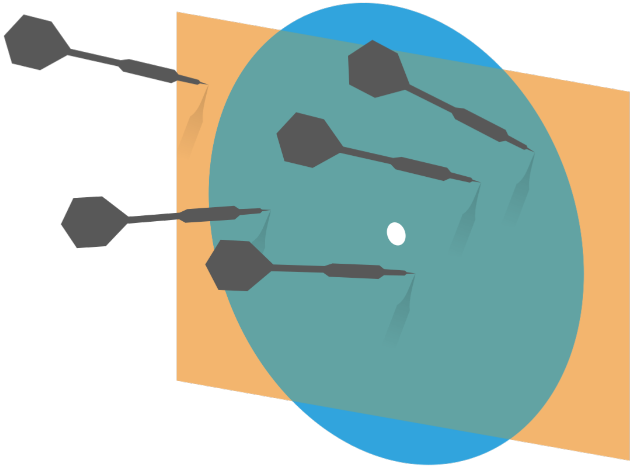
Probability and Statistics for Computer Science



Credit: wikipedia

Can we call e the exciting e ?

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$$

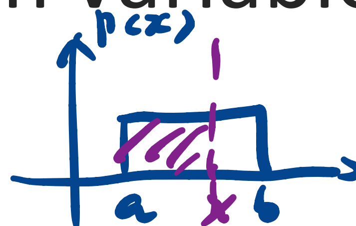
Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 3.4.2021

Last time

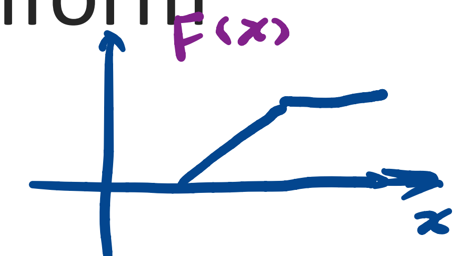
- **Poisson** distribution $P(X=k)=\frac{e^{-\lambda}\lambda^{k}}{k!}$ $k \geq 0$

- Continuous random variable; uniform distribution

pdf



cdf



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

- **Exponential** distribution

Objectives

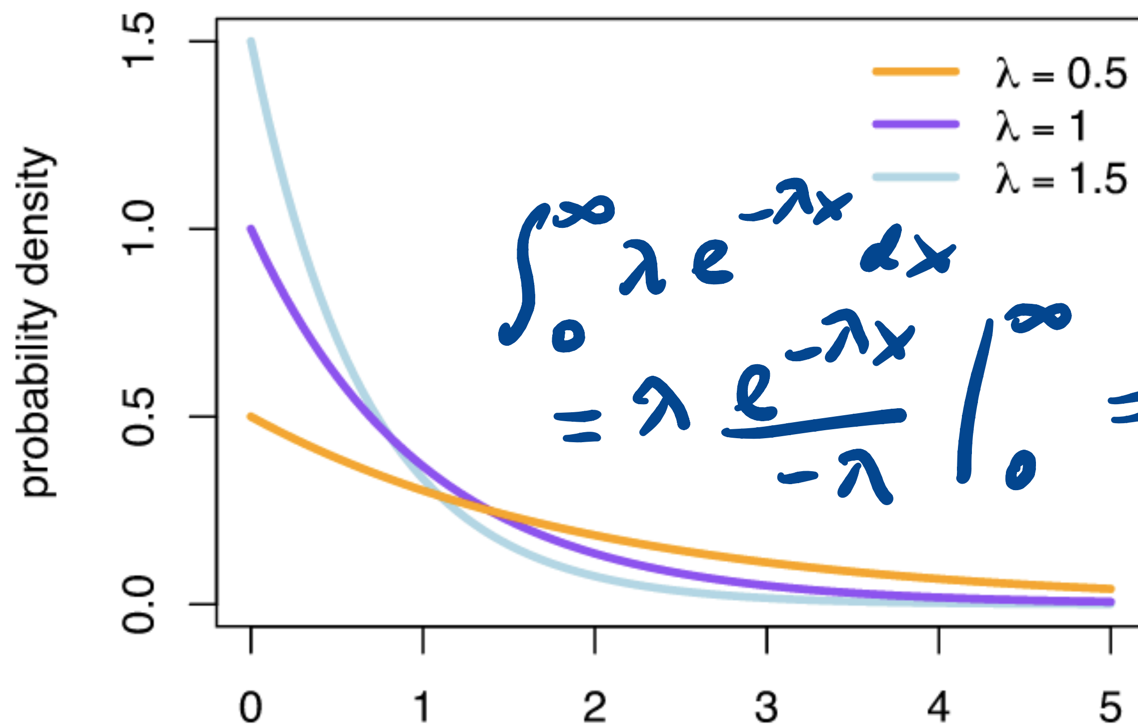
- Exponential distribution
- Normal (Gaussian) distribution

