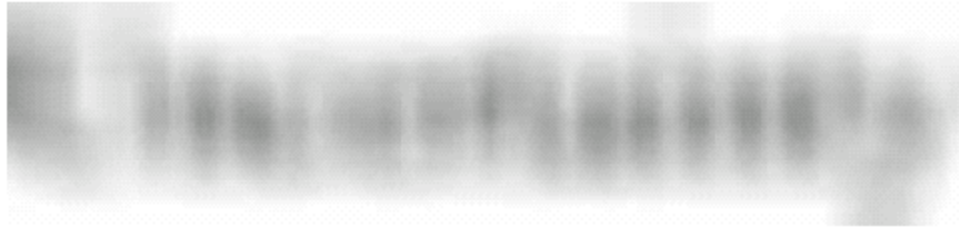
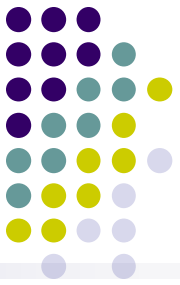


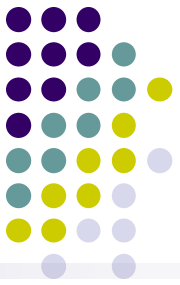
# What is this?



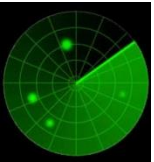
---



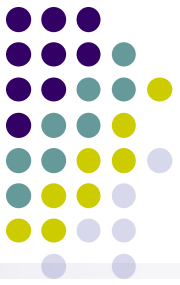
- Classical AI and ML research ignored this phenomena
- Another example
  - you want to catch a flight at 10:00am from Pitt to SF, can I make it if I leave at 8am and take a Marta at Gatech?
    - partial observability (road state, other drivers' plans, etc.)
    - noisy sensors (radio traffic reports)
    - uncertainty in action outcomes (flat tire, etc.)
    - immense complexity of modeling and predicting traffic

# Basic Probability Concepts



- A *sample space*  $\mathcal{S}$  is the set of all possible outcomes of a conceptual or physical, repeatable experiment. ( $\mathcal{S}$  can be finite or infinite.)
  - E.g.,  $\mathcal{S}$  may be the set of all possible outcomes of a dice roll:  $\mathcal{S} \equiv \{1, 2, 3, 4, 5, 6\}$ Two dice are shown, one slightly behind the other, with different faces visible.
  - E.g.,  $\mathcal{S}$  may be the set of all possible nucleotides of a DNA site:  $\mathcal{S} \equiv \{A, T, C, G\}$ A stylized DNA double helix with yellow and green strands and red, blue, and white base pairs.
  - E.g.,  $\mathcal{S}$  may be the set of all possible time-space positions of an aircraft on a radar screen:  $\mathcal{S} \equiv \{0, R_{\max}\} \times \{0, 360^\circ\} \times \{0, +\infty\}$ A green radar screen with a circular grid and several bright spots representing aircraft.
- An *event*  $\mathcal{A}$  is any subset of  $\mathcal{S}$ :
  - Seeing "1" or "6" in a dice roll; observing a "G" at a site; UA007 in space-time interval
- An *event space*  $\mathcal{E}$  is the possible worlds the outcomes can happen
  - All dice-rolls, reading a genome, monitoring the radar signal

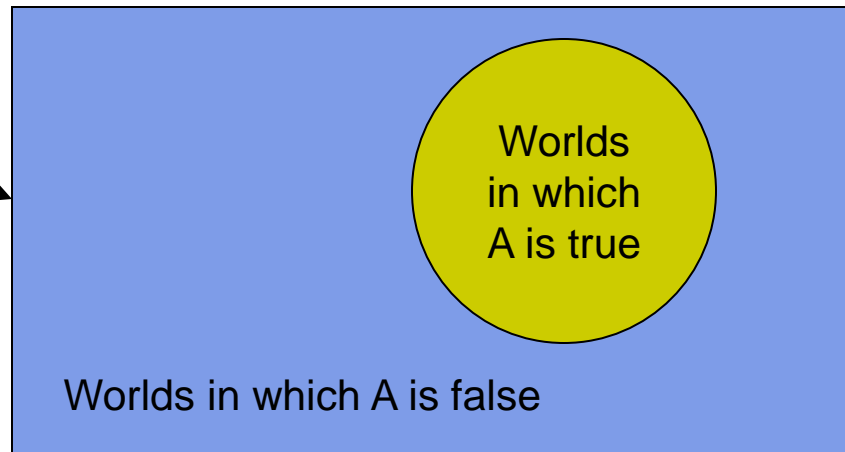
# Probability



- A *probability*  $P(A)$  is a function that maps an event  $A$  onto the interval  $[0, 1]$ .  $P(A)$  is also called the probability measure or probability mass of  $A$ .

