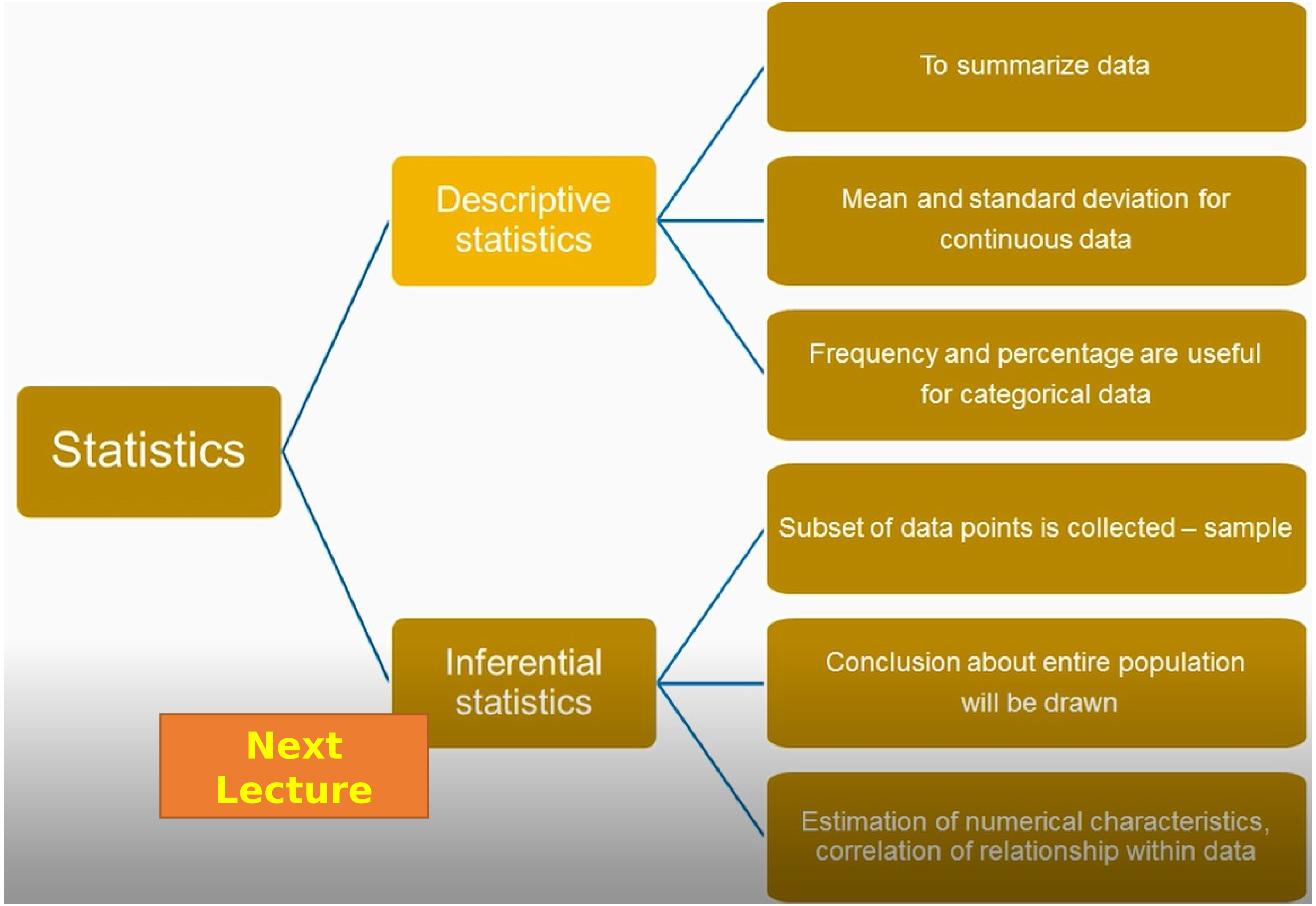
CSE 445
Lecture 4

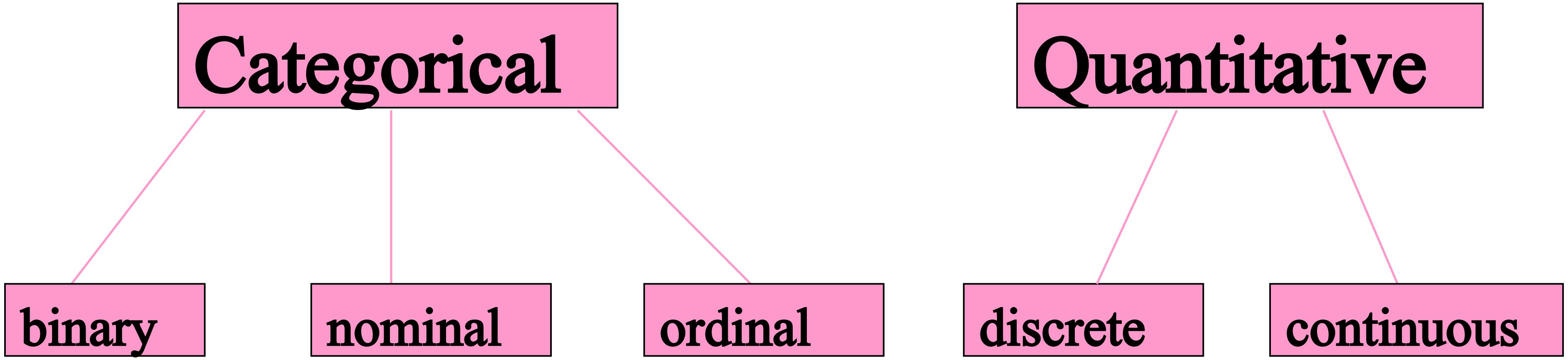
Statistics & Probability for Machine Learning

Statistics


Statistics
Descriptive statistics
Inferential statistics
To summarize data
Mean and standard deviation for continuous data
Frequency and percentage are useful for categorical data
Subset of data points is collected – sample
Conclusion about entire population will be drawn
Estimation of numerical characteristics, correlation of relationship within data
Next Lecture

Look for Pattern using Statistics

Types of Variables: Overview



2 categories +

more categories +

order matters +

numerical +

uninterrupted

Categorical Variables

- Also known as “qualitative.”

- Categories.
    - treatment groups
    - exposure groups
    - disease status

Categorical Variables

- <u>Nominal variables</u> – Named categories Order doesn't matter!
    - The blood type of a patient (O, A, B, AB)
    - Marital status
    - Occupation

Categorical Variables

- Ordinal variable – Ordered categories. Order matters!
    - Staging in breast cancer as I, II, III, or IV
    - Birth order—1st, 2nd, 3rd, etc.
    - Letter grades (A, B, C, D, F)
    - Ratings on a scale from 1-5
    - Ratings on: always; usually; many times; once in a while; almost never; never
    - Age in categories (10-20, 20-30, etc.)
    - Shock index categories (Kline et al.)

Quantitative Variables

- Numerical variables; may be arithmetically manipulated.
  - Counts
  - Time
  - Age
  - Height

Quantitative Variables

- Discrete Numbers – a limited set of distinct values, such as whole numbers.
    - Number of new AIDS cases in CA in a year (counts)
    - Years of school completed
    - The number of children in the family (cannot have a half a child!)
    - The number of deaths in a defined time period (cannot have a partial death!)
    - Roll of a die

Quantitative Variables

- Continuous Variables - Can take on any number within a defined range.
    - Time-to-event (survival time)
    - Age
    - Blood pressure
    - Serum insulin
    - Speed of a car
    - Income
    - Shock index (Kline et al.)

Looking at Data

- ü How are the data distributed?
  - Where is the center?
  - What is the range?
  - What's the shape of the distribution (e.g., Gaussian, binomial, exponential, skewed)?

- ü Are there "outliers"?

- ü Are there data points that don't make sense?

Central Tendency: Mean, Median, and Mode

- Mean:
  - Simple arithmetic average
  - Sensitive to outliers in data
- Median:
  - Midpoint of data
- Mode:
  - Most repetitive data point in data

Measure of Variation and Range

- Measures of variation:
  - Dispersion in the variation in data
  - Measures inconsistencies in values of variable
  - Dispersion provides and idea about the spread of the data rather than central values
- Range
  - Difference between maximum and minimum of value
- Variance:
  - Mean of squared deviations from mean

Central Tendency

- Mean – the average; the balancing point

calculation: the sum of values divided by the sample size

In math shorthand:

$$\overline{X} = \frac{\sum_{i=1}^{n} x}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

