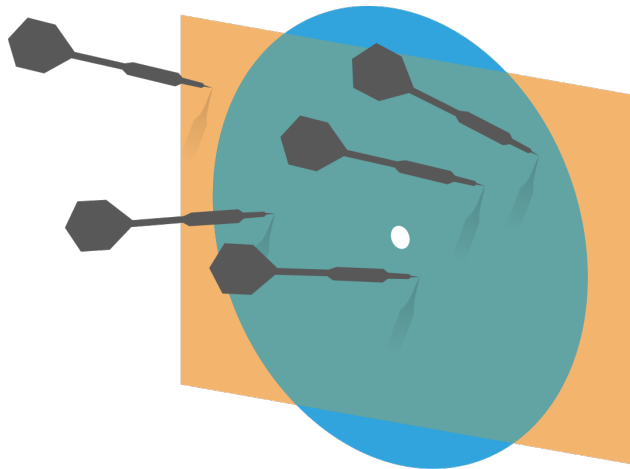# Probability and Statistics for Computer Science

↗

Can we call e the exciting e ?

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

Credit: wikipedia

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 10.1.2020

Midterm 1   Oct. 8. Thurs.
(CBTF)   12:30 pm

Conflicts & alternative

Oct 8, 8pm → CBTF

Oct. 9, 8am → CBTF

Oct 8, 4pm

Check out instructions, practices.

Last time

Poisson Distribution

Continuous Random Variable

Probability Density Function

Objectives

Cumulative Distribution Function
for continuous random variable

Normal (Gaussian) Distribution

Exponential Distribution

$$pdf \ p(x)$$

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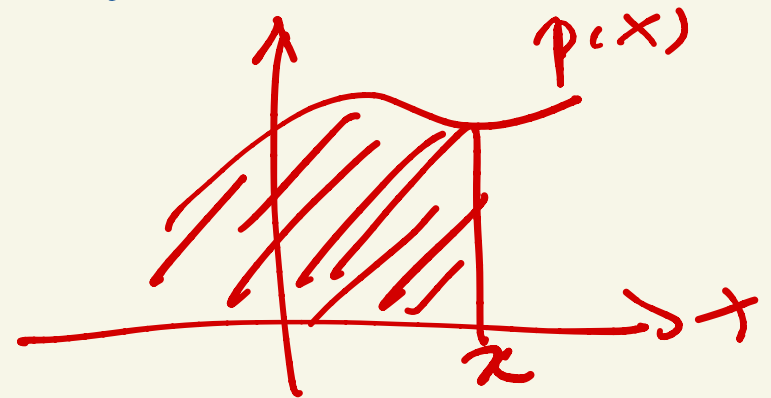
Cumulative Distribution Function