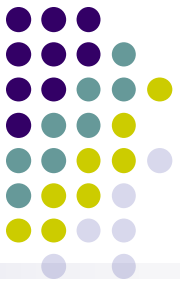
Sample space of all possible worlds.

Its area is 1

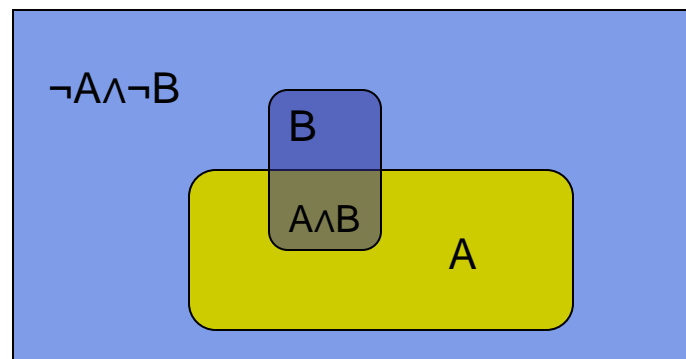


$P(a)$  is the area of the oval

# Kolmogorov Axioms

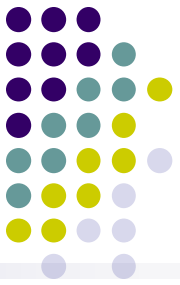


- All probabilities are between 0 and 1
  - $0 \leq P(A) \leq 1$
- $P(E) = 1$
- $P(\Phi) = 0$
- The probability of a disjunction is given by
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



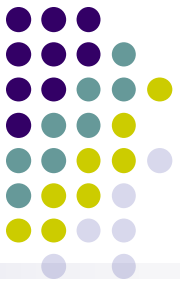
$A \vee B$  ?

# Why use probability?

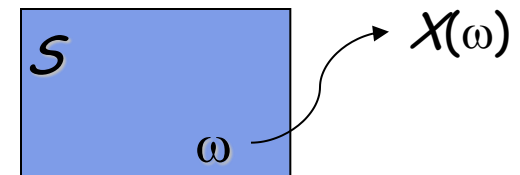


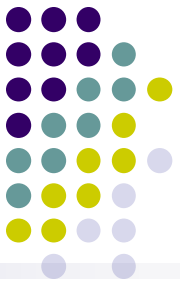
- There have been attempts to develop different methodologies for uncertainty:
  - Fuzzy logic
  - Qualitative reasoning (Qualitative physics)
  - ...
- “Probability theory is nothing but common sense reduced to calculation”
  - — Pierre Laplace, 1812.
- In 1931, de Finetti proved that it is irrational to have beliefs that violate these axioms, in the following sense:
  - If you bet in accordance with your beliefs, but your beliefs violate the axioms, then you can be guaranteed to lose money to an opponent whose beliefs more accurately reflect the true state of the world. (Here, “betting” and “money” are proxies for “decision making” and “utilities”.)
- What if you refuse to bet? This is like refusing to allow time to pass: every action (including inaction) is a bet

# Random Variable



- A *random variable* is a function that associates a unique numerical value (a token) with every outcome of an experiment. (The value of the r.v. will vary from trial to trial as the experiment is repeated)
- Discrete r.v.:
  - The outcome of a dice-roll
  - The outcome of reading a nt at site  $i$ :  $x_i$
- Binary event and indicator variable:
  - Seeing an "A" at a site  $\Rightarrow X=1$ , o/w  $X=0$ .
  - This describes the true or false outcome a *random event*.
  - Can we describe richer outcomes in the same way? (i.e.,  $X=1, 2, 3, 4$ , for being A, C, G, T)  
--- think about what would happen if we take expectation of  $X$ .
- Continuous r.v.:
  - The outcome of **recording** the **true** location of an aircraft:  $X_{true}$
  - The outcome of **observing** the **measured** location of an aircraft  $X_{obs}$