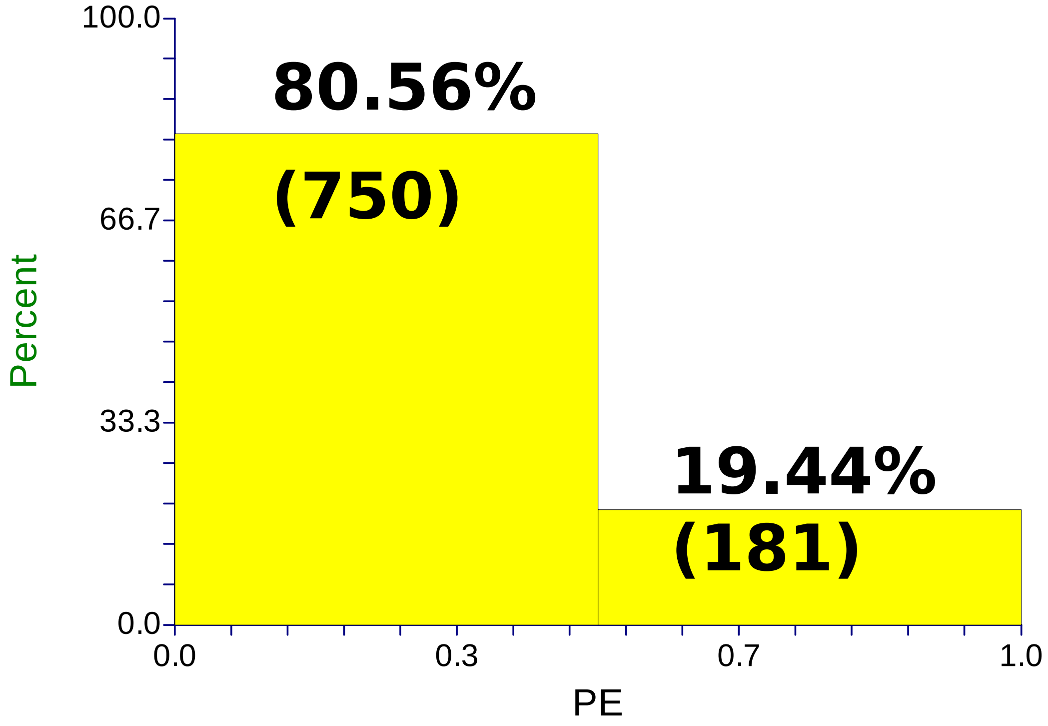
Mean: example

Some data:
Age of participants: 17 19 21 22 23 23 23 38

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{17+19+21+22+23+23+23+38}{8} = 23.25$$

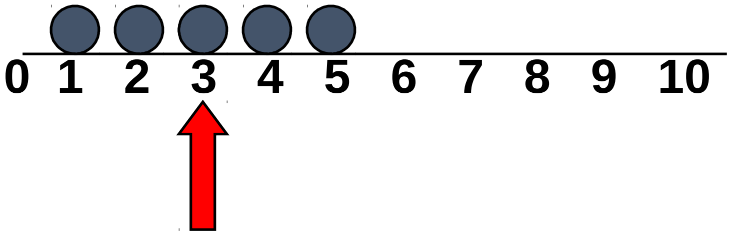
Mean of Pulmonary Embolism? (Binary variable?)

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{181*1 + 750*0}{931} = \frac{181}{931} = .1944$$



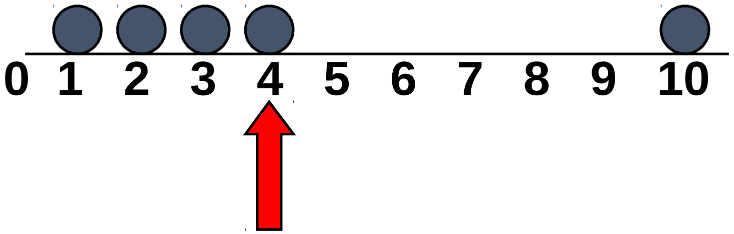
Mean

- The mean is affected by extreme values (outliers)



Mean = 3

$$\frac{1+2+3+4+5}{5}=\frac{15}{5}=3$$



Mean = 4

$$\frac{1+2+3+4+10}{5}=\frac{20}{5}=4$$

Central Tendency

- Median – the exact middle value

Calculation:

- If there are an odd number of observations, find the middle value
- If there are an even number of observations, find the middle two values and average them.

Median: example

Some data:
Age of participants: 17 19 21 22 23 23 23 38

Median = (22+23)/2 = 22.5

Central Tendency

- Mode – the value that occurs most frequently

Mode: example

Some data:

Age of participants: 17 19 21 22 23 23 23 38

Mode = 23 (occurs 3 times)

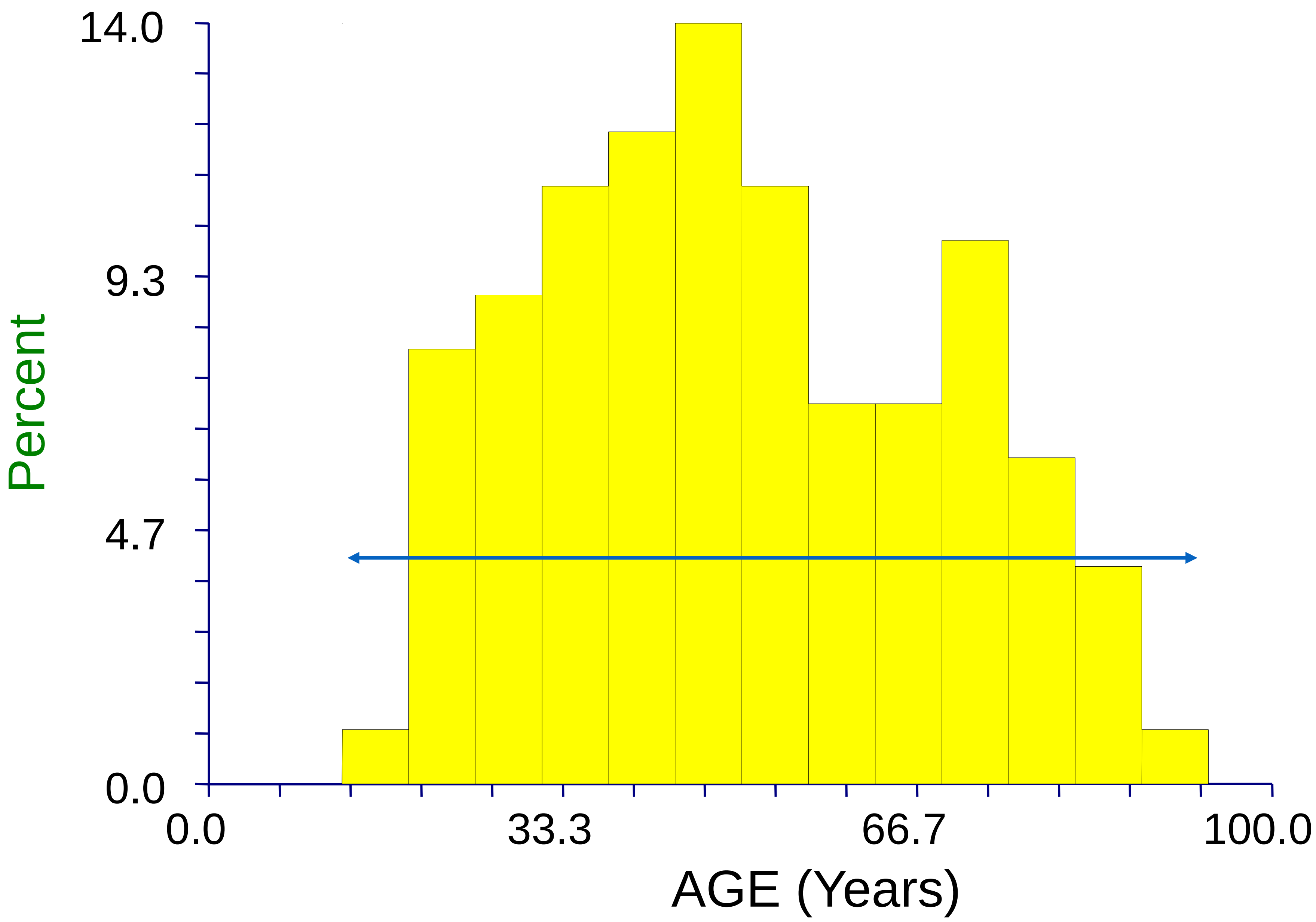
Measures of Variation/Dispersion

- Range
- Percentiles/quartiles
- Interquartile range
- Standard deviation/Variance

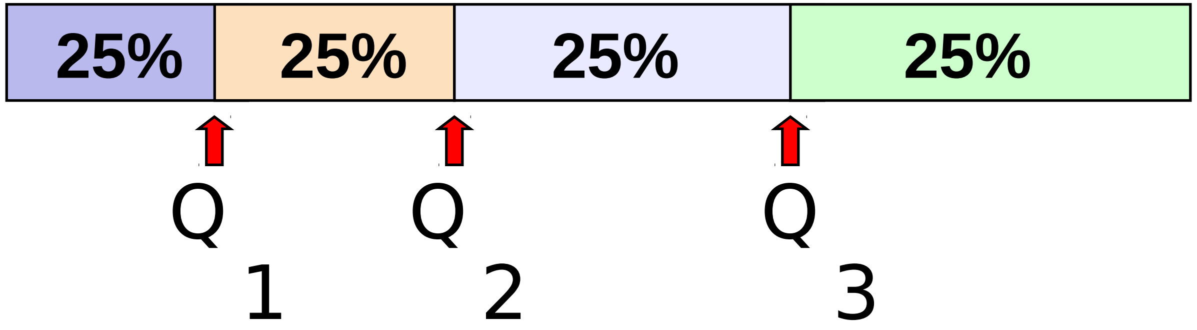
Range

- Difference between the largest and the smallest observations.

