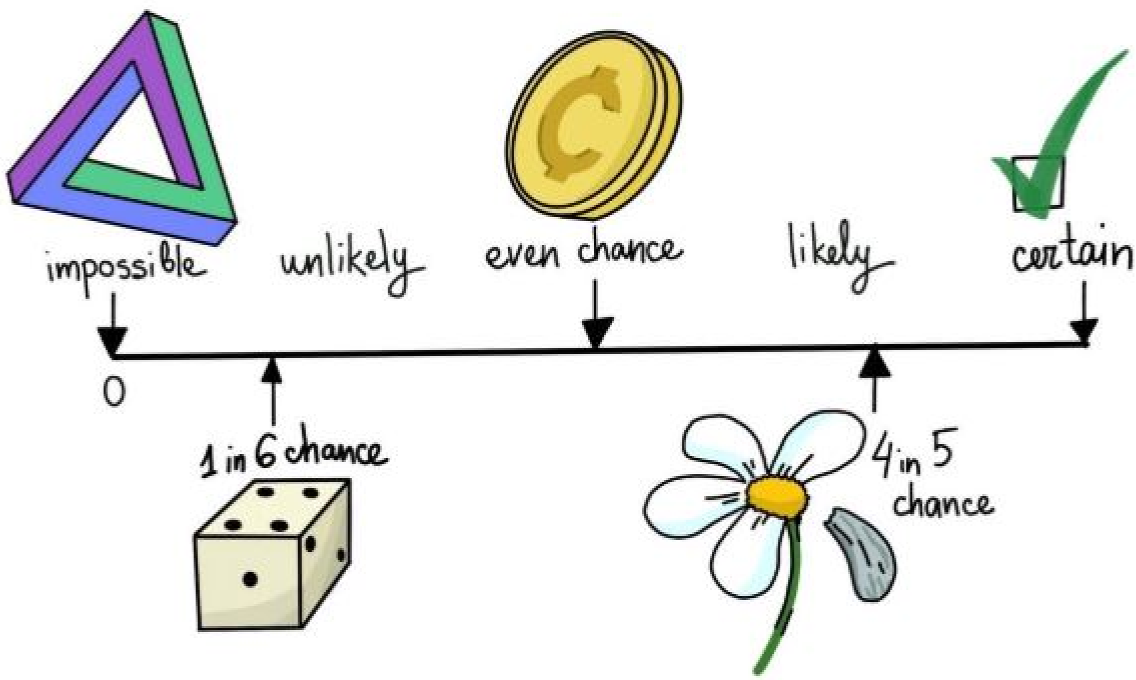
DATA AND DISTRIBUTIONS

Why Use Probability?



- data features are stochastic
- results are statistical

figure reprinted from www.kdnuggets.com with explicit permission of the editor

Discrete vs. Continuous

- discrete variables:

(a) has countable values $v_1, \ldots, v_K$
(b) values associated with probabilities $p_1, \ldots, p_K$
(c) example: age (in years)

- continuous variables:

(a) can take any value in its range
(b) probabilities described by a density function
(c) example: height

Discrete Data

- data $X = \{x_1, \ldots, x_N\}$
- assume sorted $x_1 \leq \cdots \leq x_N$
- probability p_i for each x_i
- how do we describe X?
- mean $\mu(X)$:

$$\mu = p_1 x_1 + \cdots + p_N x_N$$

- standard deviation:

$$\sigma^2(X) = p_1(x_1 - \mu)^2 + \cdots + p_N(x_N - \mu)^2$$

- mode: most frequent value
- median: value at position $\lfloor N/2 \rfloor$

Example: Chroline Atomic Weight

Isotope			Decay	
	abun-dance	half-life ($t_{1/2}$)	mode	pro-duct
^{35}Cl	76%	stable		
^{36}Cl	trace	3.01×10^5 y	β^-	^{36}Ar
			ε	^{36}S
^{37}Cl	24%	stable		

Standard atomic weight $A_{r,\,\text{standard}}$(Cl) [35.446, 35.457][1]
Conventional: 35.45

- Cl has two stable isotopes:
  1. Cl – 35: mass 34.9689 and probability 0.7577
  2. Cl – 37: mass 36.9653 and probability 0.2423

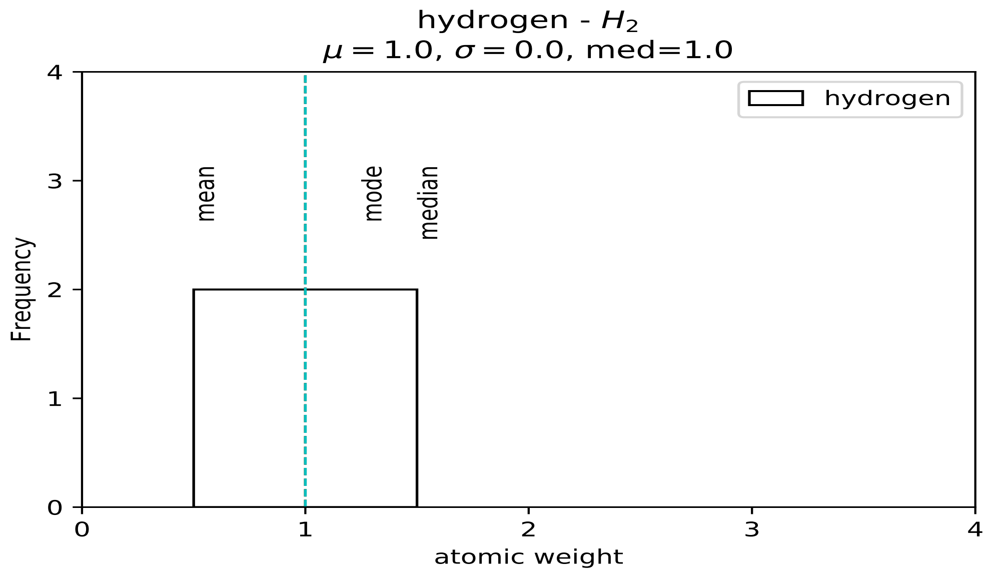
Chlorine Atomic Weight (cont'd)

