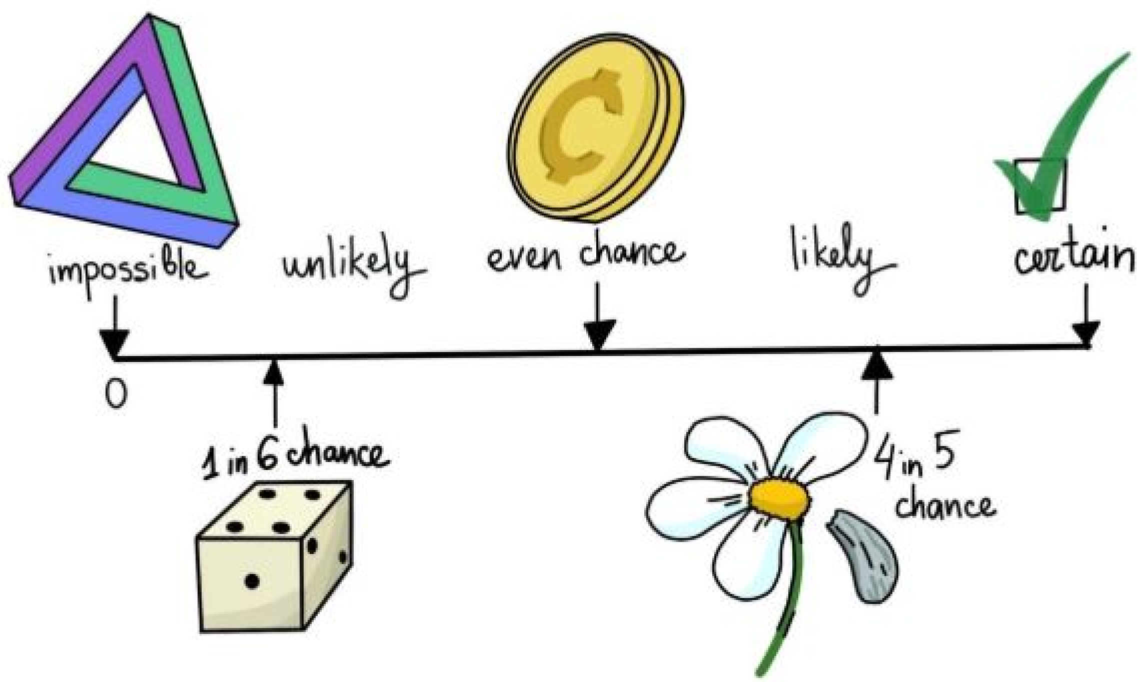
PROBABILITY

DISTRIBUTIONS

Why Use Probability?