Range of age: 94 years-15 years = 79 years



Quartiles

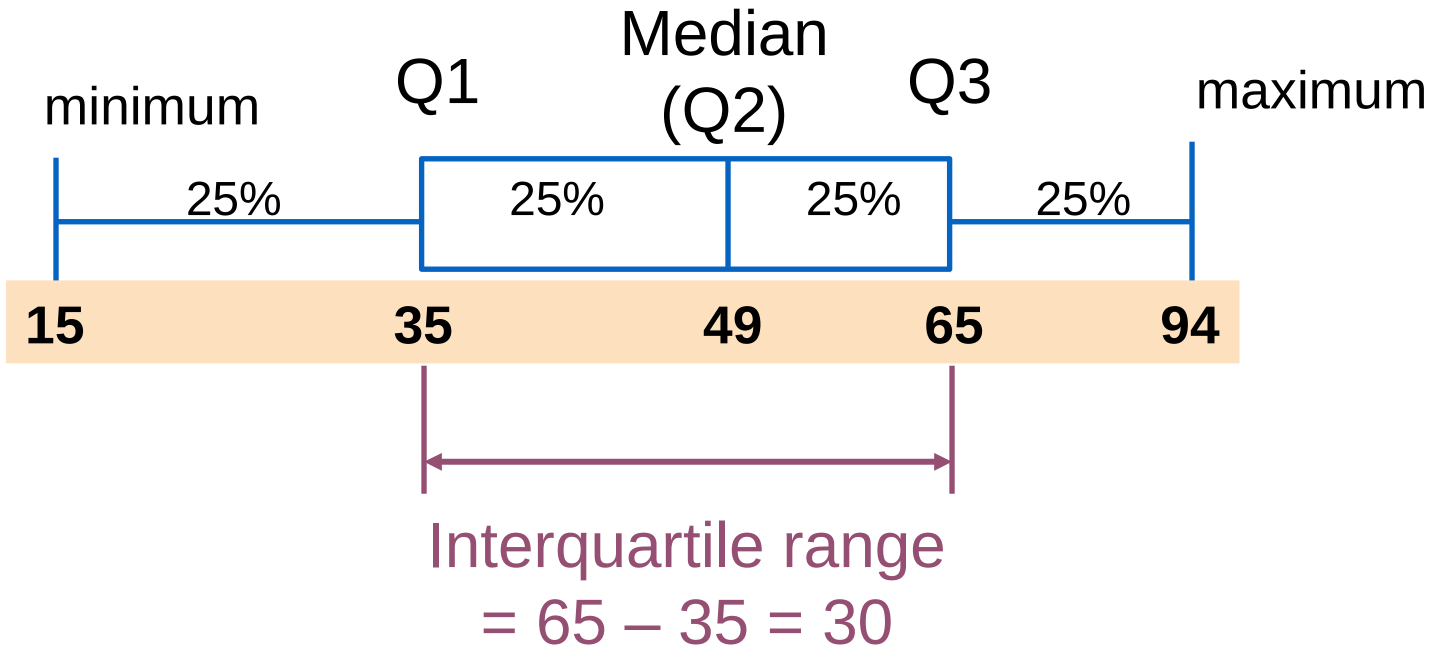


- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

Interquartile Range

Interquartile range = 3rd quartile – 1st quartile
$= Q_3 - Q_1$

Interquartile Range: age



Variance

- Average (roughly) of squared deviations of values from the mean

$$S^2 = \frac{\sum_{i}^{n} (x_i - \bar{X})^2}{n - 1}$$

Why squared deviations?

- Adding deviations will yield a sum of 0.
- Absolute values are tricky!
- Squares eliminate the negatives.

- Result:
  - Increasing contribution to the variance as you go farther from the mean.

Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

$$S = \sqrt{\frac{\sum_{i}^{n} (x_i - \overline{X})^2}{n - 1}}$$

Calculation Example: Sample Standard Deviation

Age data (n=8) : 17 19 21 22 23 23 23 38

n = 8 Mean = $\bar{X}$ = 23.25

$$S = \sqrt{\frac{(17 - 23.25)^2 + (19 - 23.25)^2 + \cdots + (38 - 23.25)^2}{8 - 1}}$$

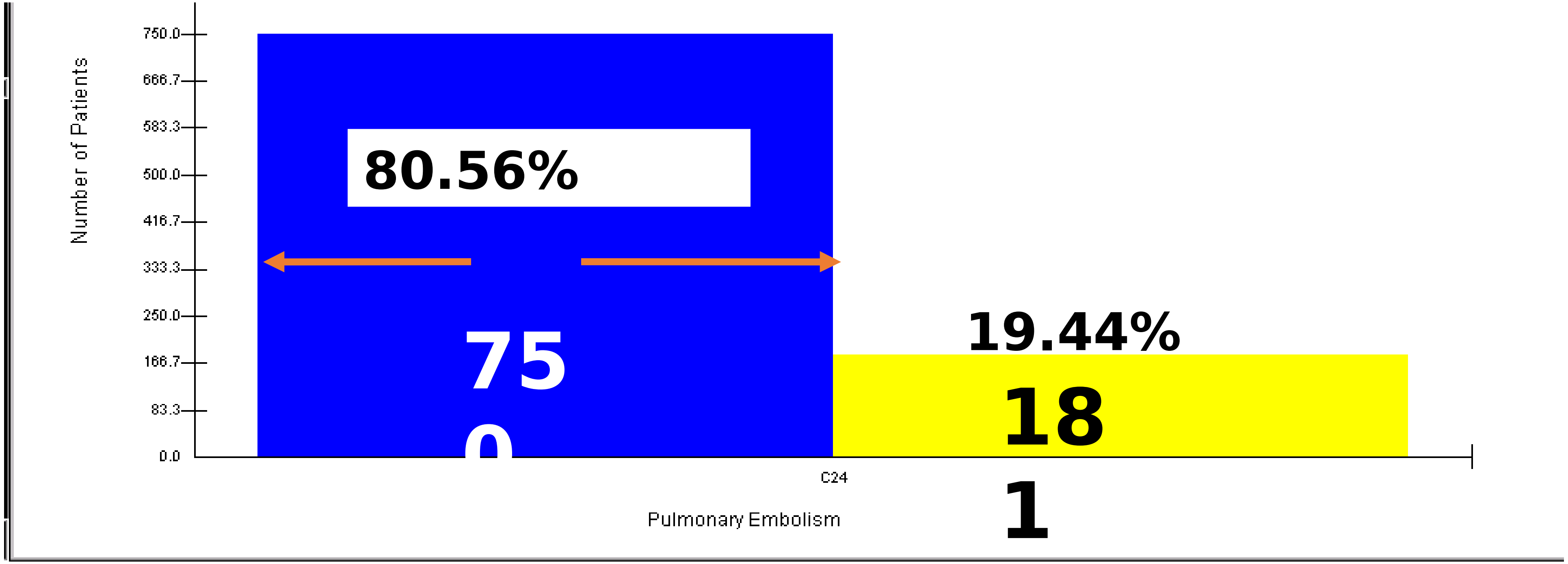
$$= \sqrt{\frac{280}{7}} = 6.3$$

Std. Dev of binary variable

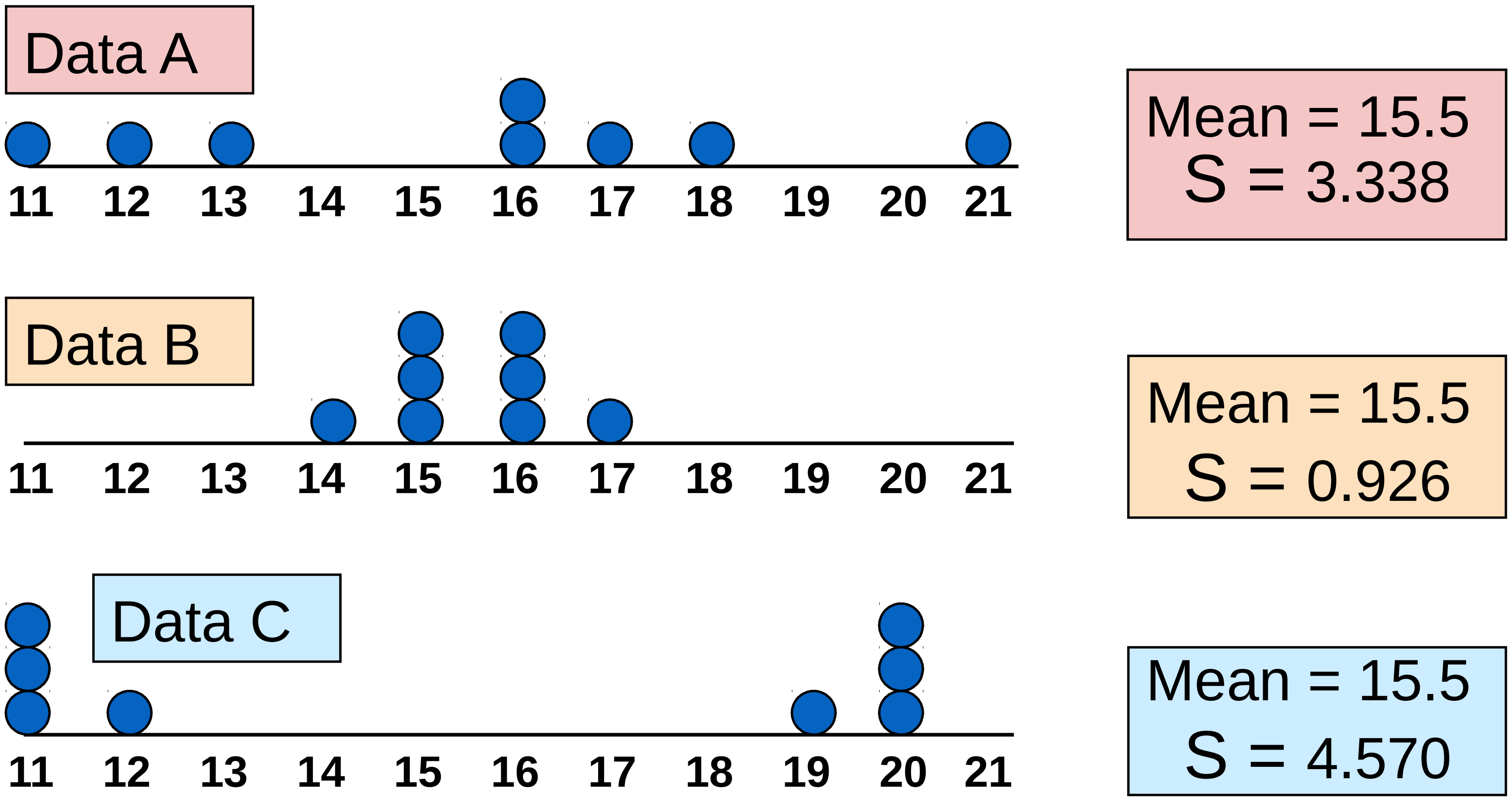
$$S = \sqrt{\frac{181 * (1 - .1944)^2 + 750 * (0 - .1944)^2}{931 - 1}}$$

