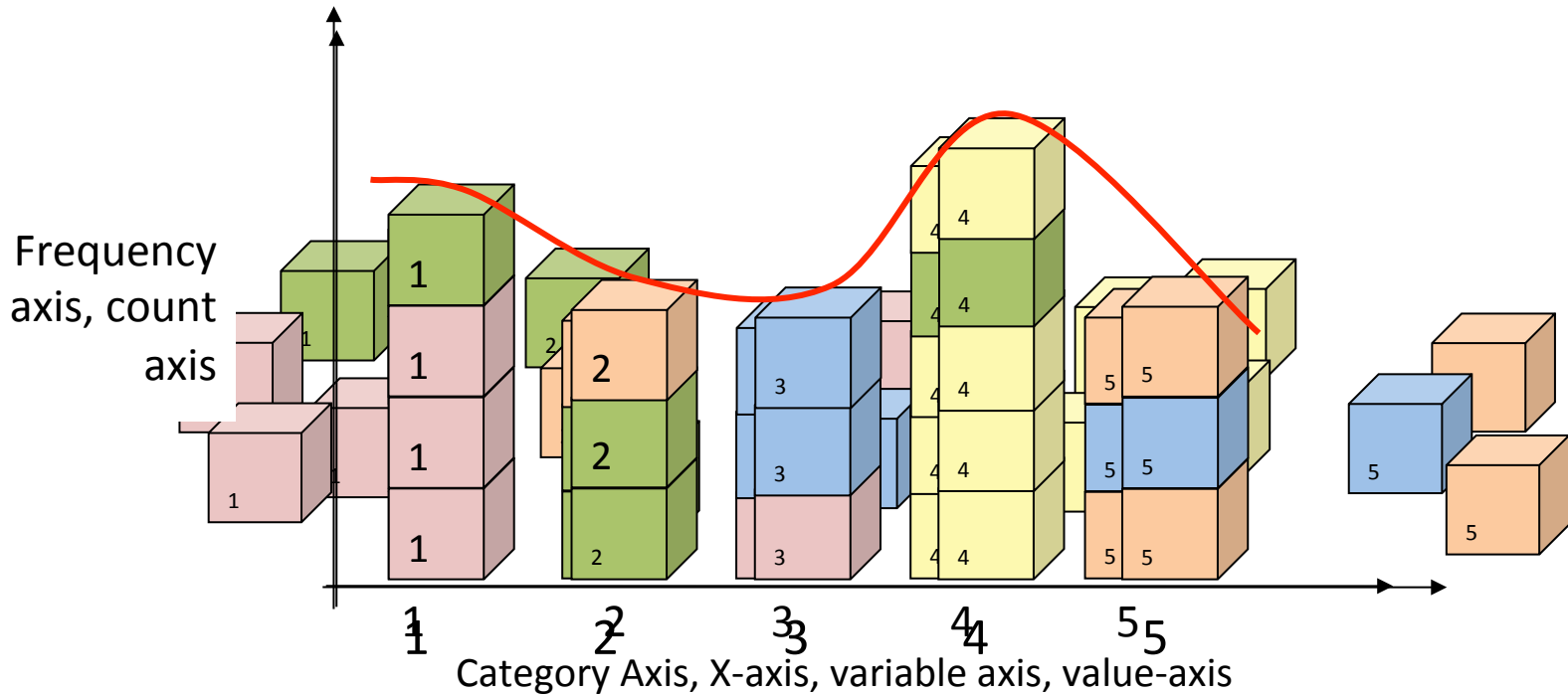


# Visualizing a Distribution

- **Bimodal** Distribution – two peaks

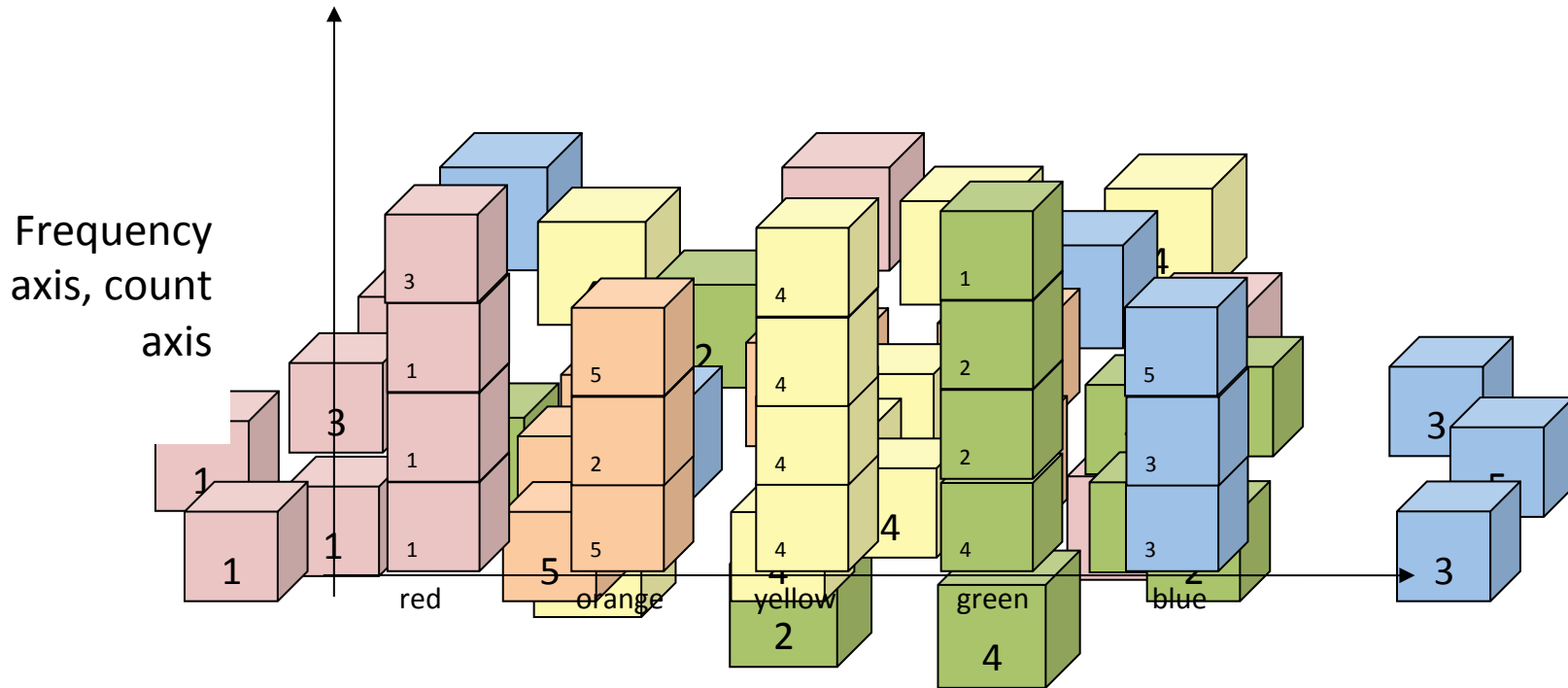


# Back to Our Data



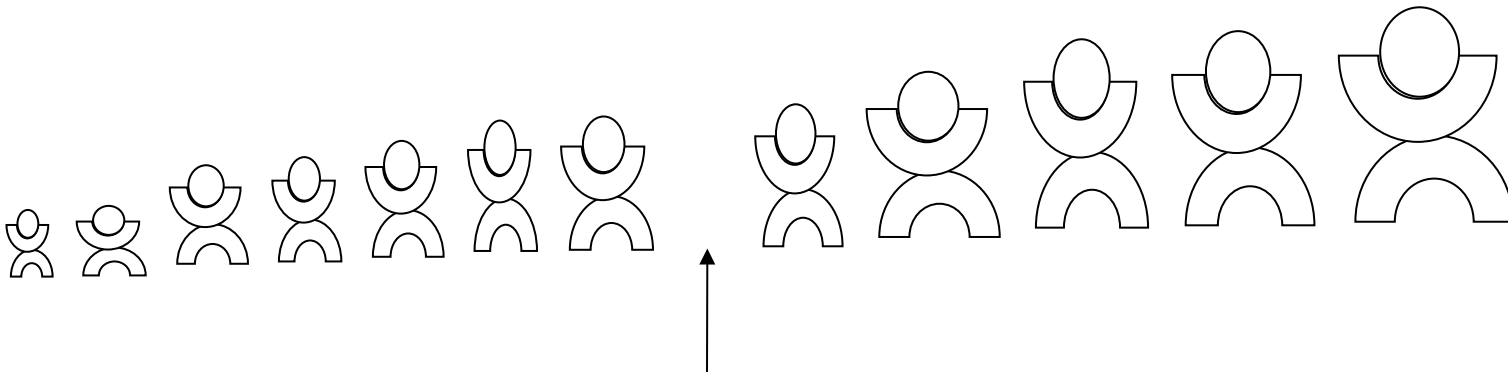


# Same Entities, Different Variable



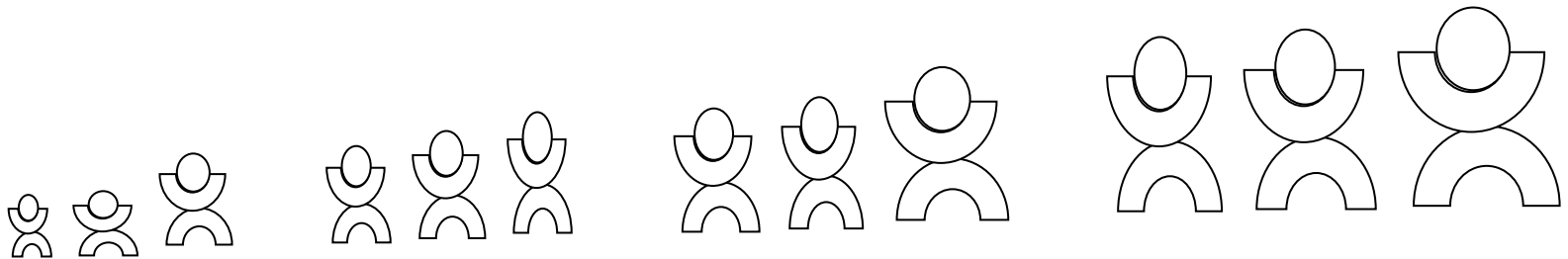
# Quantiles, etc.

- Median is the data value that separates sample into top 50% and bottom 50%



# Sometimes we want to divide more finely...

- “Quartiles” is when we look at lowest 25%, next 25%, etc.



# Important

- For quantiles we
- Sort cases by a variable
- Count cases and separate into groups
- Report **data value** at each cut