Isotope			Decay	
	abun-dance	half-life ($t_{1/2}$)	mode	pro-duct
^{35}Cl	76%	stable		
^{36}Cl	trace	3.01×10^5 y	β^-	^{36}Ar
			ε	^{36}S
^{37}Cl	24%	stable		
Standard atomic weight $A_{\mathrm{r,\ standard}}$(Cl)	[35.446, 35.457][1] Conventional: 35.45			

- atomic weight W:

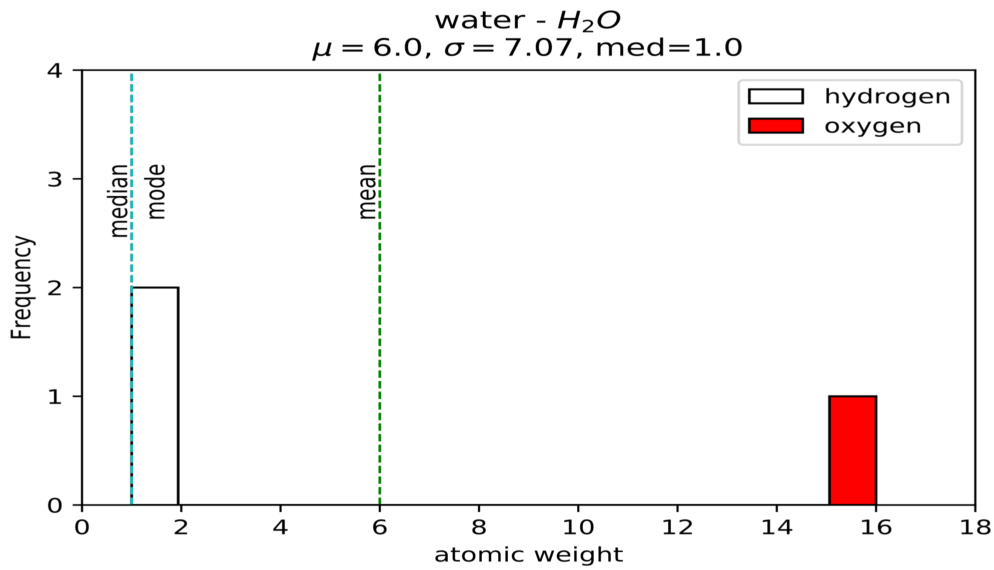
$$
\begin{aligned}
W &= 34.9689 \times 0.7577 + 36.9653 \times 0.2423 \\
&= 26.4959 + 8.9567 \\
&= 35.4526
\end{aligned}
$$

Example: Hydrogen H_2



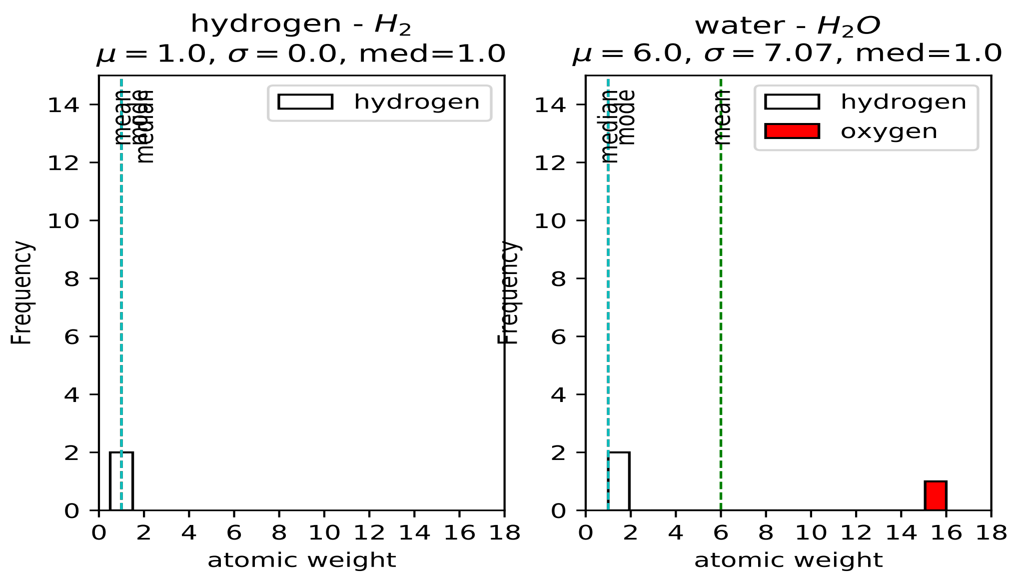
- $H_2 = \{1, 1\}$
- same mean, median, and mode
- no variation: $\sigma = 0$

Example: Water H_2O



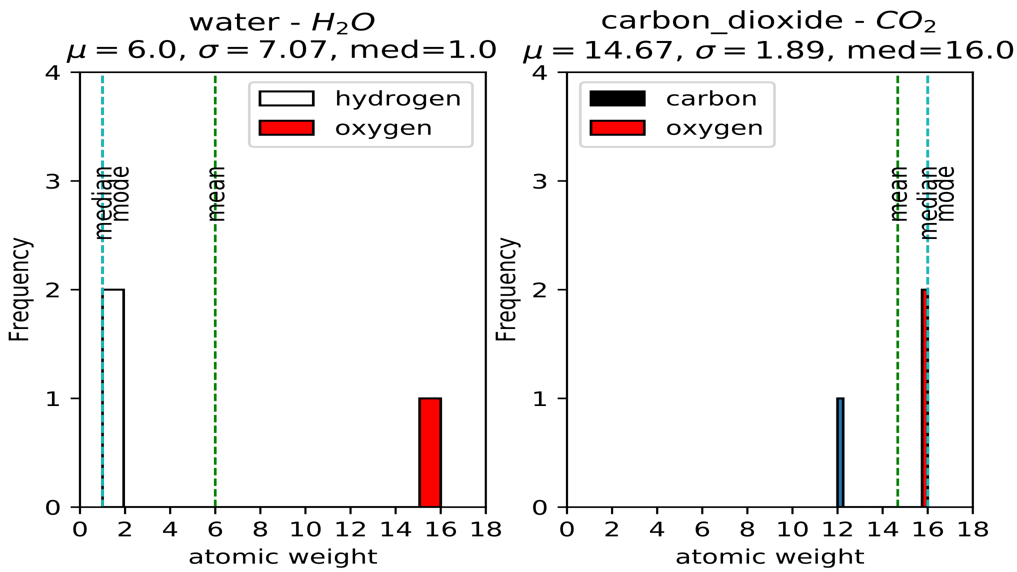
$$
\begin{aligned}
H_2O &= \{1, 1, 16\} \\
\mu(H_2O) &= \frac{1}{3} + \frac{1}{3} + \frac{16}{3} = 6 \\
\sigma^2(H_2O) &= \frac{(1-6)^2}{3} + \frac{(1-6)^2}{3} \\
&\quad + \frac{(16-6)^2}{3} = 50
\end{aligned}
$$