$$= \sqrt{\frac{145.8}{930}} = .3959$$

Std. dev is a measure of the "average" scatter around the mean.



Comparing Standard Deviations



Bienaymé-Chebyshev Rule

- Regardless of how the data are distributed, a certain percentage of values must fall within K standard deviations from the mean:

Note use of μ (mu) to represent “mean”.

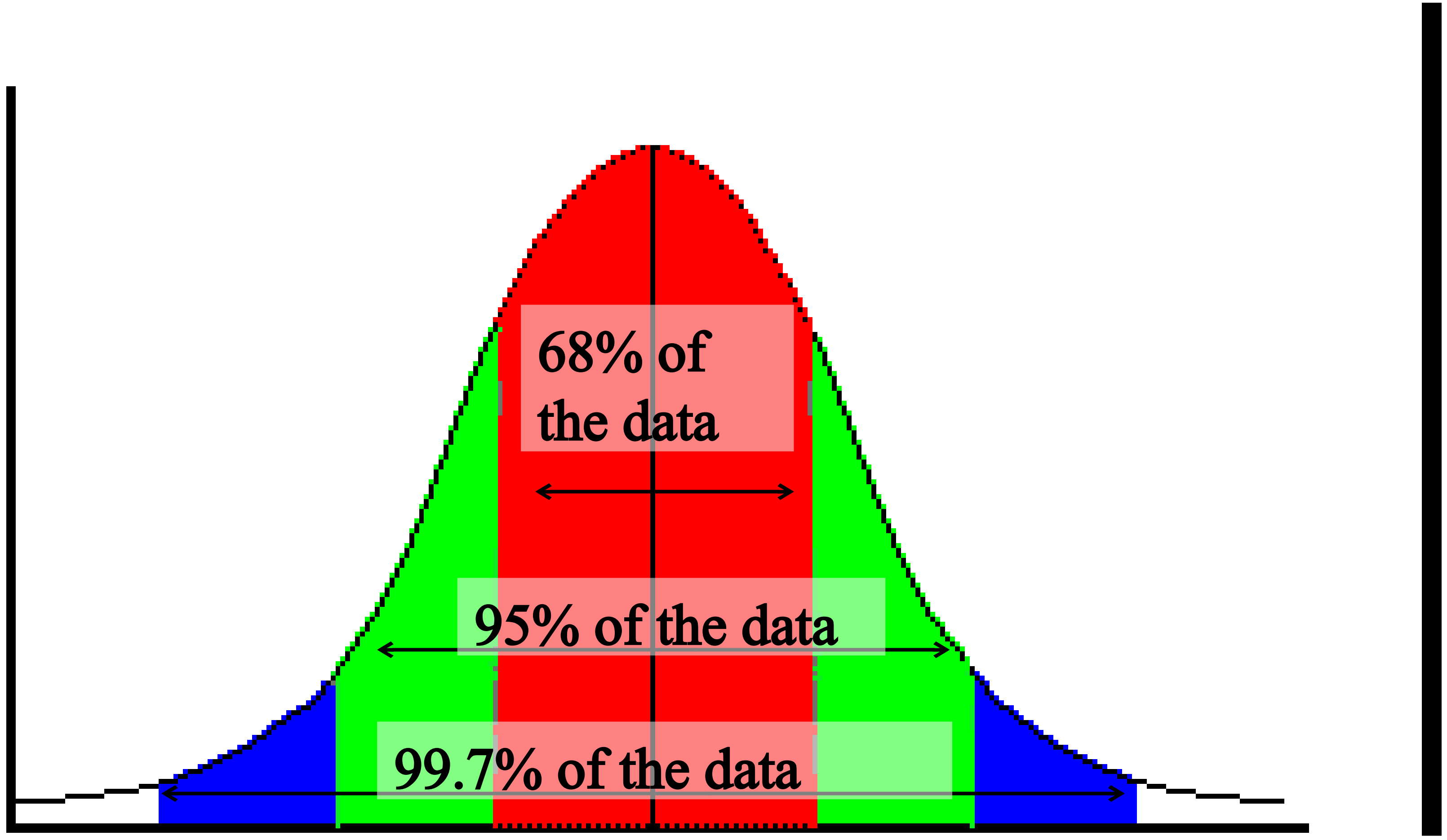
Note use of σ (sigma) to represent “standard deviation.”

At least	within
$(1 - 1/1^2) =$ 0%	$k=1$ $(\mu \pm 1\sigma)$
$(1 - 1/2^2) =$ 75%	$k=2$ $(\mu \pm 2\sigma)$

Symbol Clarification

- S = Sample standard deviation (example of a "sample statistic")
- σ = Standard deviation of the entire population (example of a "population parameter") or from a theoretical probability distribution
- X = Sample mean
- μ = Population or theoretical mean