Exponential distribution

- Common Model for waiting time
- Associated with the Poisson distribution with the same **λ**

$$p(x) = \lambda e^{-\lambda x} \quad for\ x \geq 0$$

0 ow



$$\int_0^\infty \lambda e^{-\lambda x} dx$$

$$= \lambda \frac{e^{-\lambda x}}{-\lambda} \Big|_0^\infty = 1$$

Credit: wikipedia

Exponential distribution

- A continuous random variable X is exponential if it represent the "time" until next incident in a Poisson distribution with intensity **λ**. Proof See Degroot et al Pg 324.

$$p(x) = \lambda e^{-\lambda x} \quad for \ x \geq 0$$

- It's **similar** to **Geometric distribution** – the discrete version of waiting in queue

Memoryless Property

Expectations of Exponential distribution

- A continuous random variable X is exponential if it represent the "time" until next incident in a Poisson distribution with intensity **λ**.

$$E[X] = \int_0^\infty x \cdot \lambda e^{-\lambda x} dx$$

$$p(x) = \lambda e^{-\lambda x} \quad for \;\; x \geq 0$$

$$var[X] = E[X^2] - (E[X])^2$$
$$= E[X^2] - \left(\frac{1}{\lambda}\right)^2$$

$$\boxed{E[X] = \frac{1}{\lambda} \quad \& \quad var[X] = \frac{1}{\lambda^2}}$$

Example of exponential distribution

- How long will it take until the next call to be received by a call center? Suppose it's a random variable **T**. If the number of incoming call is a Poisson distribution with intensity **λ = 20 in an hour**. What is the expected time for T?

$$E[T] = \frac{1}{20} \text{ hr}$$

$$= \frac{1}{\lambda} \quad = 3 \text{ min}$$

Q:

- A store has a number of customers coming on Sat. that can be modeled as a Poisson distribution. In order to measure the average rate of customers in the day, the staff recorded the time between the arrival of customers, can he reach the same goal?

$\boxed{\text{A.}}$ Yes B. No

$E[T] = \frac{1}{\lambda}$

Normal (Gaussian) distribution

- The most famous continuous random variable distribution. The probability density is this:

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

$\hookrightarrow e$

Carl F. Gauss
(1777-1855)
Credit: wikipedia

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

Normal (Gaussian) distribution

- The most famous continuous random variable distribution. The probability density is this:

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

Carl F. Gauss
(1777-1855)
Credit: wikipedia

$$E[X] = \mu \quad \& \quad var[X] = \sigma^2$$

$$E[X] = \int_{-\infty}^{\infty} x\, p(x)\, dx = \mu$$

$$var[X] = \int_{-\infty}^{\infty} (X - E[X])^2 p(x)\, dx = \sigma^2$$

Normal (Gaussian) distribution

- The most famous continuous random variable distribution.

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

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

DeGroot
pg 304

$$E[X] = \mu \ \ \& \ \ var[X] = \sigma^2$$

Carl F. Gauss
(1777-1855)
Credit: wikipedia

Normal (Gaussian) distribution

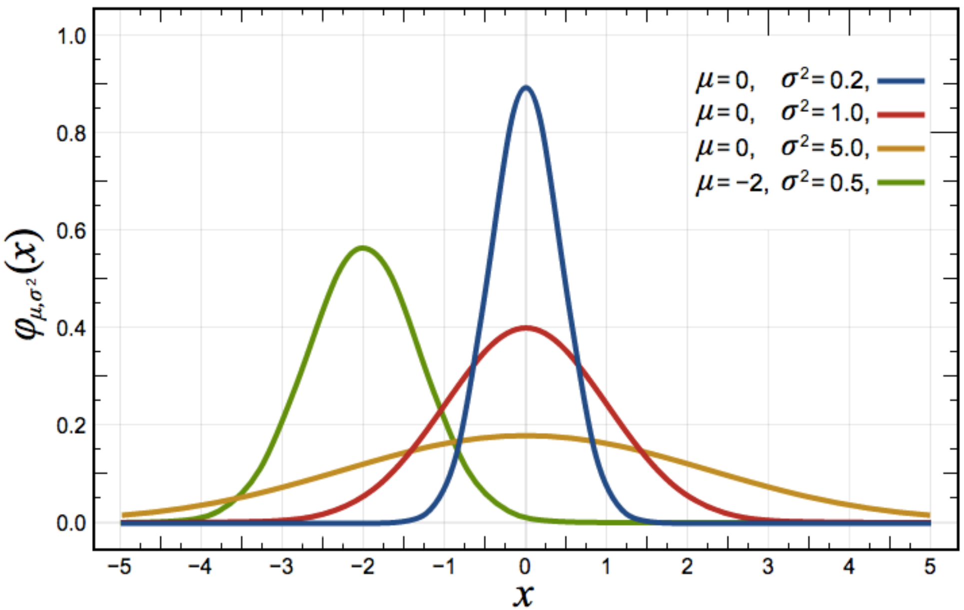
- A lot of data in nature are approximately normally distributed, ie. **Adult height**, etc.