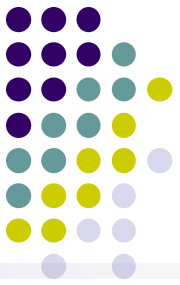




# Discrete Prob. Distribution

- A probability distribution  $P$  defined on a discrete sample space  $\mathcal{S}$  is an assignment of a non-negative real number  $P(s)$  to each sample  $s \in \mathcal{S}$  such that  $\sum_{s \in \mathcal{S}} P(s) = 1$ . ( $0 \leq P(s) \leq 1$ )
  - intuitively,  $P(s)$  corresponds to the *frequency* (or the likelihood) of getting a particular sample  $s$  in the experiments, if repeated multiple times.
  - call  $\theta_s = P(s)$  the *parameters* in a discrete probability distribution
- A discrete probability distribution is sometimes called a *probability model*, in particular if several different distributions are under consideration
  - write models as  $M_1, M_2$ , probabilities as  $P(X|M_1), P(X|M_2)$
  - e.g.,  $M_1$  may be the appropriate prob. dist. if  $X$  is from "fair dice",  $M_2$  is for the "loaded dice".
  - $M$  is usually a two-tuple of {dist. family, dist. parameters}

# Discrete Distributions



- Bernoulli distribution:  $\text{Ber}(p)$

$$P(x) = \begin{cases} 1-p & \text{for } x=0 \\ p & \text{for } x=1 \end{cases} \Rightarrow P(x) = p^x (1-p)^{1-x}$$



- Multinomial distribution:  $\text{Mult}(1, \theta)$

- Multinomial (indicator) variable:

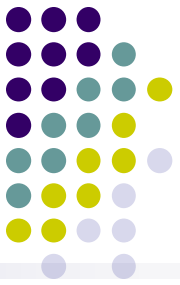
$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix}, \quad \text{where} \quad \begin{aligned} X_j &= [0,1], \quad \text{and} \quad \sum_{j \in \{1, \dots, 6\}} X_j = 1 \\ X_j &= 1 \text{ w.p. } \theta_j, \quad \sum_{j \in \{1, \dots, 6\}} \theta_j = 1 \end{aligned}$$



$$\begin{aligned} p(x(j)) &= P(\{X_j = 1, \text{ where } j \text{ index the dice-face}\}) \\ &= \theta_j = \prod_k \theta_k^{x_k} \end{aligned}$$



# Discrete Distributions



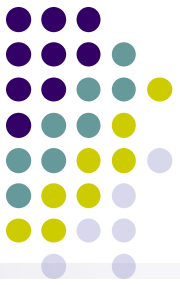
- Multinomial distribution:  $\text{Mult}(n, \theta)$
- Count variable:

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}, \quad \text{where } \sum_j x_j = n$$

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

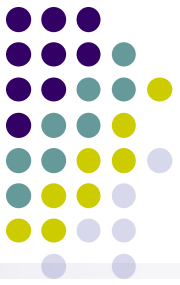
$$p(x) = \frac{n!}{x_1! x_2! \cdots x_k!} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_k^{x_k} = \frac{n!}{x_1! x_2! \cdots x_k!} \theta^x$$



# Continuous Prob. Distribution

- A **continuous random variable**  $X$  is defined on a continuous sample space: an interval on the real line, a region in a high dimensional space, etc.
  - $X$  usually corresponds to a real-valued measurements of some property, e.g., length, position, ...
  - It is meaningless to talk about the probability of the random variable assuming a particular value ---  $P(x) = 0$
  - Instead, we talk about the probability of the random variable assuming a value within a given interval, or half interval, or arbitrary Boolean combination of basic propositions.
    - $P(X \in [x_1, x_2])$
    - $P(X < x) = P(X \in [-\infty, x])$
    - $P(X \in [x_1, x_2] \cup [x_3, x_4])$

# Probability Density



- If the prob. of  $x$  falling into  $[x, x+dx]$  is given by  $p(x)dx$  for  $dx$ , then  $p(x)$  is called the **probability density** over  $x$ .
- If the probability  $P(x)$  is differentiable, then the probability density over  $x$  is the derivative of  $P(x)$ .
  - The probability of the random variable assuming a value within some given interval from  $x_1$  to  $x_2$  is equivalent to the area under the graph of the probability density function between  $x_1$  and  $x_2$ .