68-95-99.7 Rule



Plots: Frequency Plots

Categorical variables

- Bar Chart

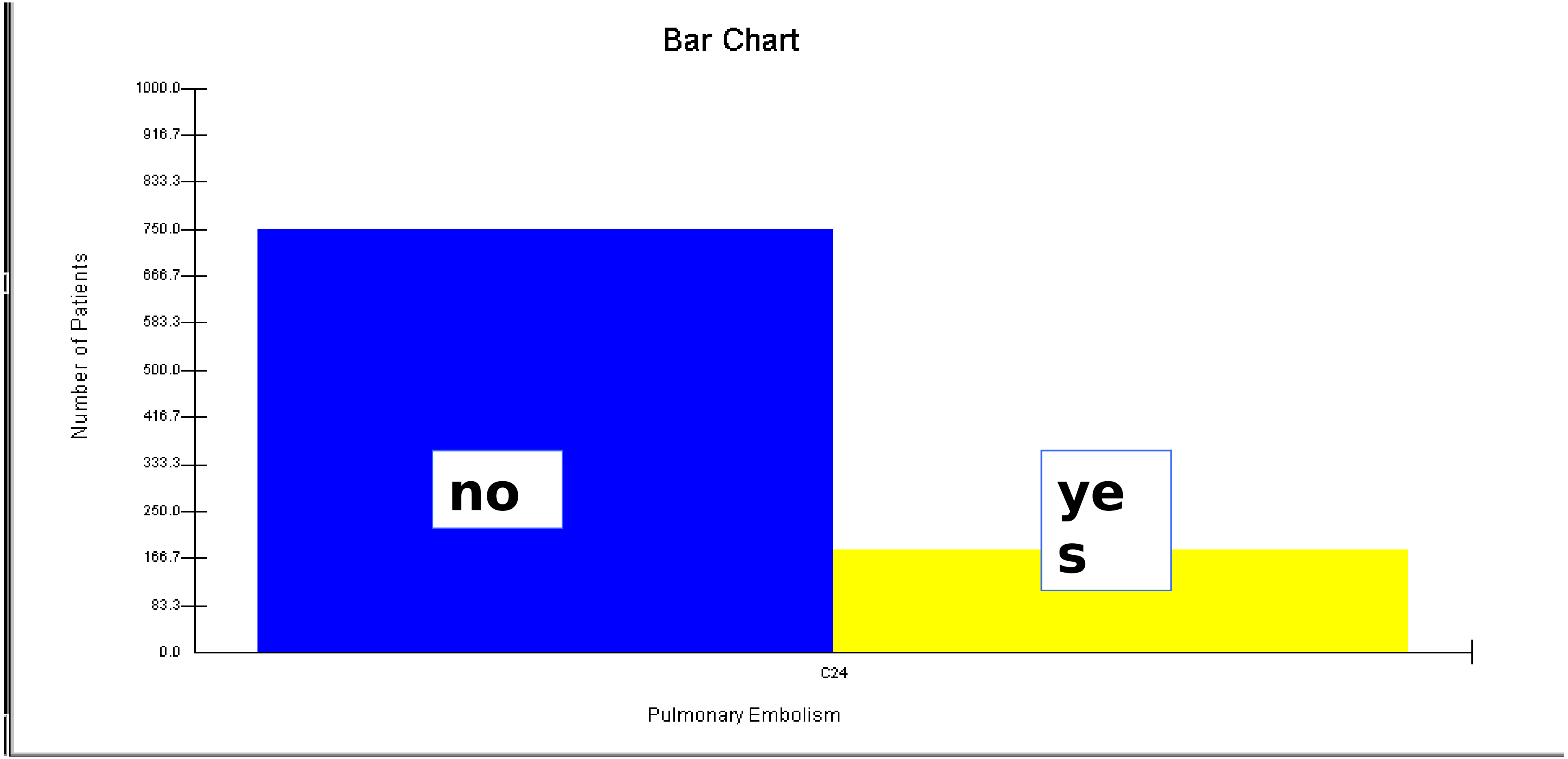
Continuous variables

- Box Plot
- Histogram

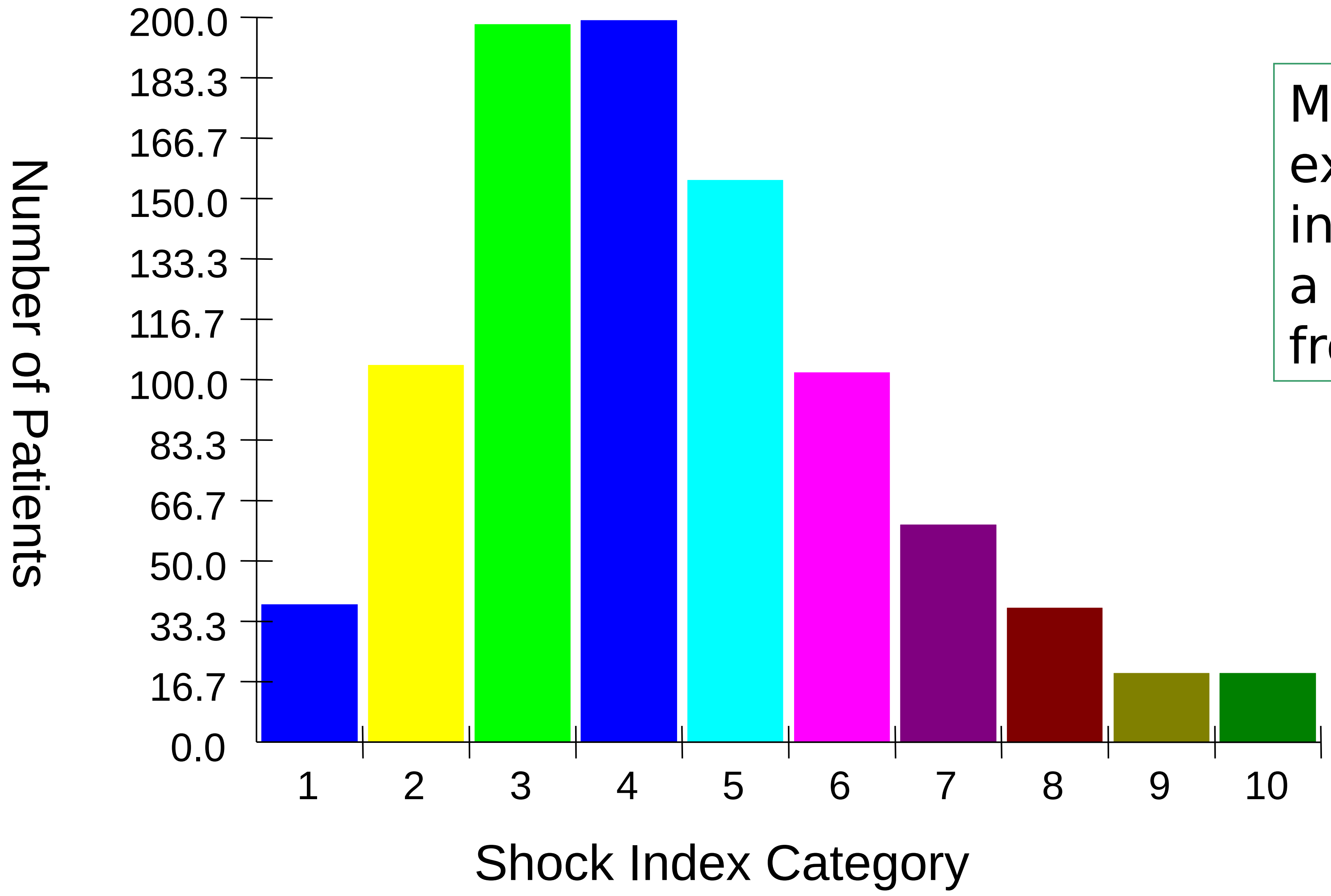
Bar Chart

- Used for categorical variables to show frequency or proportion in each category.
- Translate the data from frequency tables into a pictorial representation…

Bar Chart: binary categorical variables



Bar Chart: nominal categorical variables

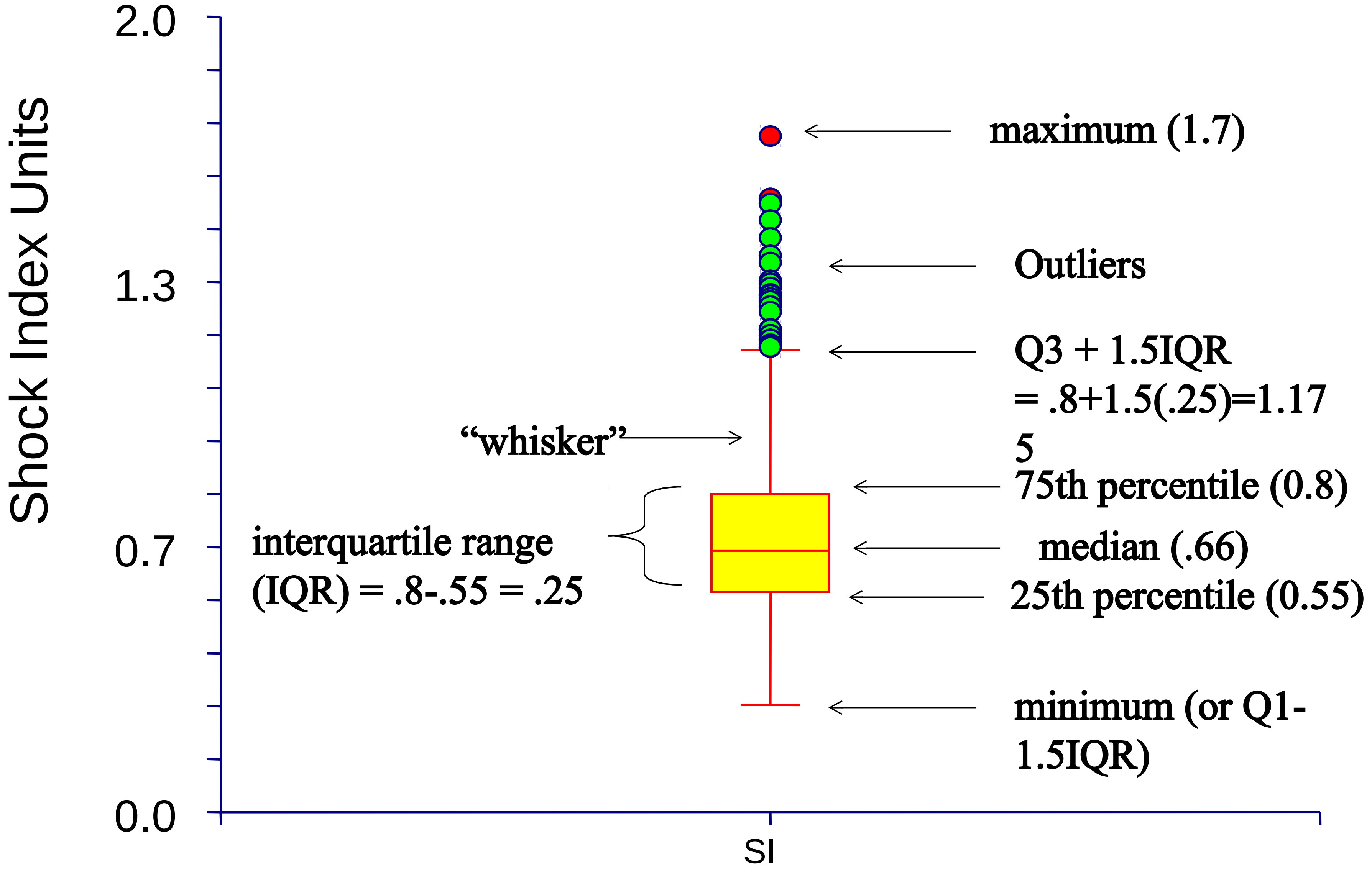


Much easier to extract information from a bar chart than from a table!