Hydrogen vs. Water



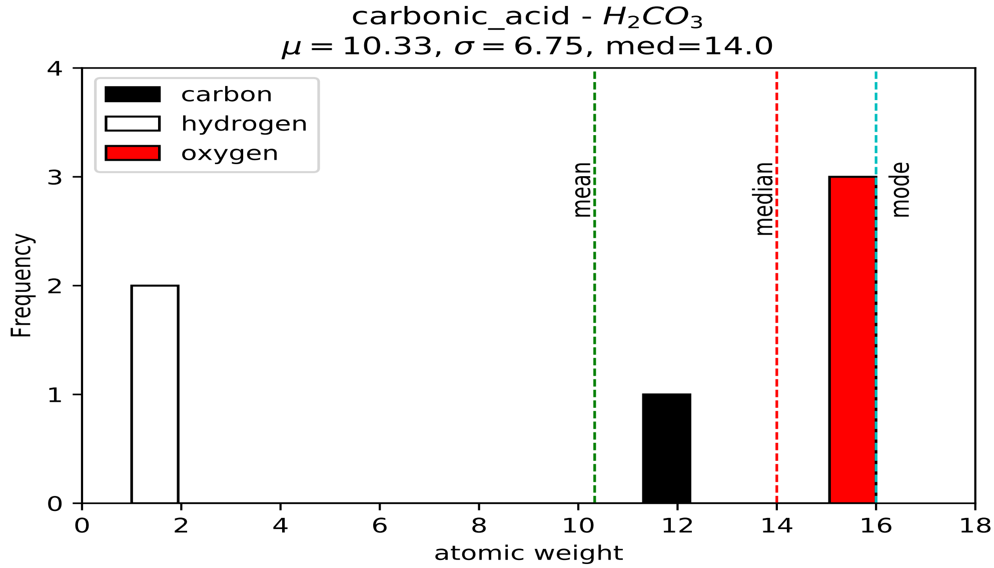
- $H_2 = \{1, 1\}$, $H_2O = \{1, 1, 16\}$
- same median, and mode
- positive $\sigma(H_2O)$

Water vs. Carbon Dioxide



- $H_2O = \{1, 1, 16\}$, $CO_2 = \{12, 16, 16\}$
- $\mu(H_2O) < \mu(CO_2)$ larger values
- $\sigma(H_2O) > \sigma(CO_2)$ less variation