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

$$E[X] = \mu \quad \& \quad var[X] = \sigma^2$$

Carl F. Gauss
(1777-1855)
Credit: wikipedia

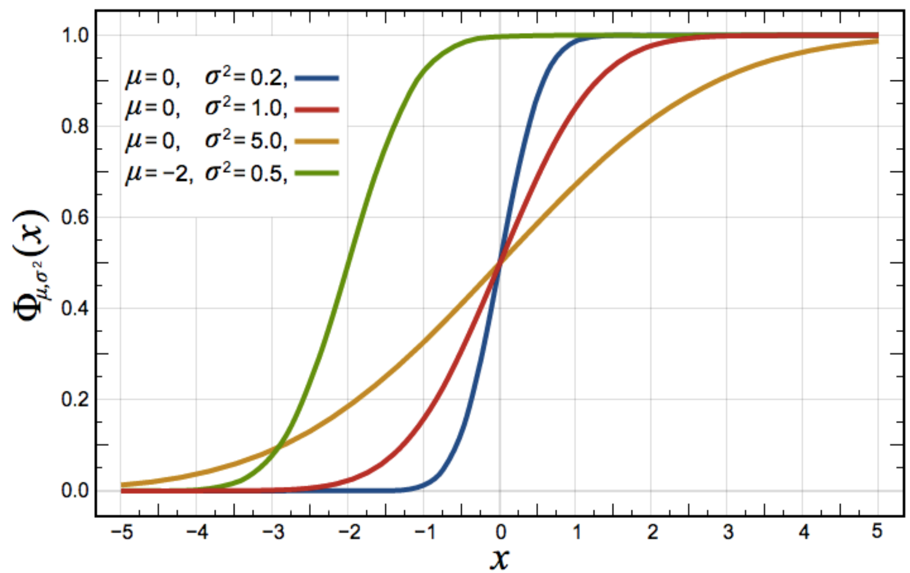
PDF and CDF of normal distribution curves



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

$F(x) = \int_{-\infty}^{x} p(x)\,dx$ CDF

$= P(X \leq x)$

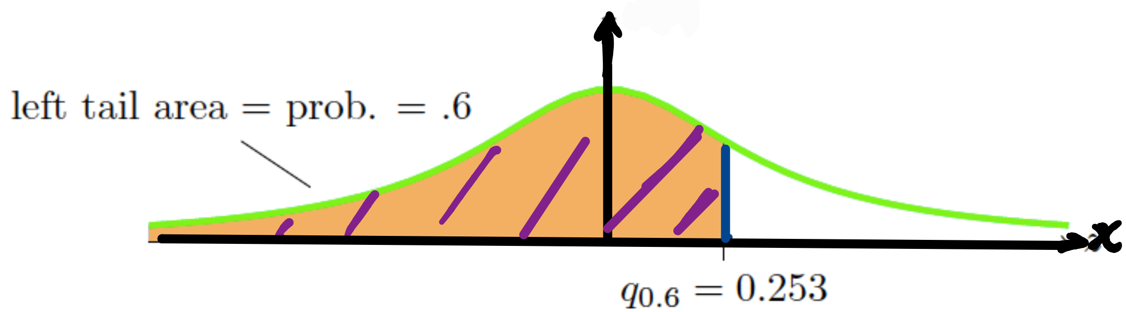


Credit: wikipedia

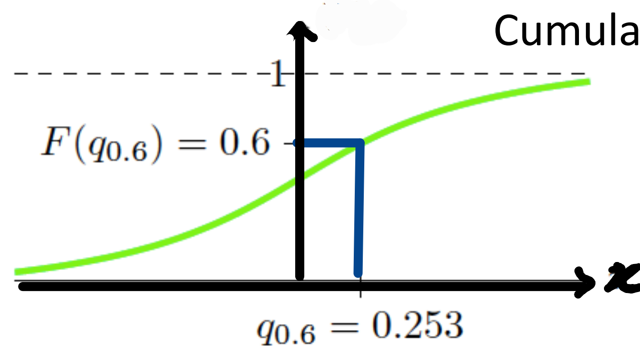
Quantile

- Quantiles give a measure of location, the median is the 0.5 quantile

Probability density



Cumulative Probability (CDF)



Credit:
J. Orloff et al

$q_{0.6}$: left tail area $= 0.6 \Leftrightarrow F(q_{0.6}) = 0.6$

Q.