Def.: $P(X \leq x)$

Discrete R.V. $\quad P(X \leq x) = \sum_{-\infty}^{x} P(X = x)$

$$\sum \longrightarrow \int$$

Continuous RV

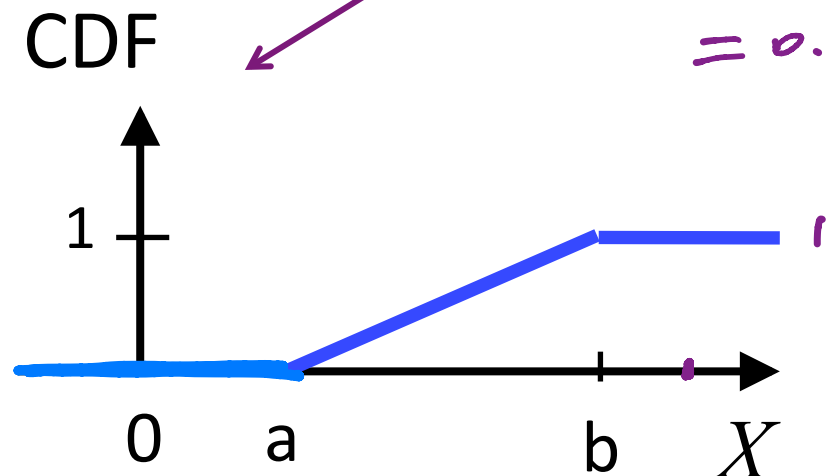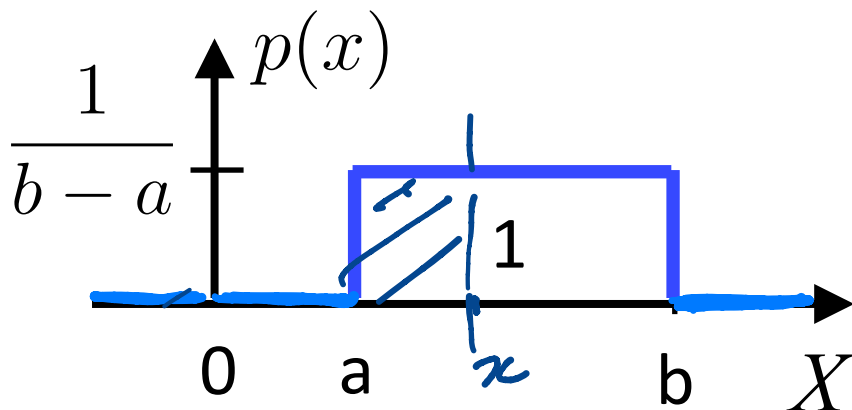$$P(X \leq x) = \int_{-\infty}^{x} p(x) \, dx$$

Cumulative distribution of continuous uniform distribution

✳ Cumulative distribution function (CDF)

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

of a uniform random variable X is:

CDF $\left(x = \frac{a+b}{2}\right)$ = 0.5

❋ ~~What is the constant **c** given the spin angle θ has the following pdf?~~ *What is the CDF ?*

$CDF(z)$

$p(\theta)$

$\frac{1}{\pi}$ **c**

A. 1

B. 1/π

C. 2/π

D. 4/π

E. 1/2π

0   π   2π   θ

Quantile

CDF (x)