- Probability mass:  $P(X \in [x_1, x_2]) = \int_{x_1}^{x_2} p(x)dx$ ,

note that  $\int_{-\infty}^{+\infty} p(x)dx = 1$ .

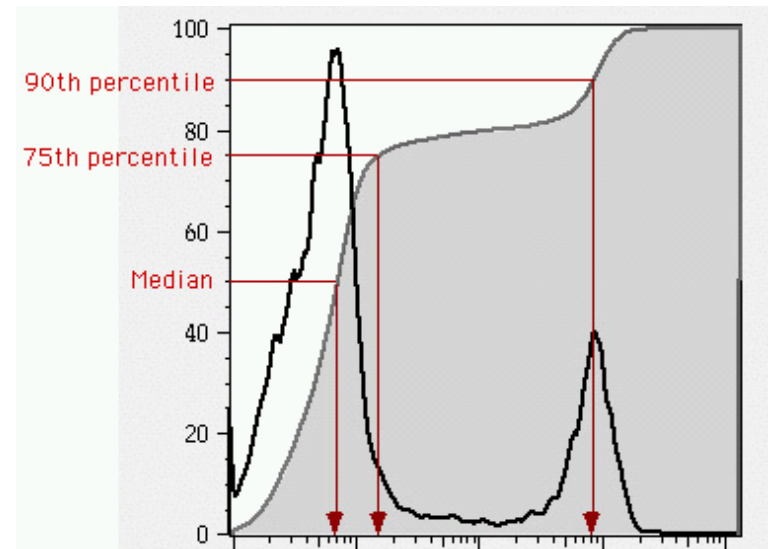
- Cumulative distribution function (CDF):

$$P(x) = P(X < x) = \int_{-\infty}^x p(x')dx'$$

- Probability density function (PDF):

$$p(x) = \frac{d}{dx} P(x)$$

$$\int_{-\infty}^{+\infty} p(x)dx = 1; \quad p(x) > 0, \forall x$$



Car flow on Liberty Bridge (cooked up!)

# The intuitive meaning of $p(x)$



- If

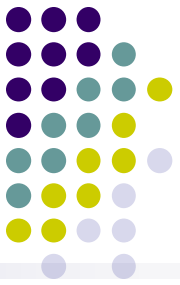
$$p(x_1) = a \text{ and } p(x_2) = b,$$

then when a value  $X$  is sampled from the distribution with density  $p(x)$ , you are  $a/b$  times as likely to find that  $X$  is “very close to”  $x_1$  than that  $X$  is “very close to”  $x_2$ .

- That is:

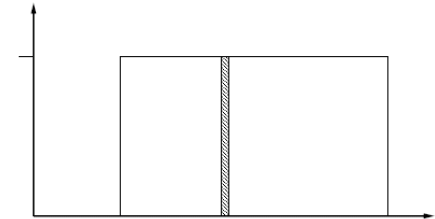
$$\lim_{h \rightarrow 0} \frac{P(x_1 - h < X < x_1 + h)}{P(x_2 - h < X < x_2 + h)} = \frac{\int_{x_1-h}^{x_1+h} p(x) dx}{\int_{x_2-h}^{x_2+h} p(x) dx} = \frac{p(x_1) \times 2h}{p(x_2) \times 2h} = a/b$$

# Continuous Distributions



- Uniform Density Function

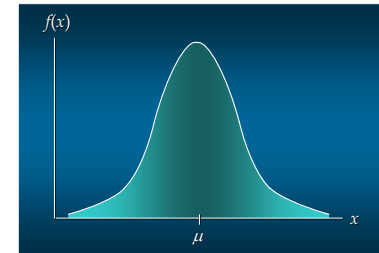
$$p(x) = 1/(b-a) \quad \text{for } a \leq x \leq b$$
$$= 0 \quad \text{elsewhere}$$



- Normal (Gaussian) Density Function

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

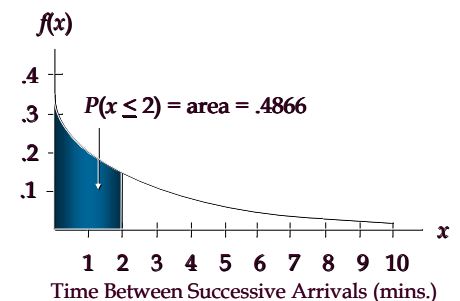
- The distribution is symmetric, and is often illustrated as a bell-shaped curve.
- Two parameters,  $\mu$  (mean) and  $\sigma$  (standard deviation), determine the location and shape of the distribution.
- The highest point on the normal curve is at the mean, which is also the median and mode.