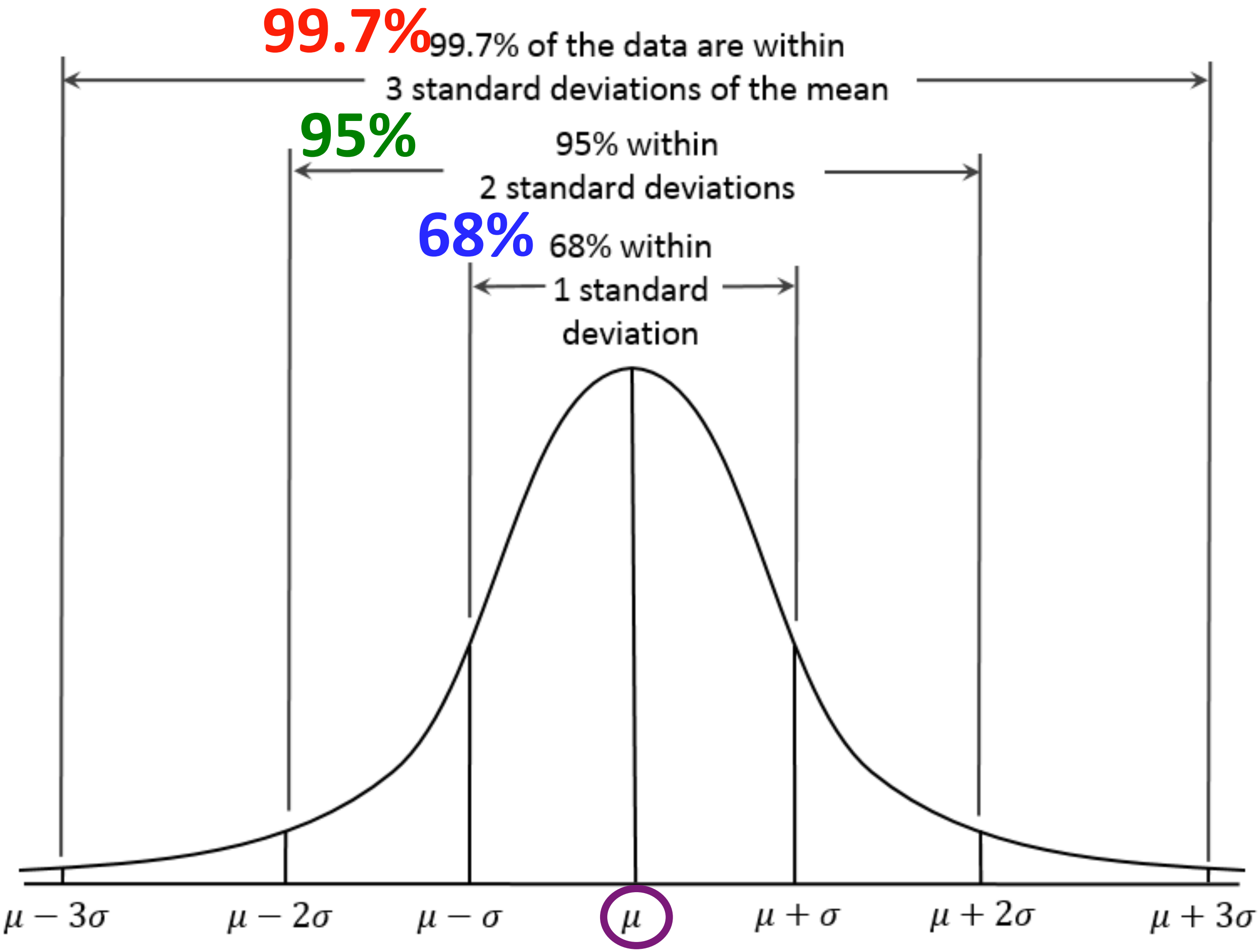
- What is the value of 50% quantile in a standard normal distribution?

A. -1

B. 0

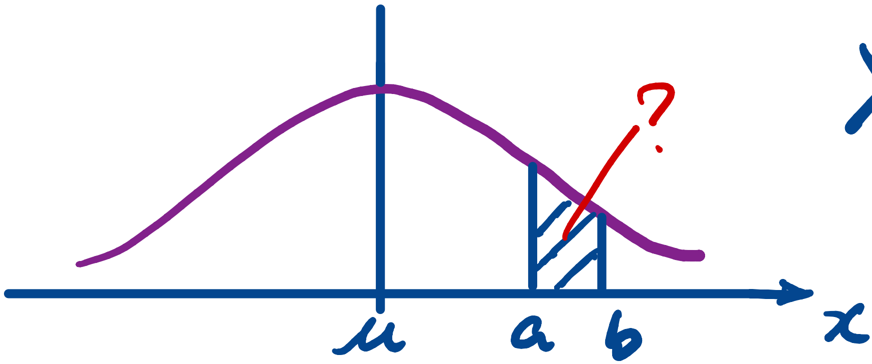
C. 1

Spread of normal (Gaussian) distributed data



Credit:
wikipedia

What is this probability?