pdf (x)
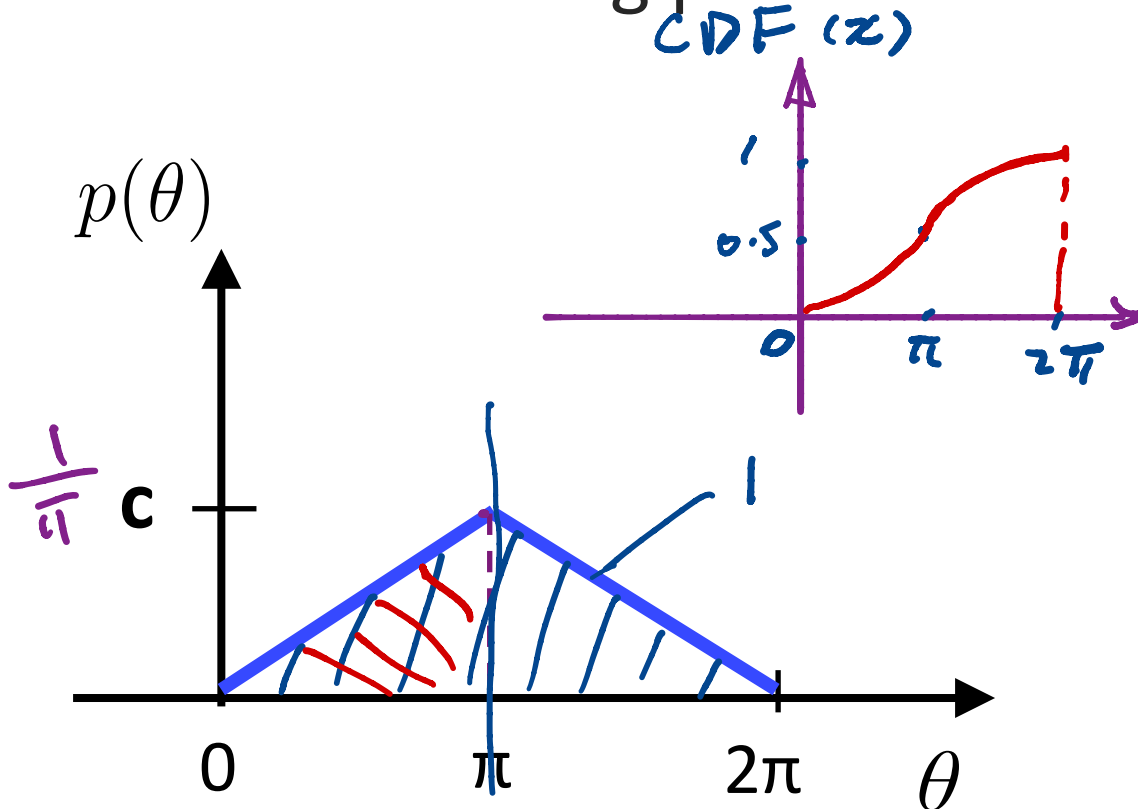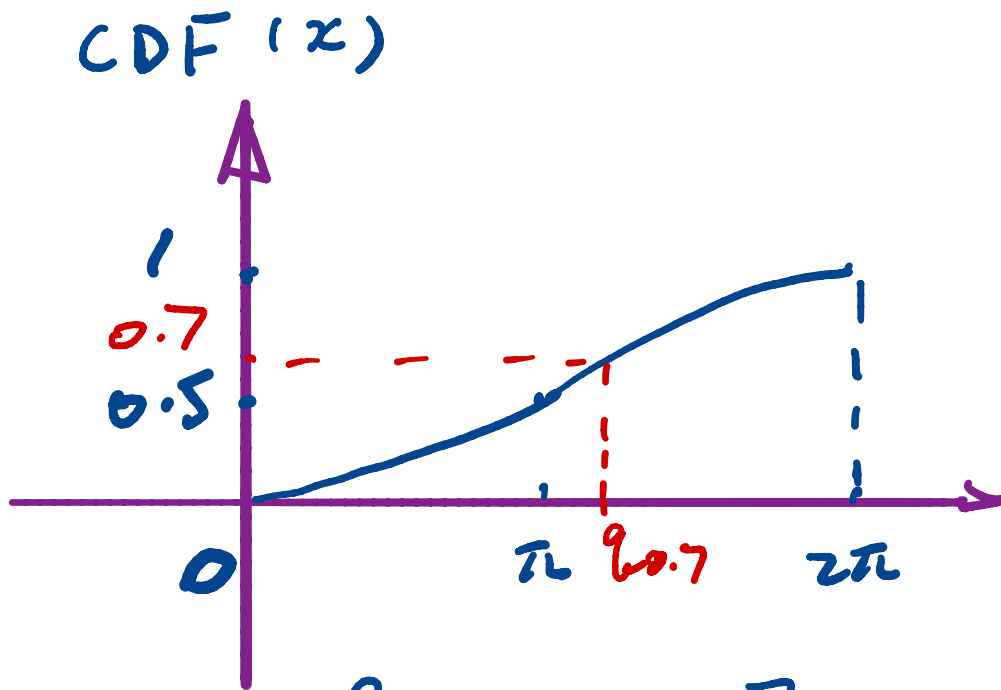
$q_{0.7} \neq 0.7$