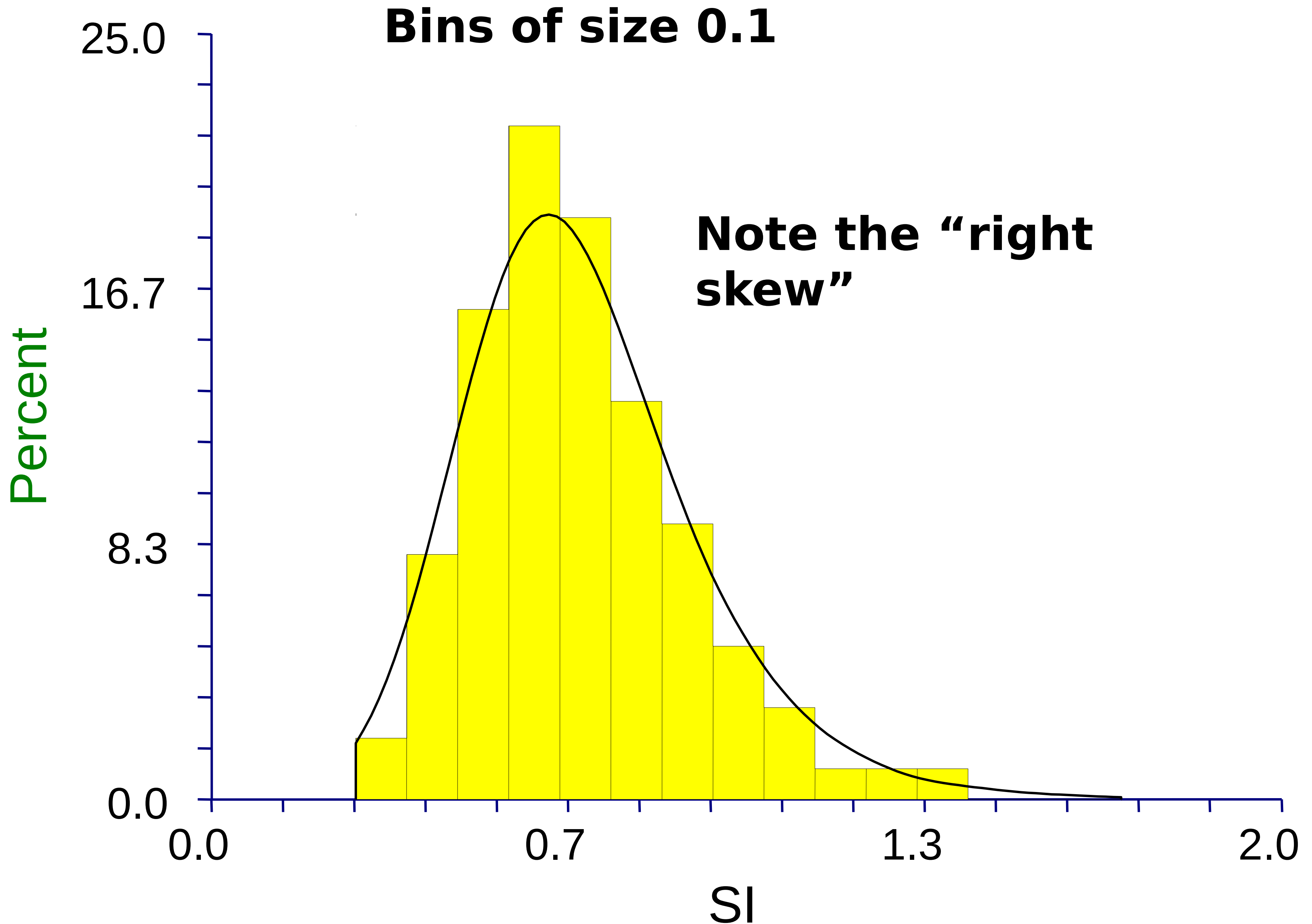
Box plot and histograms: for continuous variables

- Robust Bar chart for continuous variables
- Reveal the underlying distribution
- To show the distribution parameters
    - shape, center, range, variation of continuous variables.