Carbonic Acid H_2CO_3



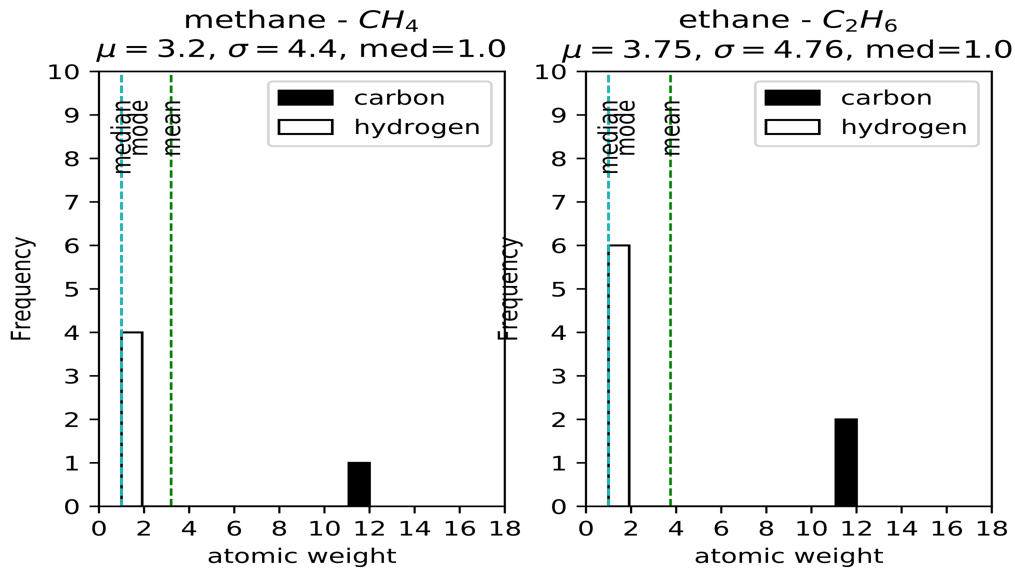
$$H_2CO_3 \leftrightarrow CO_2 + H_2O$$

$$\{1, 1, 12, 16, 16, 16\} = \{12, 16, 16\} + \{1, 1, 16\}$$

$$\mu(H_2CO_3) = 0.5 \cdot \mu(CO_2) + 0.5 \cdot \mu(H_2O)$$

- mean of a mixture

Methane and Ethane

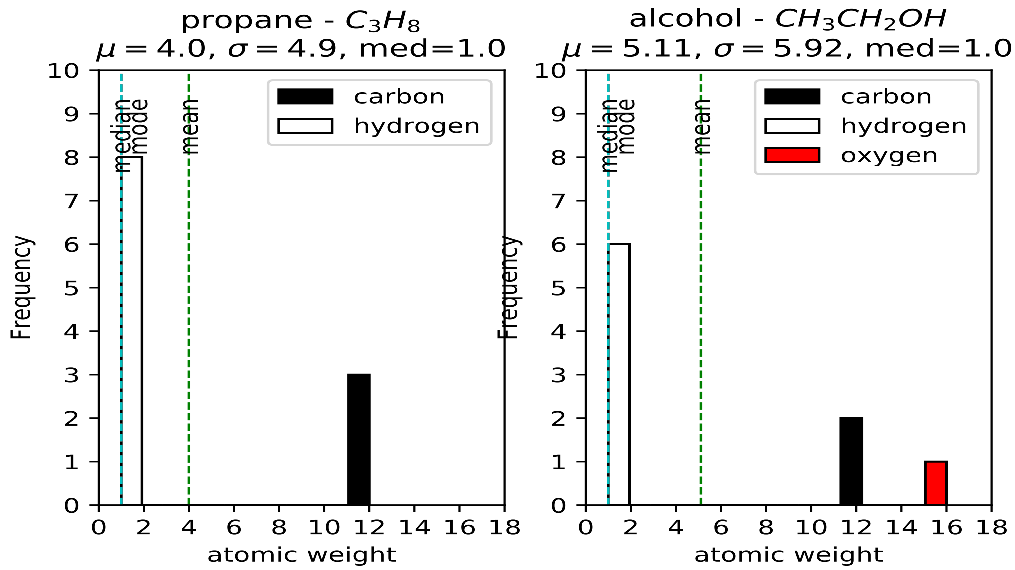


$$\text{methane } CH_4 = \{1, 1, 1, 1, 12\}$$

$$\text{ethane } C_2H_6 = \{1, 1, 1, 1, 1, 1, 12, 12\}$$

- same mode and median

Propane and Alcohol

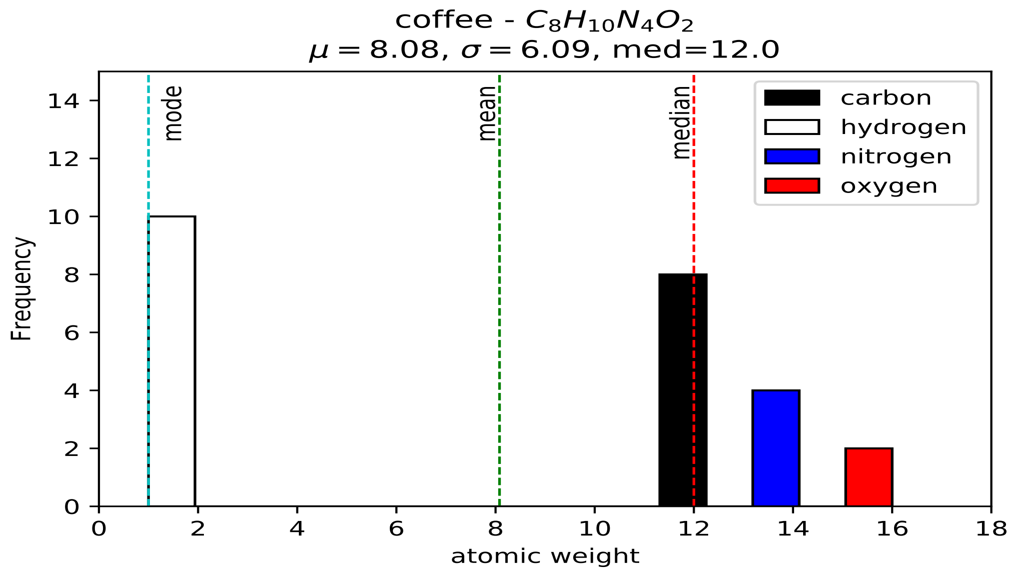


- $C_3H_8 = \{1, 1, 1, 1, 1, 1, 1, 1, 12, 12, 12\}$
- $CH_3CH_2OH = \{1, 1, 1, 1, 1, 1, 12, 12, 16\}$
- no change to mode and median

Coffee

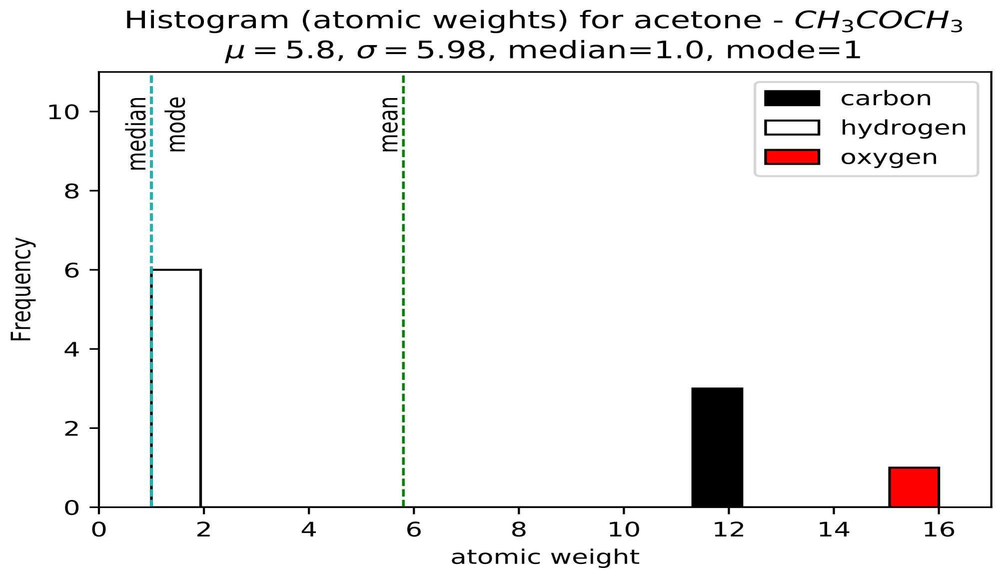
- formula: $C_8H_{10}N_4O_2$
- 24 atoms
- 16 non-hydrogen atoms