$X \sim N(\mu, \sigma^2)$

$$P(a<X<b) = \int_a^b \frac{e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2}}{\sqrt{2\pi}\,\sigma}\,dx$$

No analytical solution!

Standard normal distribution

- If we standardize the normal distribution (by subtracting μ and dividing by σ), we get a random variable that has standard normal distribution.

$\hat{X} = \frac{x - \mu}{\sigma}$

- A continuous random variable X is **standard normal** if

$$p(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2})$$

Derivation of standard normal distribution

$$\hat{x} = \frac{x-\mu}{\sigma}$$

$$e^{(\frac{x-\mu}{\sigma})^2/2}$$

$$\int_{-\infty}^{+\infty} p(x)\ dx$$

$$= \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(x-\mu)^2}{2\sigma^2})\ dx$$

$$= \int_{-\infty}^{+\infty} \frac{1}{\cancel{\sigma}\sqrt{2\pi}} exp(-\frac{\hat{x}^2}{2})\cancel{\sigma}\ d\hat{x}$$

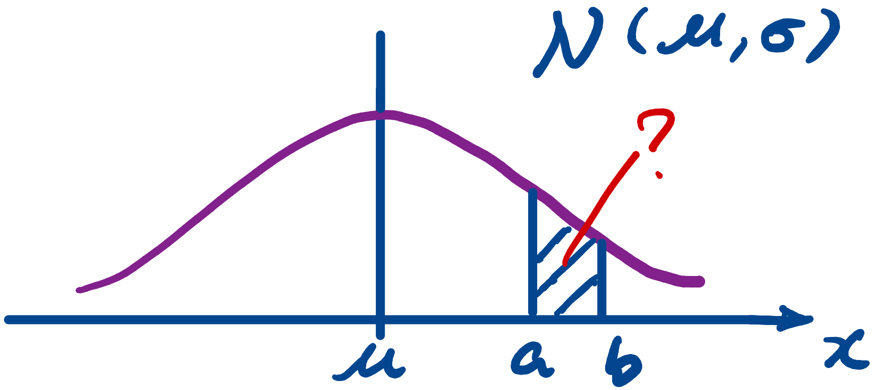
$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} exp(-\frac{\hat{x}^2}{2})\ d\hat{x}$$

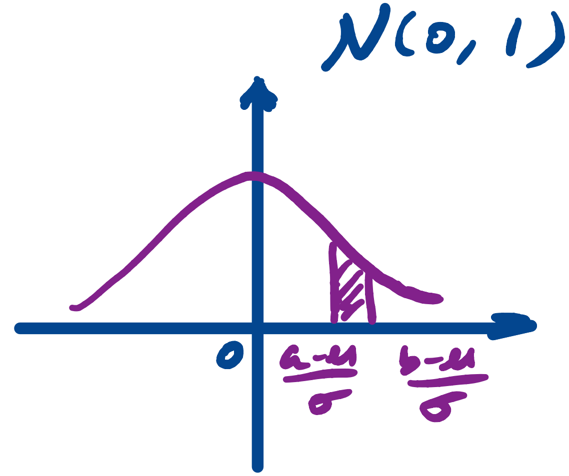
$$= \int_{-\infty}^{+\infty} p(\hat{x})\ dx$$

Call this standard and omit using a **hat**

$$\boxed{p(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2})}$$

What is this probability?





$$P(a < X < b) = \int_a^b \frac{e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2}}{\sqrt{2\pi}\,\sigma}\,dx = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{e^{-\hat{x}^2/2}}{\sqrt{2\pi}}\,d\hat{x}$$

Q. What is the mean of standard normal?

A. 0
B. 1

Q. What is the standard deviation of standard normal?

A. 0
B. 1

Standard normal distribution

- If we standardize the normal distribution (by subtracting μ and dividing by σ), we get a random variable that has standard normal distribution.