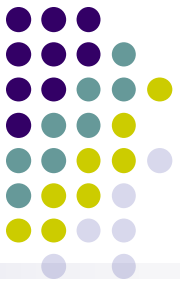
- Exponential Distribution

$$\text{PDF: } p(x) = \frac{1}{\mu} e^{-x/\mu},$$

$$\text{CDF: } P(x \leq x_0) = 1 - e^{-x_0/\mu}$$



# Gaussian (Normal) density in 1D



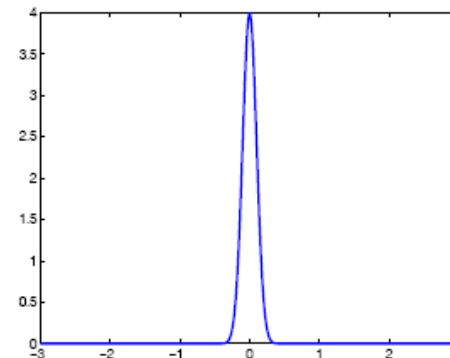
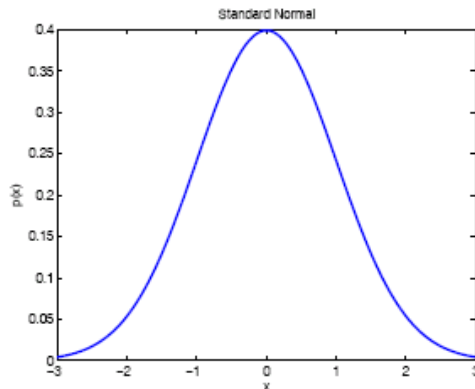
- If  $X \sim N(\mu, \sigma^2)$ , the probability density function (pdf) of  $X$  is defined as

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2 / 2\sigma^2}$$

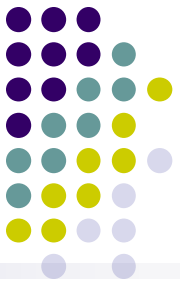
- We will often use the precision  $\lambda = 1/\sigma^2$  instead of the variance  $\sigma^2$ .
- Here is how we plot the pdf in matlab

```
xs=-3:0.01:3;
```

```
plot(xs,normpdf(xs,mu,sigma));
```



- Note that a density evaluated at a point can be larger than 1.

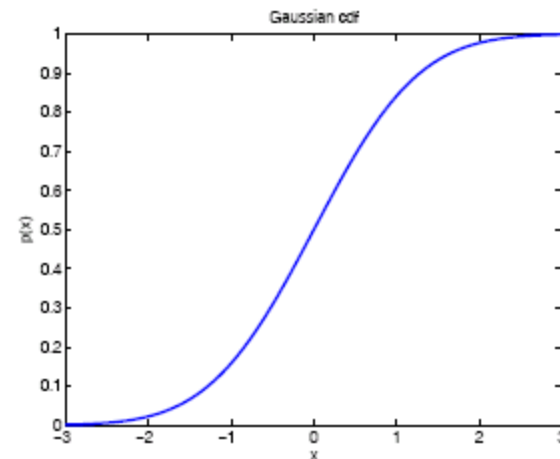
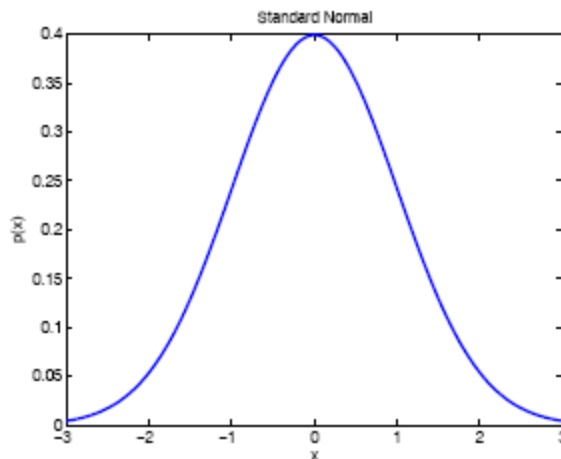