Coffee



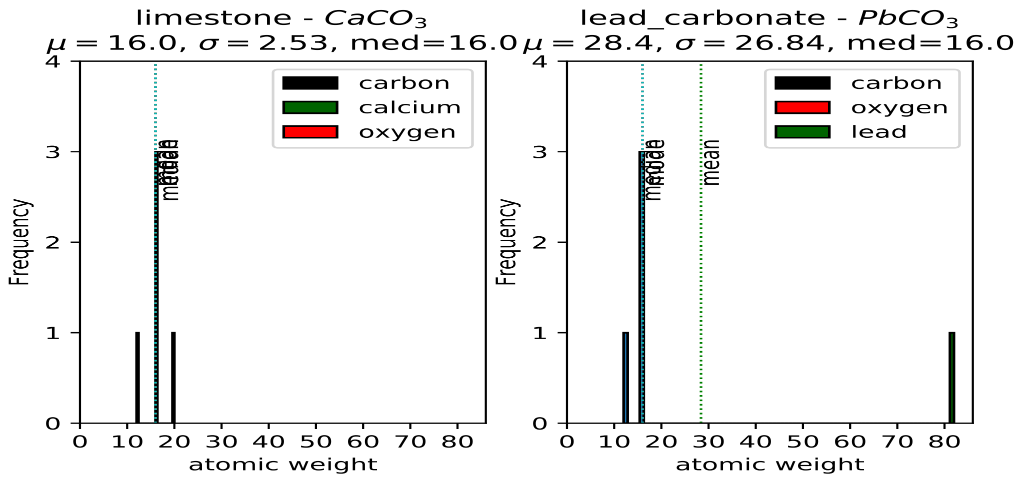
- same mode as ethane, propane, methane etc.
- $\mu <$ median
- most values larger than mean

Acetone



- formula: CH_3COCH_3
- same mode as ethane, propane, methane and coffee!
- $\mu >$ median
- most values smaller than mean

Impact of Outliers



$$CaCO_3 = \{12, 16, 16, 16, 20\}$$
$$PbCO_3 = \{12, 16, 16, 16, 82\}$$

- outliers - values outside of "normal" range
- may not impact median
- huge effect on μ, σ

Example of a Discrete Distribution

- coffeine formula: $C_8H_{10}N_4O_2$
- $K = 4$ elements $\{C, H, N, O\}$

$$P_C = \frac{8}{24}, P_H = \frac{10}{24}, P_N = \frac{4}{24}, P_O = \frac{2}{24}$$

Example of a Discrete Distribution (cont'd)

- values (atomic weights):

$$v_C = 12, v_H = 1, v_N = 14, v_O = 16$$

- 24 atoms of 4 elements
- mean weight per atom:

$$\begin{aligned}\mu &= v_C \cdot p_C + v_H \cdot p_H + v_N \cdot p_N + v_O \cdot p_O \\ &= 12 \cdot \frac{8}{24} + 1 \cdot \frac{10}{24} + 14 \cdot \frac{4}{24} + 16 \cdot \frac{2}{24} \\ &= \frac{194}{24} = 8.08\end{aligned}$$

- (weight) mode: 1 (hydrogen)
- (weight) median: 12 (carbon)

Prob. Distributions

- have sample points from X
- know distribution of $X \mapsto$ better prediction
- example: X has mean μ, variance σ^2
- Chebyshev's inequality (valid for any distribution)

$$P\Big(\mu - k\sigma \leq X \leq \mu + k\sigma\Big) \leq 1/k^2$$

Prob. Distributions (cont'd)

- for $k = 2$ for any X

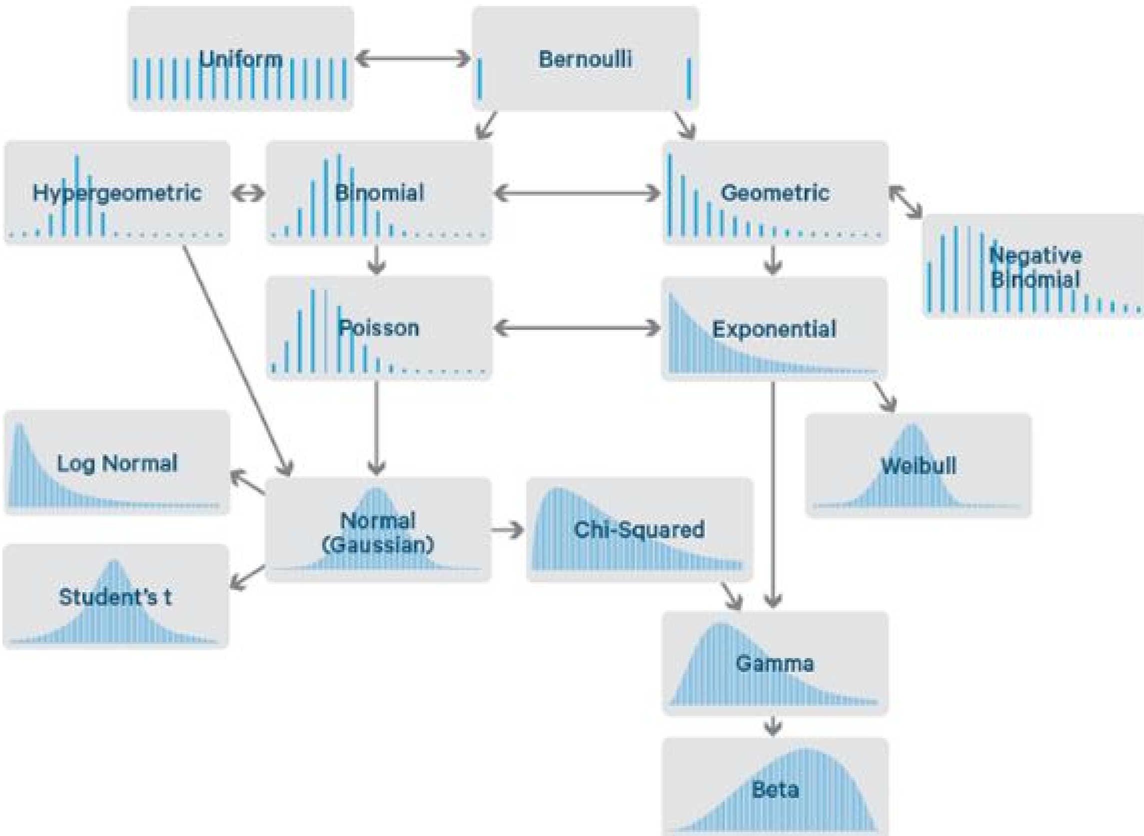
$$P\left(\mu - 2\sigma \leq X \leq \mu + 2\sigma\right) \leq 0.25$$