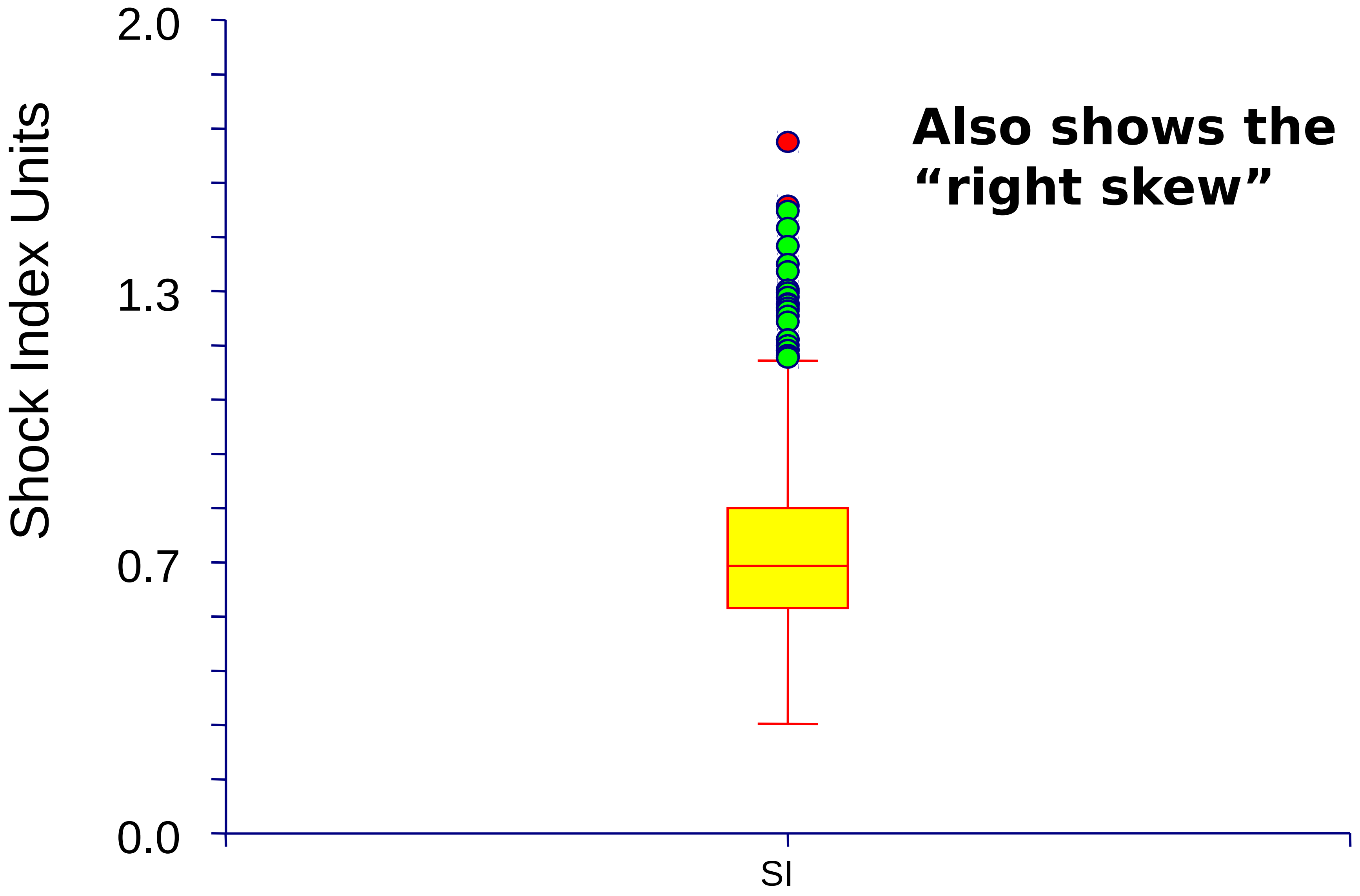
Box Plot: Shock Index



Histogram of SI

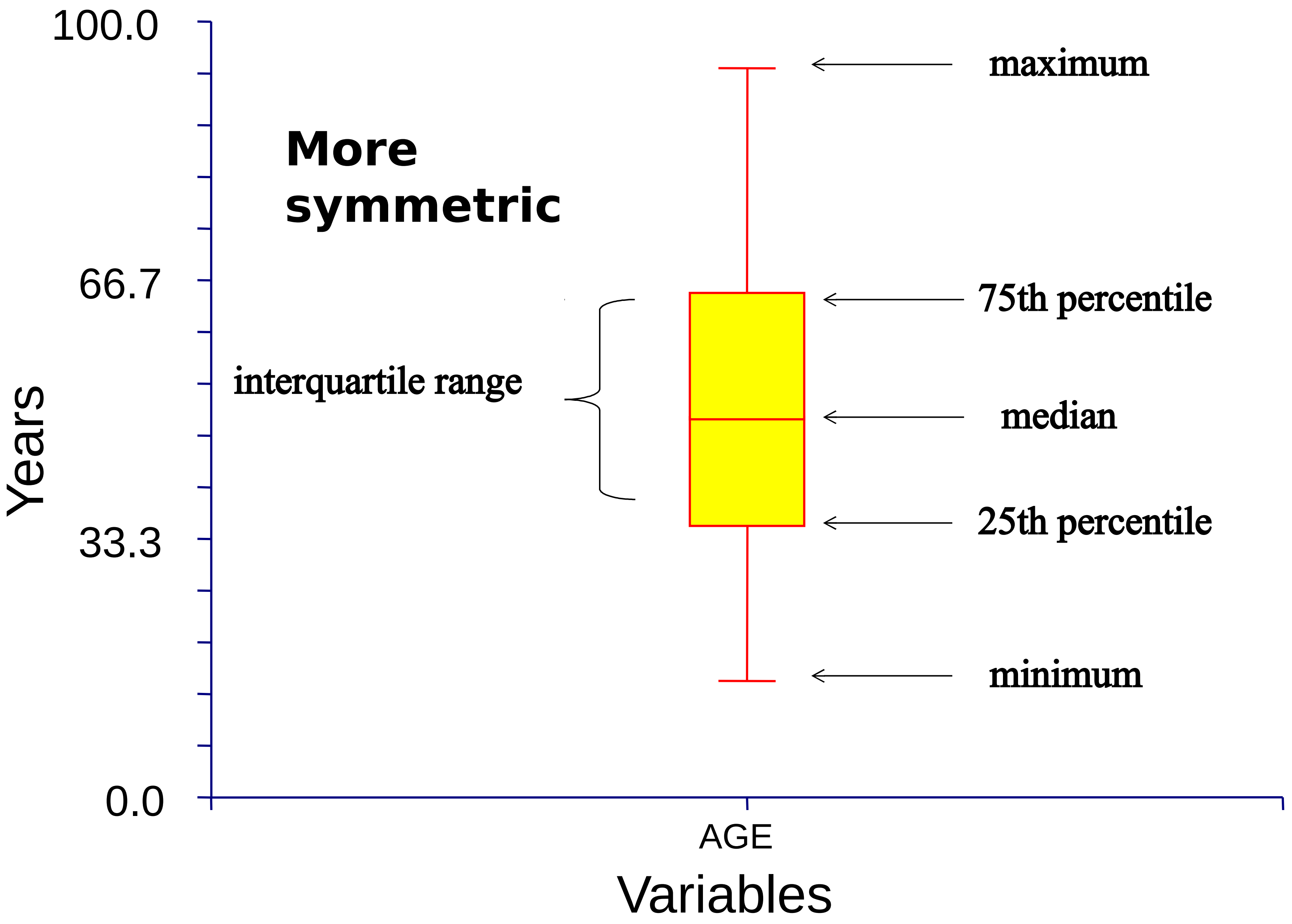


Box Plot: Shock Index



Also shows the "right skew"

Box Plot: Age



100.0
66.7
33.3
0.0
Years
More symmetric
interquartile range
maximum
75th percentile
median
25th percentile
minimum
AGE
Variables

Probability

Probability

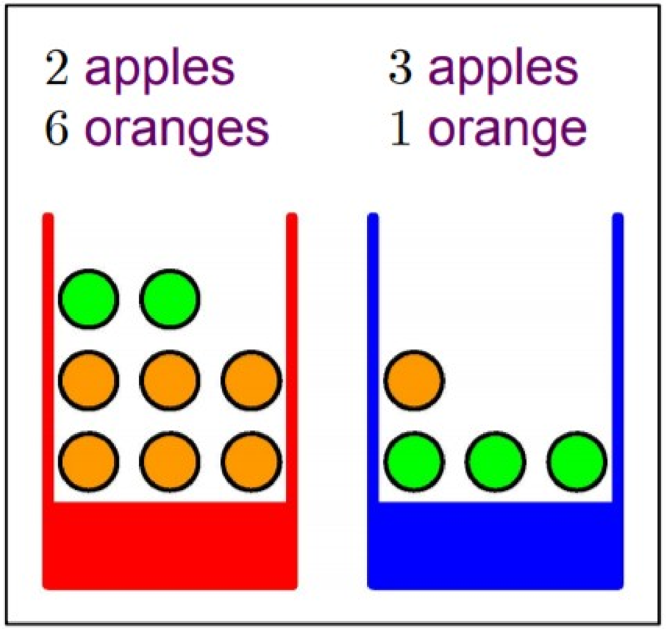
- Probability is key concept in dealing with uncertainty
  - Arises due to finite size of data sets and noise on measurements
- Probability Theory
  - Framework for quantification and manipulation of uncertainty
  - One of the central foundations of machine learning

Random Variables

- Takes values subject to chance –
  - E.g., **X** is the result of coin toss with values Head and Tail which are non - numeric
- X can be denoted by a random variable ***X*** which has values of 1 and 0
  - Each value of x has an associated probability
- Probability Distribution
  - Mathematical function that describes
    - Possible values of a random variable
    - Associated probabilities

Probability with two variables

- Key concepts:
  - conditional & joint probabilities of variables
- Random Variables: B and F
  - Box B, Fruit F
    - F has two values orange (o) or apple (a)
    - B has values red (r) or blue (b)



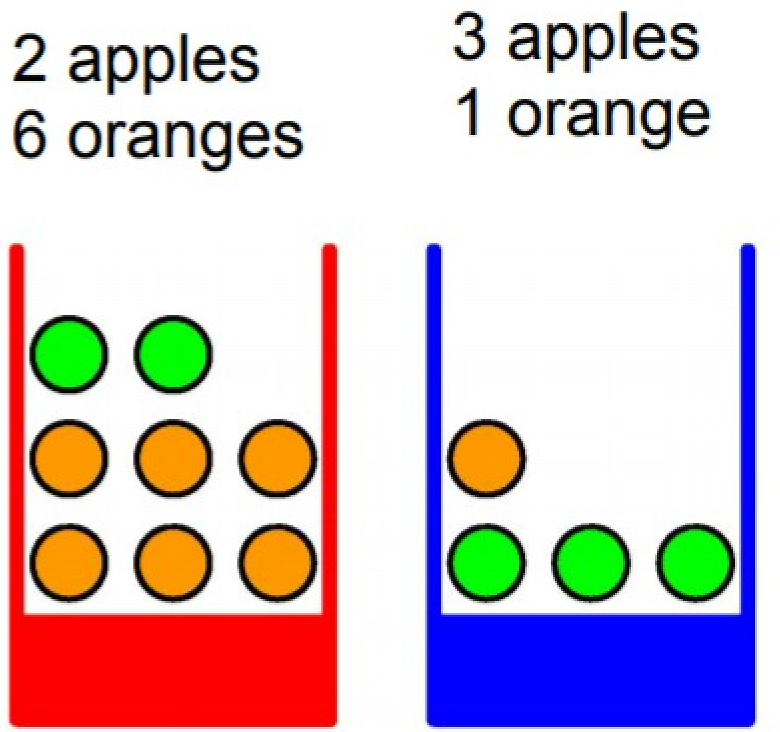
$P(F{=}o){=}3/4$ and $P(F{=}a){=}1/4$

Let $p(B{=}r){=}4/10$ and $p(B{=}b){=}6/10$

Given the above data we are interested in several probabilities of interest: *marginal, conditional and joint* Described next

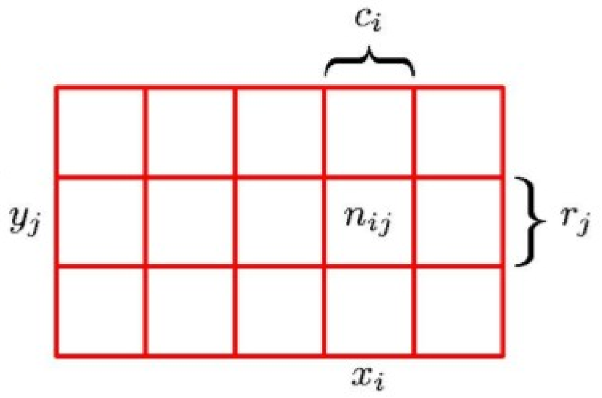
Probabilities of interest

- Marginal Probability
  - What is the probability of an apple? P(F=a)
- Note that we have to consider P(B)
- Conditional Probability
  - Given that we have an orange what is the probability that we chose the blue box? P(B=b|F=o)
- Joint Probability
  - What is the probability of orange AND blue box? P(B=b,F=o)



Sum rule of probability

- Consider two random variables
- X can take on values $x_i,\ i=1,,\ M$
- Y can take on values $y_i,\ i=1,..L$
- N trials sampling both X and Y
- No of trials with $X=x_i$ and $Y=y_j$ is n_{ij}



Joint Probability $p(X = x_i, Y = y_j) = \dfrac{n_{ij}}{N}$

- Marginal Probability $p(X = x_i) = \dfrac{c_i}{N}$

Since $c_i = \sum_j n_{ij}$, $\boxed{p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)}$

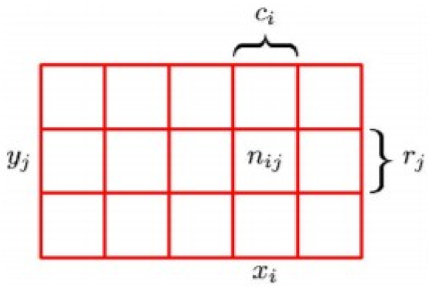
Product rule of probability

- Consider only those instances for which $X = x_i$
- Then fraction of those instances for which $Y = y_j$ is written as $p(Y = y_j | X = x_i)$
- Called conditional probability
- Relationship between joint and conditional probability:

$$p(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$$

$$\boxed{p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}} = \frac{n_{ij}}{ci} \bullet \frac{c_i}{N}$$

$$= \boxed{p(Y = y_j \mid X = x_i)\, p(X = x_i)}$$



Baye's theorem

- From the product rule together with the symmetry property $p(X,Y)=p(Y,X)$ we get

$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)}$$

- Which is called Bayes' theorem
- Using the sum rule the denominator is expressed as