- data features are stochastic
- results are statistical

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

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

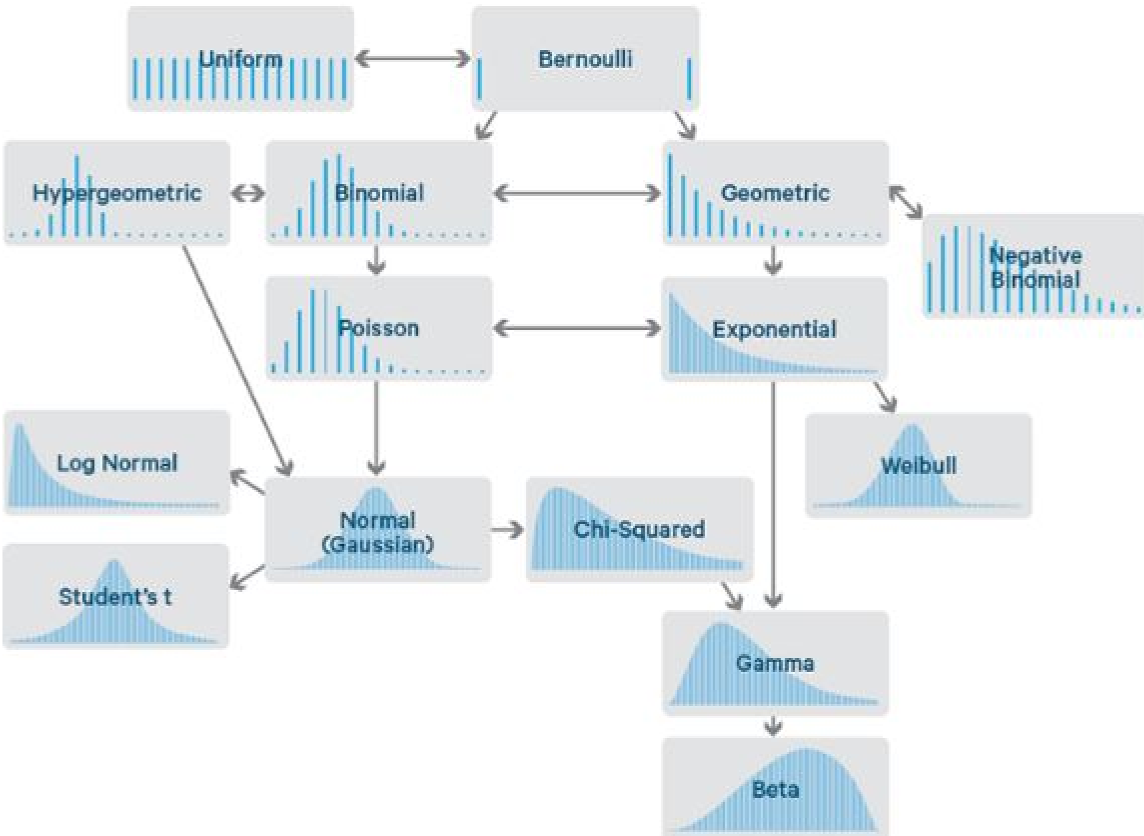
$$P\left(\mu - 2\sigma \leq X \leq \mu + 2\sigma\right) \leq 0.25$$

- 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

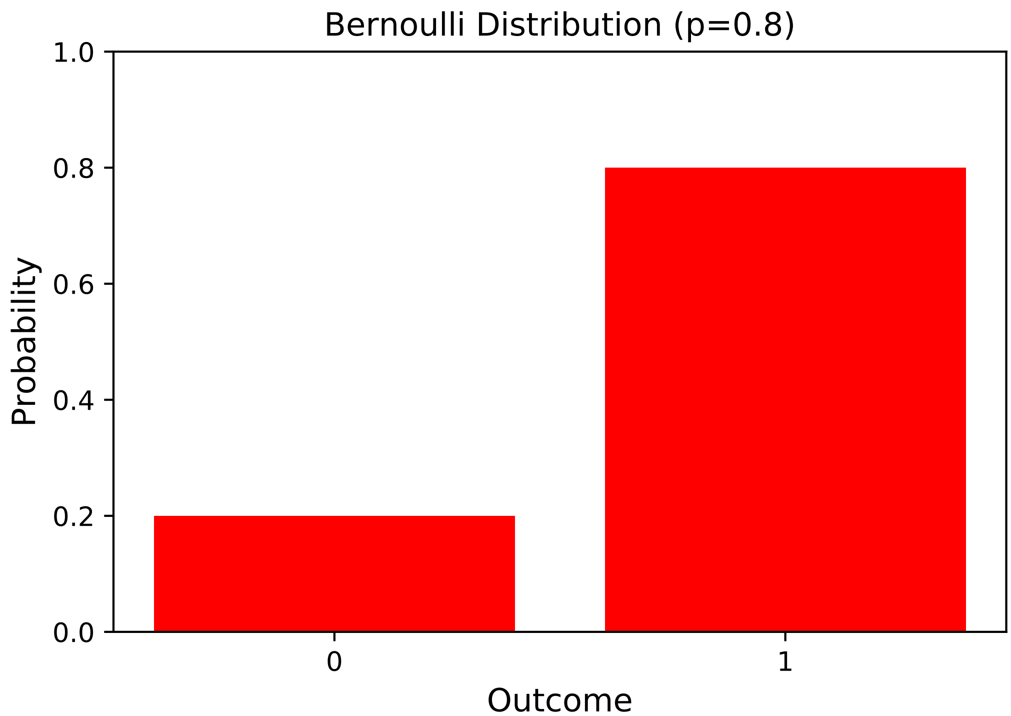
- discrete distribution
- value 1 with probability p
- value 0 with probability $q = 1 - p$
- result of a single experiment