- A continuous random variable X is **standard normal** if

$$p(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2})$$

$$\boxed{E[X] = 0 \quad \& \quad var[X] = 1}$$

Another way to check the spread of normal distributed data

- Fraction of **normal** data within **1** standard deviation from the mean.

$$\int_{\frac{\mu-\sigma-\mu}{\sigma}}^{\frac{\mu+\sigma-\mu}{\sigma}} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$$

$$\frac{1}{\sqrt{2\pi}} \int_{-1}^{1} exp(-\frac{x^2}{2})dx \simeq 0.68$$

- Fraction of **normal** data within **k** standard deviations from the mean.

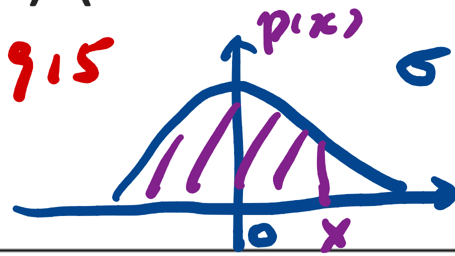
$$\frac{1}{\sqrt{2\pi}} \int_{-k}^{k} exp(-\frac{x^2}{2})dx$$

Using the standard normal's table to calculate for a normal distribution's probability

- If $X \sim \boldsymbol{N}(\mu=3, \sigma^2=16)$ (normal distribution)

$$P(X \leq 5) = ?$$

0.6915



$\sigma = 1$

$P\left(\frac{X-\mu}{\sigma} < \frac{5-\mu}{\sigma}\right)$

CDF table of $N(0,1)$

$\int_{-\infty}^{x} p(x)dx$

$P\left(\hat{X} \leq \frac{5-3}{4}\right)$

$P(\hat{X} \leq x_0)$

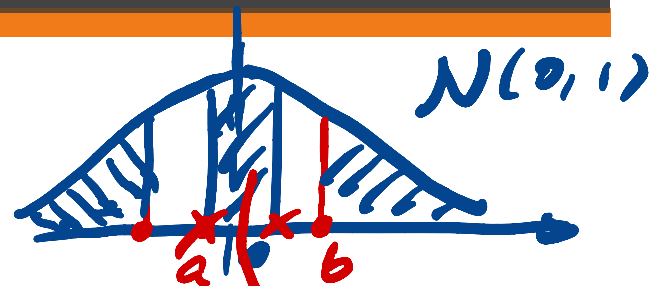
$x_0 = \frac{2}{4} = 0.5$

$x_0 = 0.51$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015

Q. Is the table with only positive x values enough?

A. Yes   B. No.



CDF

CDF(b)
-CDF(a)

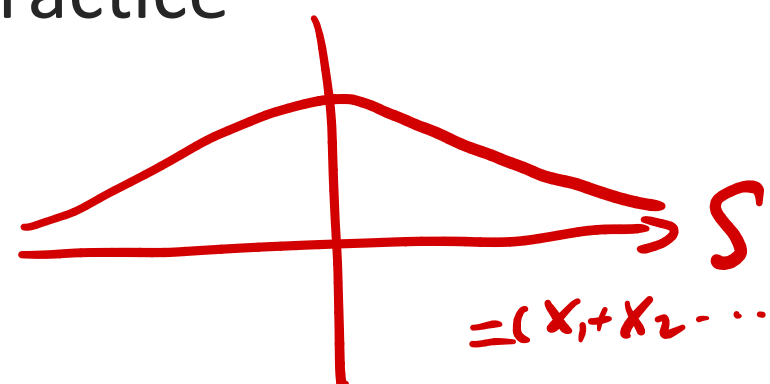
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

Central limit theorem (CLT)

- The distribution of the **sum** of ***N*** independent identical (IID) random variables tends toward a **normal** distribution as ***N*** $\longrightarrow \infty$

- Even when the component random variables are not exactly IID, the result is approximately true and very useful in practice

$\bar{X} = \frac{X_1}{N} + \frac{X_2}{N} + \cdots$



$= (X_1 + X_2 \cdots X_N)$

Central limit theorem (CLT)

- CLT helps explain the prevalence of normal distributions in nature
- A binomial random variable tends toward a normal distribution when ***N*** is large due to the fact it is the sum of IID Bernoulli random variables