- suppose we know X is normal $N(\mu, \sigma)$

$$P\left(\mu - 2\sigma \leq X \leq \mu + 2\sigma\right) \leq 0.05$$

- much sharper bound
- it is important to model data

Distributions



- important for *parametric* modeling of data

figure reprinted from www.kdnuggets.com with explicit permission of the editor

Bernouilli

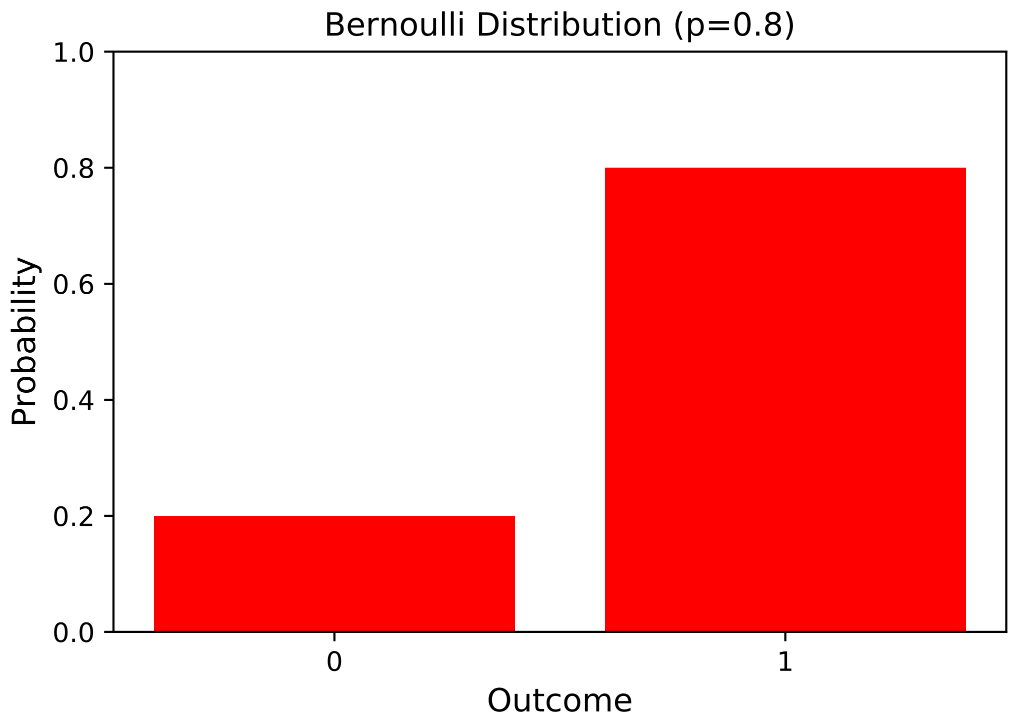
- discrete distribution
- value 1 with probability p
- value 0 with probability $q = 1 - p$
- result of a single experiment

Bernouilli (cont'd)

```
import numpy as np
import matplotlib.pyplot as plt

p = 0.8
q = 0.2
plt.xticks([1, 0])
plt.bar([1, 0], np.array([p, q]),
            color="red")
plt.title("Bernoulli (p=0.8)",
            fontsize=12)
plt.ylabel('Probability', fontsize=12)
plt.xlabel('Outcome', fontsize=12)
plt.ylim([0,1])
plt.show()
```

Bernouilli (cont'd)



Uniform

- values are equally likely
- discrete case:

(a) n values $v_1, \ldots, v_n$

(b) $P(X = v_i) = 1/n$

- continuous case:

(a) any value in interval $[a, b]$

(b) $P(a \leq X \leq b) = 1/(b - a)$

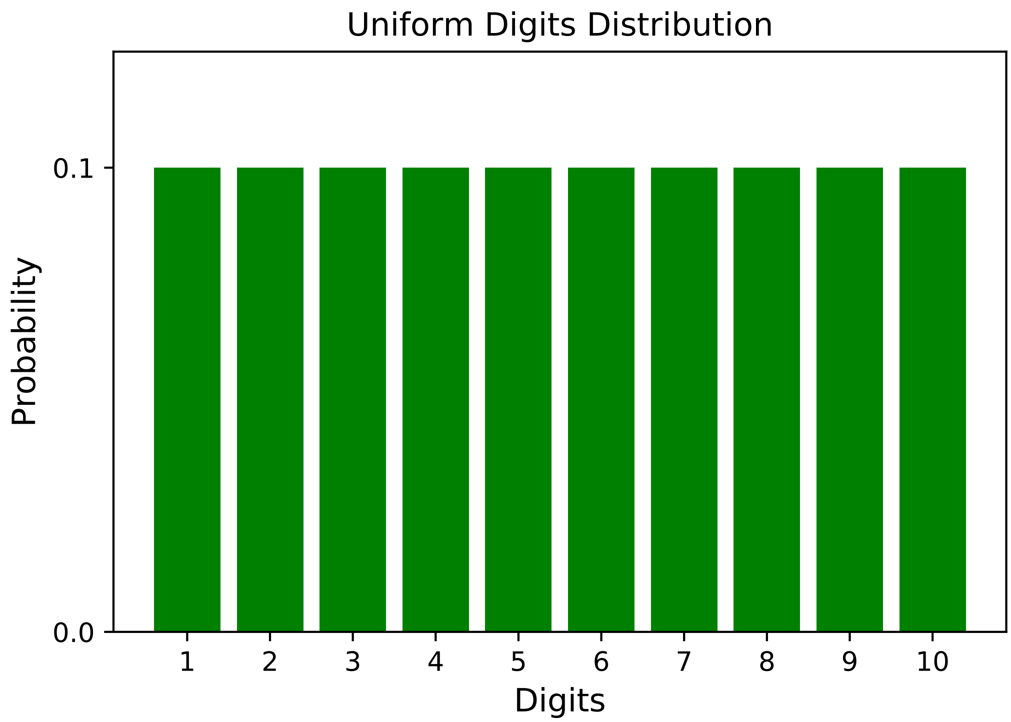
Uniform (cont'd)

```
import numpy as np
import matplotlib.pyplot as plt

prob = np.full((10), 1/10)
x = range(1,11,1)
plt.xticks(x)
plt.yticks([0, 0.1, 0.2])
plt.bar(x, prob, color="green")
plt.ylabel("Probability", fontsize=12)
plt.xlabel("Digits", fontsize=12)
plt.title("Uniform Digits Distribution",
          fontsize=12)
plt.ylim([0,0.125])
plt.show()
```