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

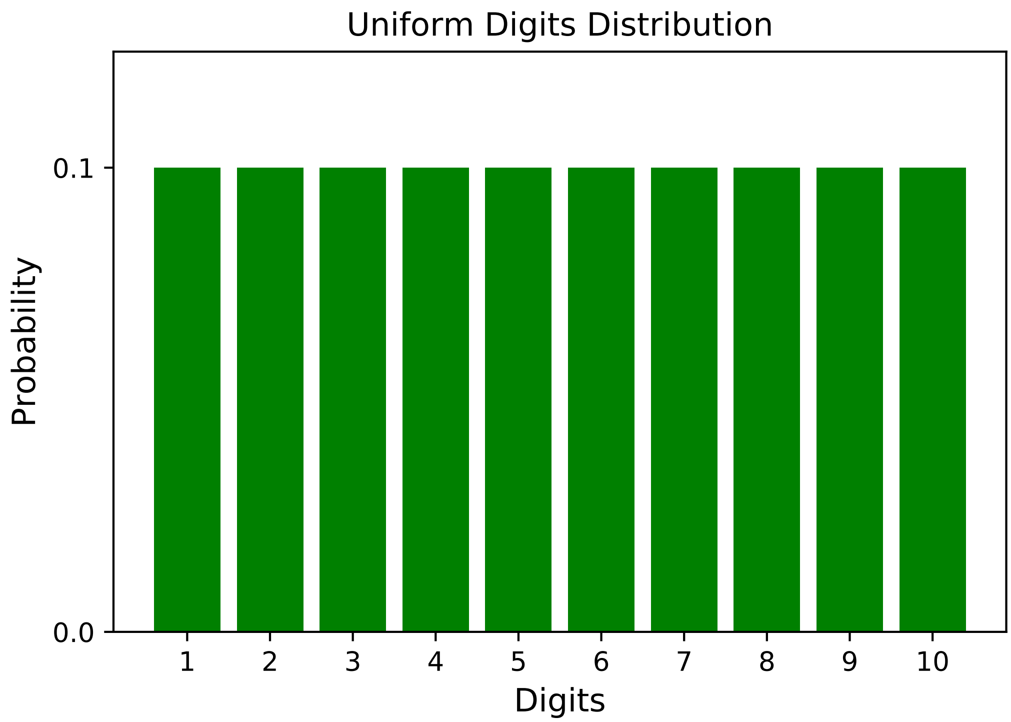
Uniform (cont'd)

```
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

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

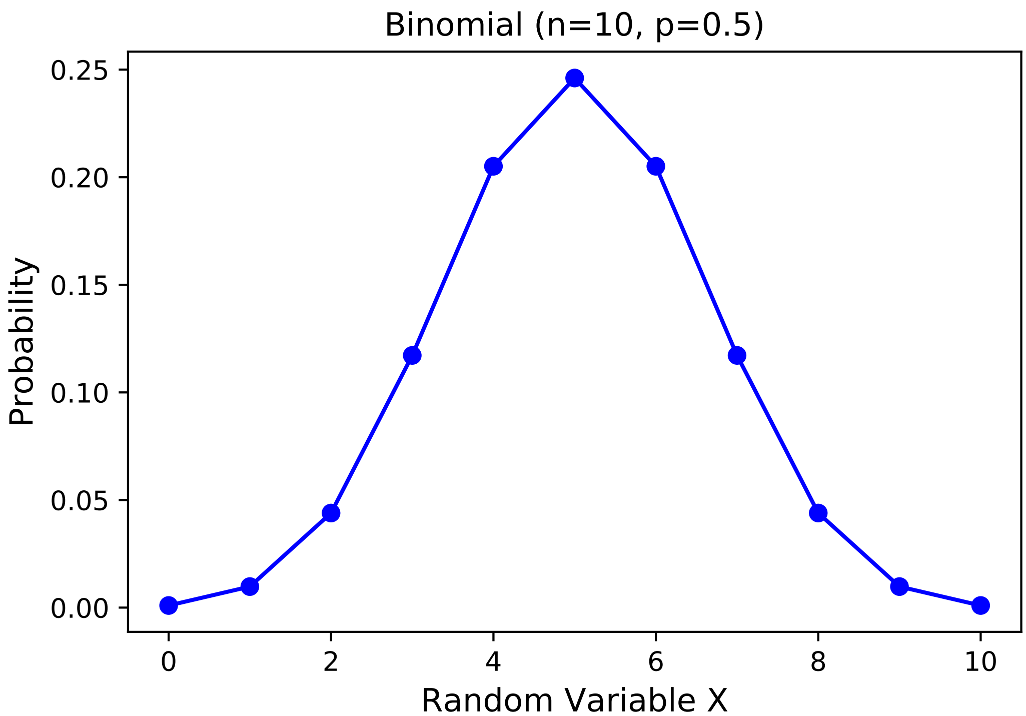
- $n = 1$ is the Bernouilli distribution

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=
plt.ylabel("Probability", fontsize=12)
plt.title("Binomial (n=10, p=0.5)")
plt.show()
```