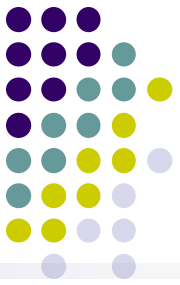
# Gaussian CDF

- If  $Z \sim N(0, 1)$ , the cumulative density function is defined as

$$\begin{aligned}\Phi(x) &= \int_{-\infty}^x p(z) dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-z^2/2} dz\end{aligned}$$

- This has no closed form expression, but is built in to most software packages (eg. `normcdf` in matlab stats toolbox).





# More on Gaussian Distribution

- If  $X \sim N(\mu, \sigma^2)$ , then  $Z = (X - \mu)/\sigma \sim N(0, 1)$ .
- How much mass is contained inside the  $[-2\sigma, 2\sigma]$  interval?

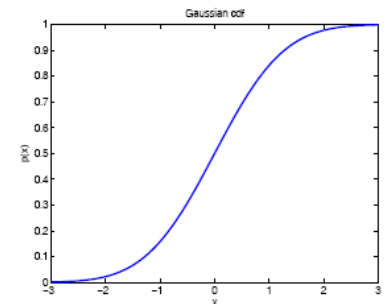
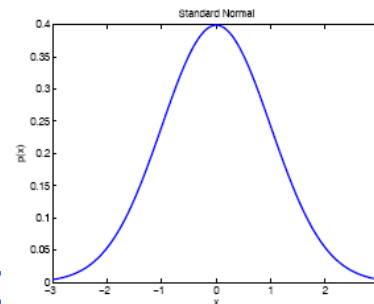
$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

- Since

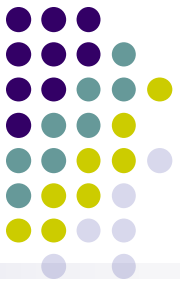
$$p(Z \leq -2) = \text{normcdf}(-2) = 0.025$$

we have

$$P(-2\sigma < X - \mu < 2\sigma) \approx 1 - 2 \times 0.025 =$$







# Statistical Characterizations

- **Expectation:** the centre of mass, mean value, first moment):

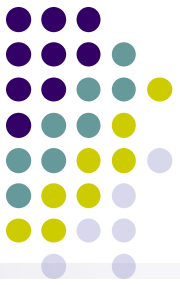
$$E(X) = \begin{cases} \sum_{i \in \mathcal{S}} x_i p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} x p(x) dx & \text{continuous} \end{cases}$$

- Sample mean:  $\mu = \frac{1}{N} \sum_{i=1}^N x_i$

- **Variance:** the spreadness:

$$Var(X) = \begin{cases} \sum_{x \in \mathcal{S}} [x - E(X)]^2 p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} [x - E(X)]^2 p(x) dx & \text{continuous} \end{cases}$$

- Sample variance  $\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2$



# Central limit theorem

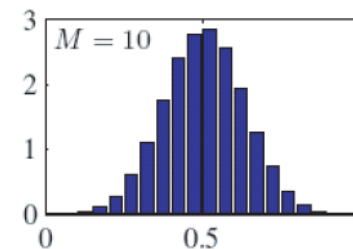
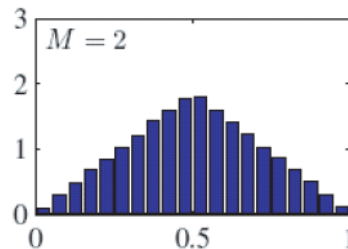
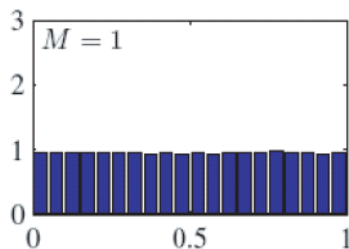
- If  $(X_1, X_2, \dots, X_n)$  are i.i.d. continuous random variables

- Define

$$\bar{X} = f(X_1, X_2, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i$$

- As  $n \rightarrow$  infinity,

$p(\bar{X}) \rightarrow$  Gaussian with mean  $E[X_i]$  and variance  $\text{Var}[X_i]$



- Somewhat of a justification for assuming Gaussian noise is common