- assume every digit is equally likely

Uniform (cont'd)



Binomial

- discrete ditribution
- number m of successes in n trials
- each trial has success probability p

$$P(X = m) = \binom{n}{m} p^m (1 - p)^{n-m}$$

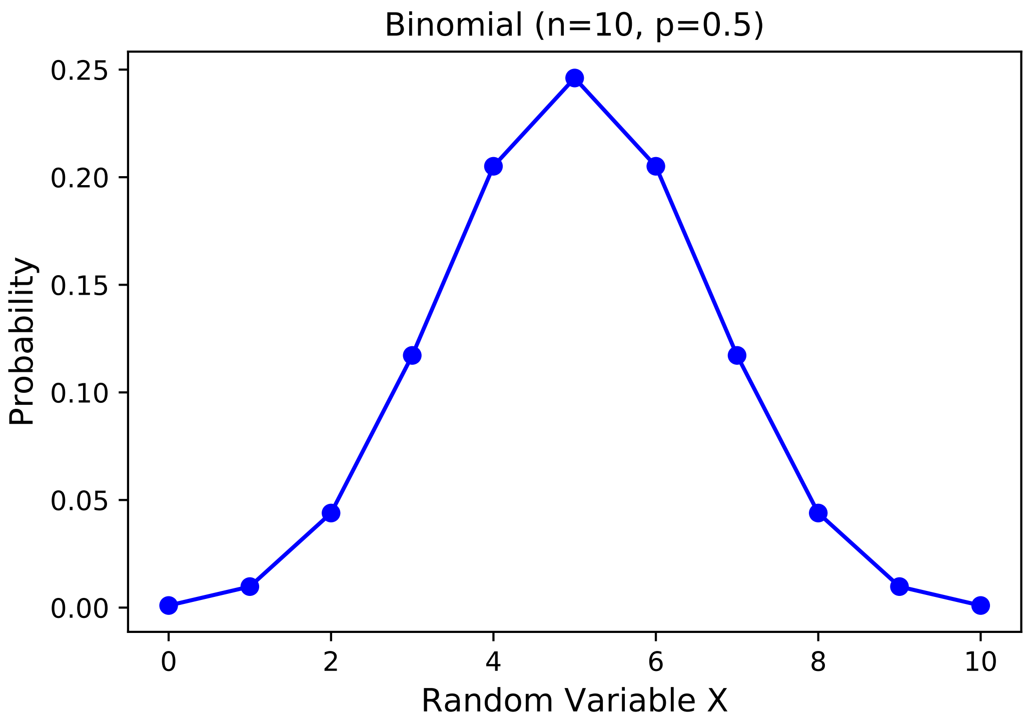
- $n = 1$ is the Bernouilli distribution
- mean: $\mu = np$
- variance: $\sigma^2 = np(1 - p)$

Binomial (cont'd)

```
import numpy as np
import matplotlib.pyplot as plt

p = 0.5
n = 10
x = np.arange(0, n + 1)
prob = stats.binom.pmf(x, n, p)
plt.plot(x, prob, "-o", color="blue")
plt.xlabel("Random Variable X",
                            fontsize=12)
plt.ylabel("Probability", fontsize=12)
plt.title("Binomial (n=10, p=0.5)")
plt.show()
```

Binomial (cont'd)



Poisson

- discrete distribution
- independent events
- events per time is constant λ
- prob. of k events in time T:

$$P\Big(X = k\Big) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- mean is λ, variance is λ

Poisson Example

- average number of goals in World Cup is 2.5
- model as Poisson $\lambda = 2.5$

$$P(k=0) = \frac{\lambda^0 e^{-\lambda}}{0!} = \frac{2.5^0 e^{-\lambda}}{0!} = 0.082$$

$$P(k=1) = \frac{\lambda^1 e^{-\lambda}}{1!} = \frac{2.5^1 e^{-\lambda}}{1!} = 0.205$$

$$P(k=2) = \frac{\lambda^2 e^{-\lambda}}{2!} = \frac{2.5^2 e^{-\lambda}}{2!} = 0.257$$

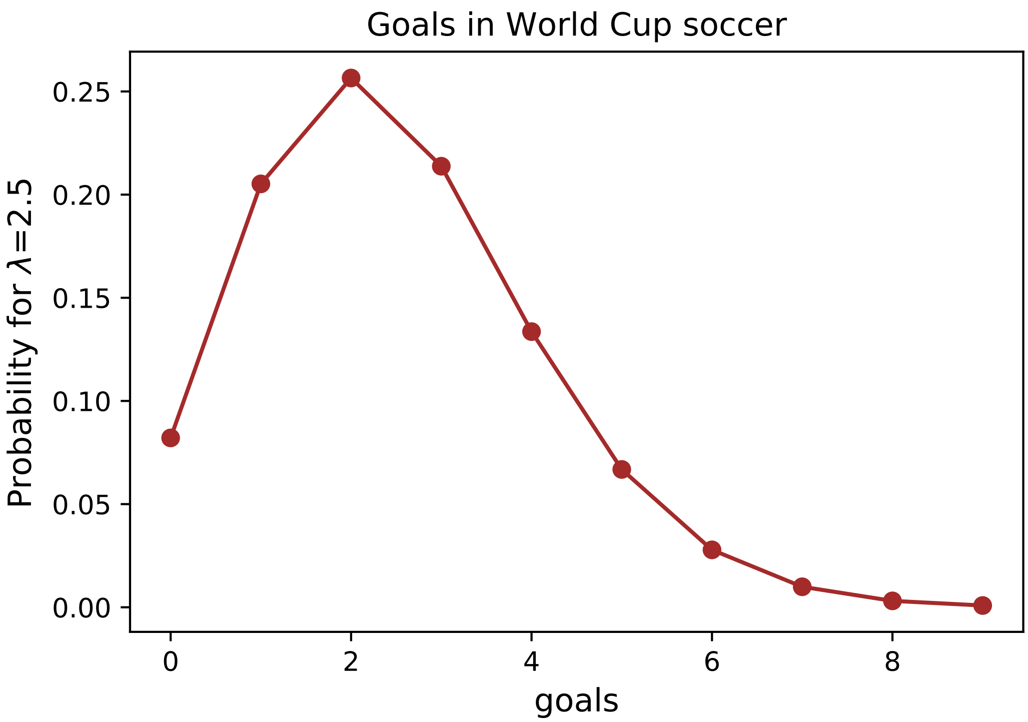
$$P(k=3) = \frac{\lambda^3 e^{-\lambda}}{3!} = \frac{2.5^3 e^{-\lambda}}{3!} = 0.213$$