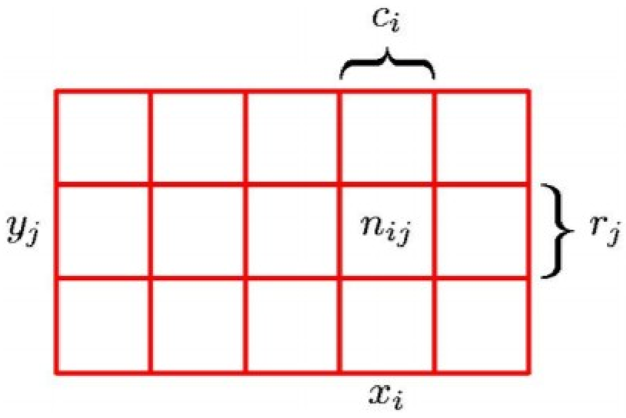
$$p(X) = \sum_{Y} p(X \mid Y)p(Y)$$

Normalization constant to ensure sum of conditional probability on LHS sums to 1 over all values of Y

Rules of probability

- Given random variables X and Y
- Sum Rule gives Marginal Probability

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j) = \frac{c_i}{N}$$



- Product Rule: joint probability in terms of conditional and marginal

$$p(X,Y) = \frac{n_{ij}}{N} = p(Y \mid X)p(X) = \frac{n_{ij}}{c_i} \times \frac{c_i}{N}$$

- Combining we get Bayes Rule

$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)}$$

where

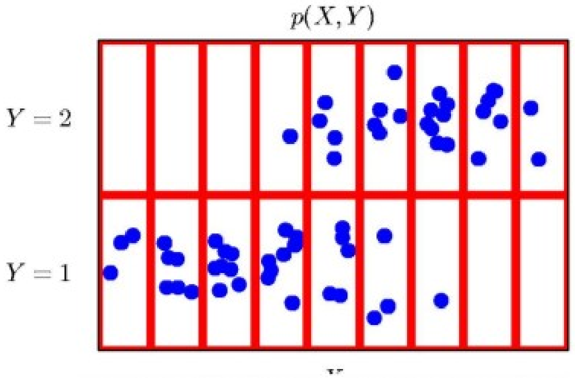
$$p(X) = \sum_{Y} p(X \mid Y)p(Y)$$

Viewed as
Posterior a likelihood x prior

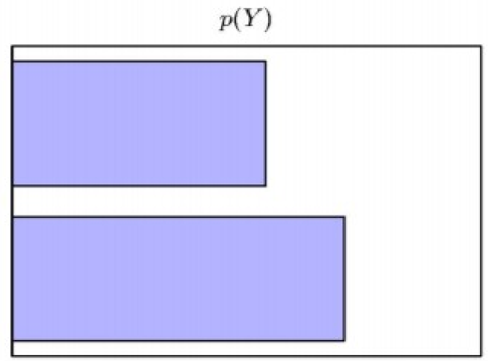
Joint distribution over two random variables

X takes nine possible values, Y takes two values

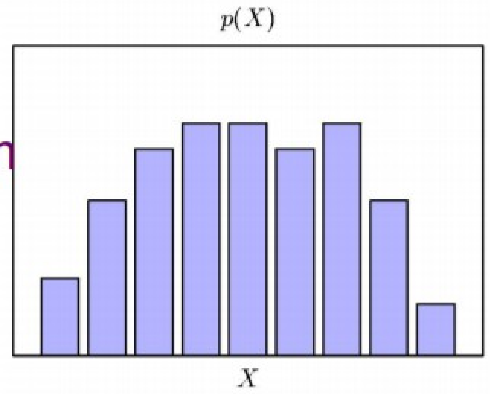
$N = 60$ data points



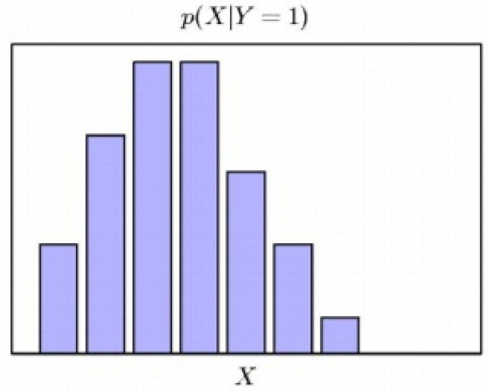
Histogram of Y (Fraction of data points having each value of Y)



Histogram of X



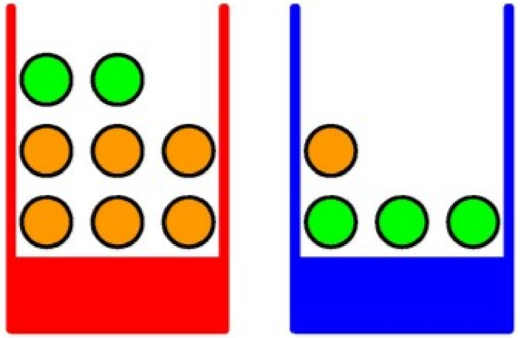
Histogram of X given Y=1



Fractions would equal the probability as $N \rightarrow$oo

Baye's rule example

- Probability that box is red given that fruit picked is orange



$$p(B=r \mid F=o) = \frac{p(F=o \mid B=r)\,p(B=r)}{p(F=o)}$$

$$= \frac{\frac{3}{4} \times \frac{4}{10}}{\frac{9}{20}} = \boxed{\frac{2}{3} = 0.66}$$

The *a posteriori* probability of 0.66 is different from the *a priori* probability of 0.4

- Probability that fruit is orange
  - From sum and product rules

$$p(F=o) = p(F=o, B=r) + p(F=o, B=b)$$

$$= p(F=o \mid B=r)\,p(B=r) + p(F=o \mid B=b)\,p(B=b)$$

$$= \frac{6}{8} \times \frac{4}{10} + \frac{1}{4} \times \frac{6}{10} = \boxed{\frac{9}{20} = 0.45}$$

The *marginal* probability of 0.45 is lower than average probability of 7/12=0.58

Independent variables

- If $p(X,Y)=p(X)p(Y)$ then X and Y are said to be independent
- Why?
- From product rule: $\boxed{p(Y \mid X) = \frac{p(X,Y)}{p(X)} = p(Y)}$

- In fruit example if each box contained same fraction of apples and oranges then $p(F|B)=p(F)$

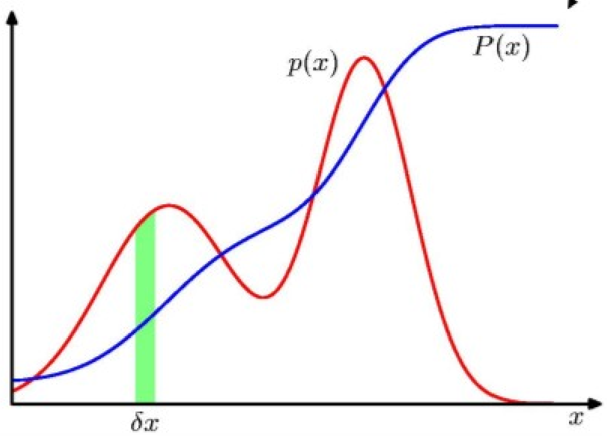
Probability density function

- Continuous Variables
- If probability that x falls in interval $(x, x+\delta x)$ is given by $p(x)\,dx$ for $\delta x \rightarrow 0$

  then $p(x)$ is a pdf of x
- Probability x lies in interval (a,b) is

$$p(x \in (a,b)) = \int_a^b p(x)\,dx$$

Cumulative Distribution Function



Probability that x lies in Interval $(-\infty, z)$ is

$$P(z) = \int_{-\infty}^{z} p(x)\,dx$$

Several variables

- If there are several continuous variables $x_1, \ldots, x_D$ denoted by vector $\mathbf{x}$ then we can define a joint probability density $p(\mathbf{x}) = p(x_1, \ldots, x_D)$
- Multivariate probability density must satisfy

$$p(\mathbf{x}) \geq 0$$

$$\int_{-\infty}^{\infty} p(\mathbf{x})\, d\mathbf{x} = 1$$

Expectation

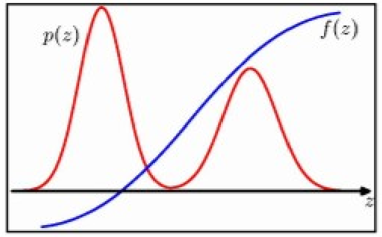
- Expectation is *average* value of some function $f(x)$ under the probability distribution $p(x)$ denoted $E[f]$
- For a discrete distribution

$$E[f] = \sum_x p(x)\, f(x)$$

- For a continuous distribution

$$E[f] = \int p(x) f(x)\, dx$$

Examples of $f(x)$ of use in ML:



$f(x)=x$; $E[f]$ is mean
$f(x)=\ln p(x)$; $E[f]$ is entropy
$f(x)=-\ln[q(x)/p(x)]$; K-L divergence

- If there are N points drawn from a pdf, then expectation can be approximated as

$$E[f] = (1/N)\sum_{n=1}^{N} f(x_n)$$

This approximation is extremely important when we use
sampling to determine expected value

- Conditional Expectation with respect to a conditional distribution

$$E_x[f] = \sum_x p(x|y)\, f(x)$$

Variance

- Measures how much variability there is in $f(x)$ around its mean value $E[f(x)]$
- Variance of $f(x)$ is denoted as

$$\mathrm{var}[f] = E[(f(x) - E[f(x)])^2]$$

- *Expanding the square*

$$\mathrm{var}[f] = E[(f(x)^2] - E[f(x)]^2$$

- Variance of the variable x itself

$$\mathrm{var}[x] = E[x^2] - E[x]^2$$