Binomial (cont'd)



Poisson

- discrete distribution
- prob. of k events in time T:

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

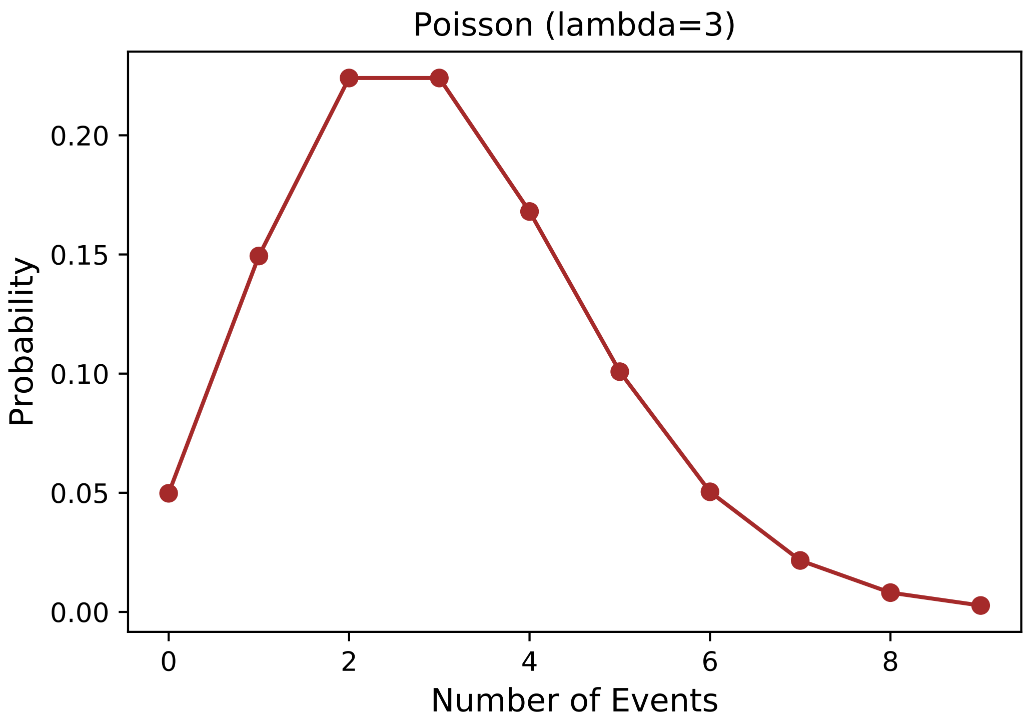
- mean is λ, variance is λ

Poisson (cont'd)