$$X_{Bino} = \sum_{i}^{N} X_i$$

$$X_i = \begin{cases} 1 \\ 0 \end{cases}$$

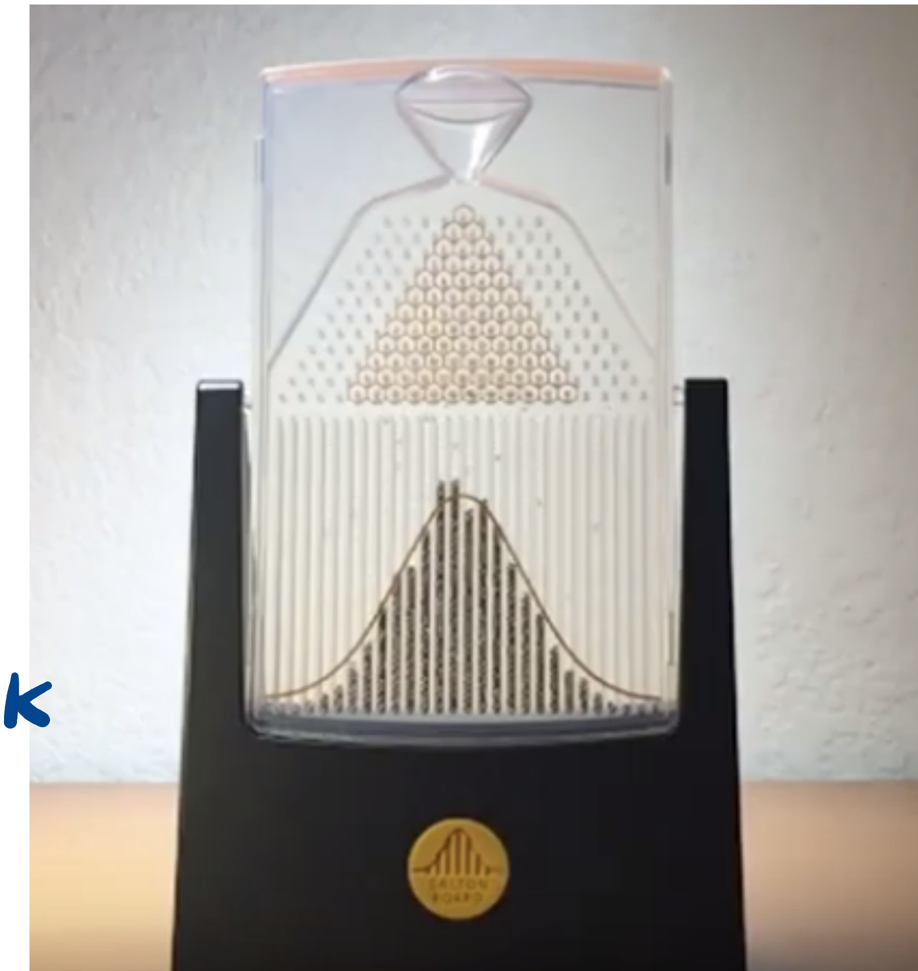
$$P(X_i) = \begin{cases} p & x_i = 1 \\ 1-p & x_i = 0 \end{cases}$$