Poisson (cont'd)

```python
import numpy as np
import matplotlib.pyplot as plt

# use lambda_ not keyword lambda
lambda_ = 2.5
n = np.arange(0, 10)
prob = stats.poisson.pmf(n, lambda_)
plt.plot(n, prob, '-o', color="brown")
plt.xlabel('Number of Events',
                fontsize=12)
plt.ylabel('Probability',
                fontsize=12)
plt.title("Poisson (lambda=3)")
plt.show()
```

Poisson (cont'd)



Normal (Gaussian)

- continous distribution
- most widely used
- mean μ, variance σ^2

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

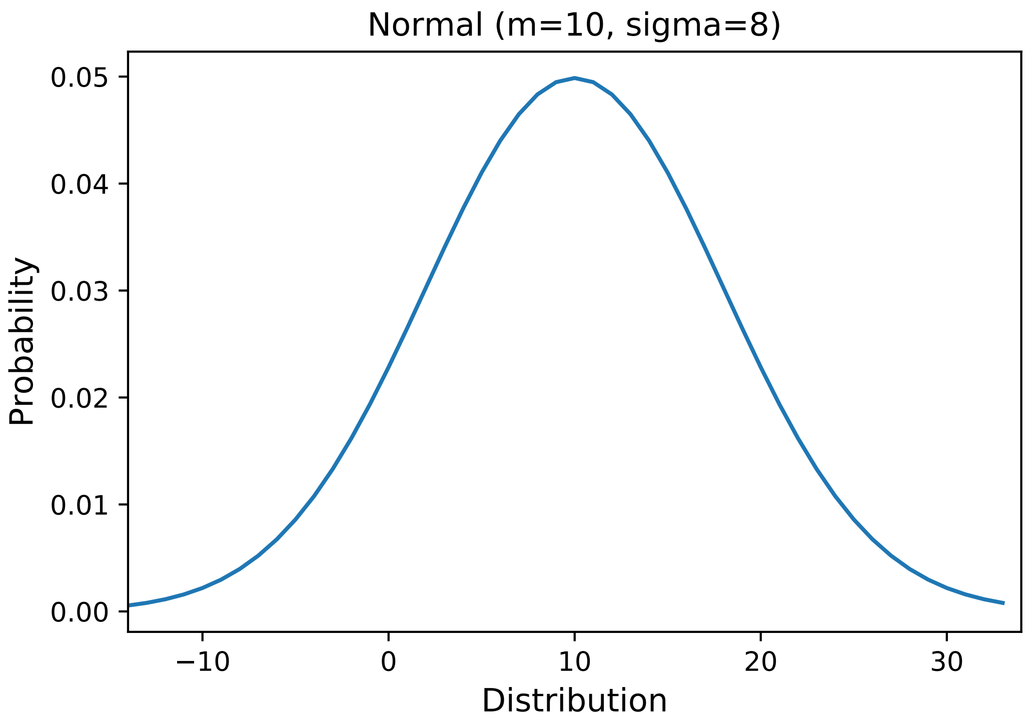
- symmetric

Normal (cont'd)

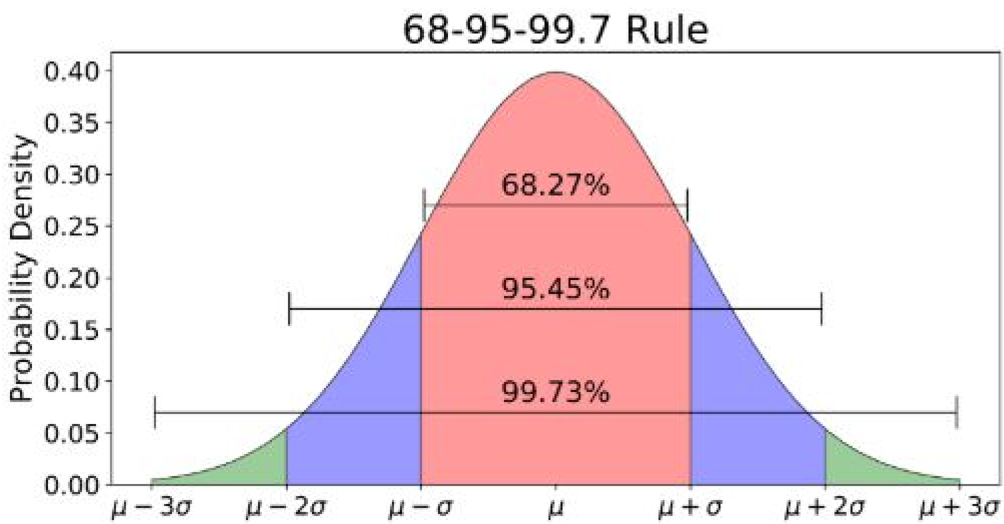
```
import numpy as np
import matplotlib.pyplot as plt

mean = 10
st_dev = 8
n = np.arange(mean - 3*st_dev,
                  mean + 3*st_dev)
normal = stats.norm.pdf(n, mean, st_dev)
plt.plot(n, normal)
plt.xlabel("Random variable X",
                   fontsize=12)
plt.ylabel("Probability",
                   fontsize=12)
plt.title("Normal (m=10, sigma=8)")
plt.xlim([mean - 3*st_dev,
                    mean + 3*st_dev])
plt.show()
```

Normal (cont'd)



68-95-99 Rule



- have explicit bounds
- much sharper than general non-parametric bounds

figure reprinted from www.kdnuggets.com with explicit permission of the editor

Concepts Check:

(a) discrete vs. continuous data
(b) probability distributions
(c) mean and standard deviation
(d) Bernouilli, uniform, binomial, Poisson, Normal
(e) outliers
(f) bounds