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

# use lambda_ not keyword lambda
lambda_ = 3
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

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

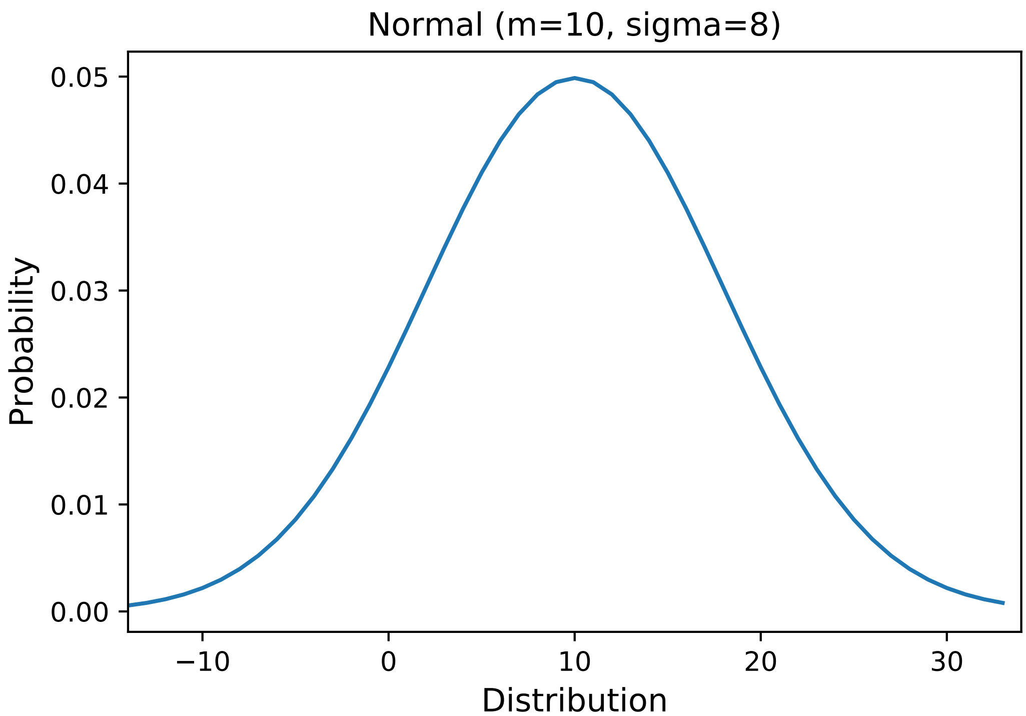
- symmetric

Normal (cont'd)

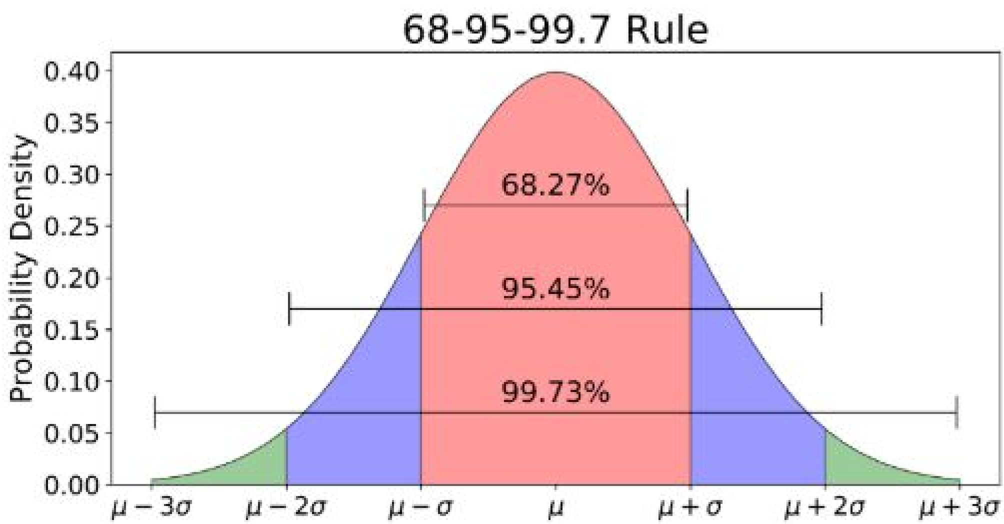
```
import numpy as np
import matplotlib.pyplot as plt

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) bounds