The Binomial distributed beads of the Galton Board

The Binomial distribution looks very similar to Normal when N is large

$$P(X_{Bino}=K) = \binom{N}{K} p^{K}(1-p)^{N-K}$$

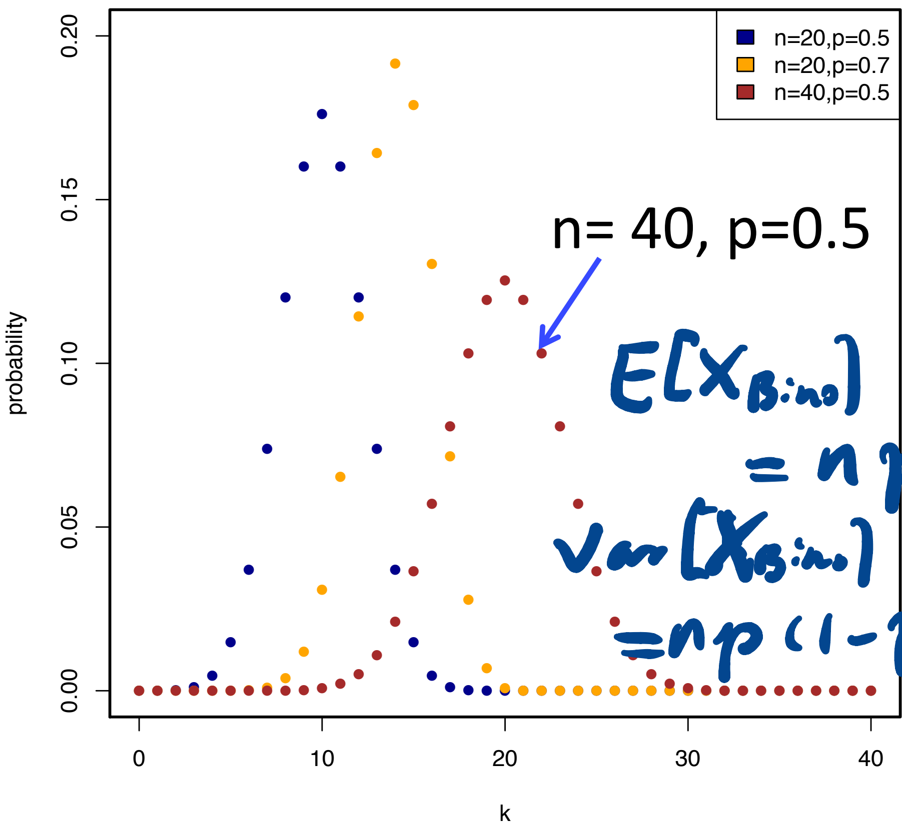
Σ



Binomial approximation with Normal

$p \neq 0.5$

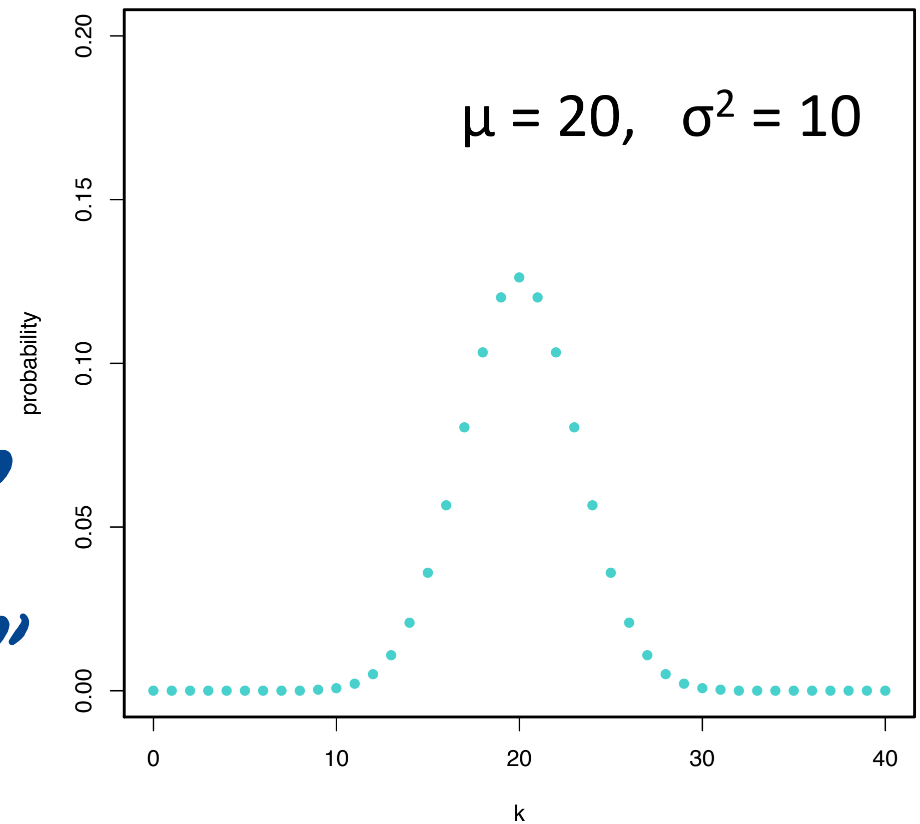
Binomial distribution



$E[X_{Bino}] = np$

$Var[X_{Bino}] = np(1-p)$

Approximation with Normal



Binomial approximation with Normal

- Let k be the number of heads appeared in 40 tosses of fair coin

- The goal is to estimate the following with normal

$$P(10 \le k \le 25) = \sum_{k=10}^{25} \binom{40}{k} 0.5^k 0.5^{40-k}$$

$$= \sum_{k=10}^{25} \binom{40}{k} 0.5^{40} \simeq 0.96$$

$p(x) = CDF'(x)$

$$E[k] = np = 40 \cdot 0.5 = 20$$

$$std[k] = \sqrt{np(1-p)}$$
$$= \sqrt{40 \cdot 0.5 \cdot 0.5} = \sqrt{10}$$

Binomial approximation with Normal

- Use the same mean and standard deviation of the original binomial distribution.

$$\mu = 20 \qquad \sigma = \sqrt{10} \simeq 3.16$$

- Then standardize the normal to do the calculation

$$P(10 \leq k \leq 25) \simeq \frac{1}{\sigma\sqrt{2\pi}} \int_{10}^{25} exp(-\frac{(x-\mu)^2}{2\sigma^2})dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\frac{10-20}{3.16}}^{\frac{25-20}{3.16}} exp(-\frac{x^2}{2})dx$$

$$\simeq 0.94$$

use the table

Assignments this week

- Week 6 quiz
- HW5
- Prepare for Midterm1:
  - Practice exams
  - Read through instructions on Compass

Additional References

- Charles M. Grinstead and J. Laurie Snell "Introduction to Probability”
- Morris H. Degroot and Mark J. Schervish "Probability and Statistics”

See you next time

See You!