Normal (Gaussian) distribution

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

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

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

$$(exp) \rightarrow \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

$$\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

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

Normal (Gaussian) distribution

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

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

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

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
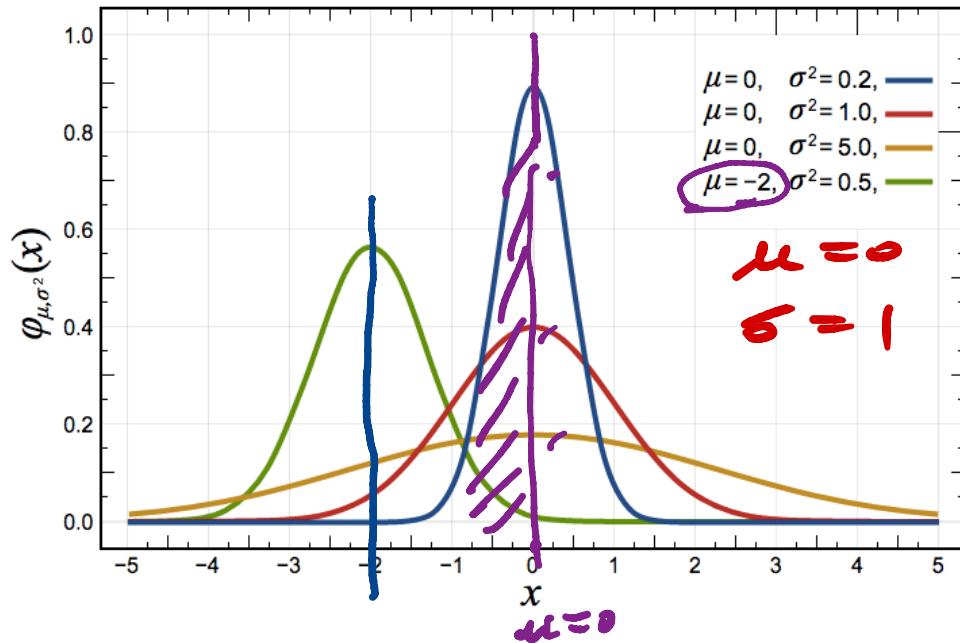
Probability density func. (PDF)

Cumulative distribution func. (CDF)

$\mu = 0$
$\sigma = 1$

$\mu = 0$
$\sigma = 1$

$\mu = 0$

Credit: wikipedia
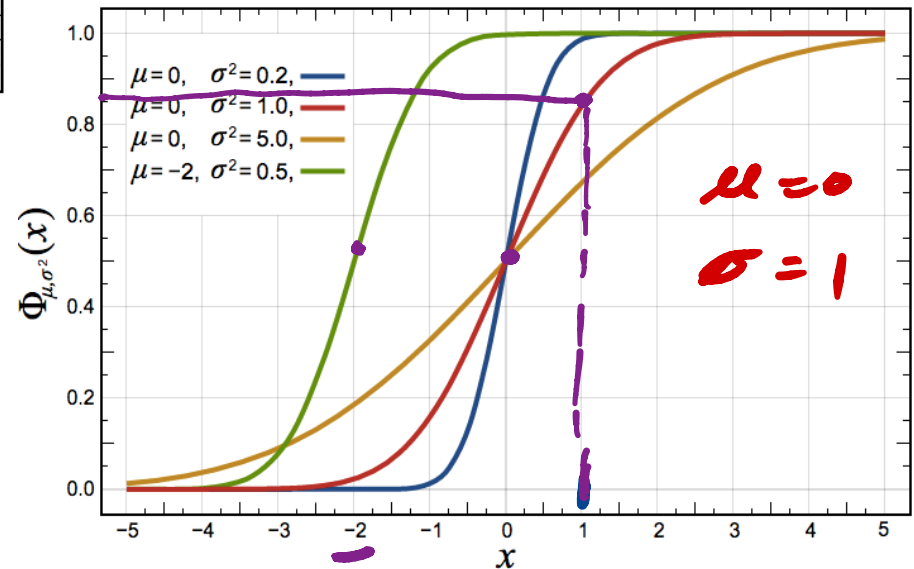
Quantile

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

$$q_{0.6} = 0.253$$

Credit:
J. Orloff et al

$\phi(z) \to$ pdf

Probability density

$\phi(z)$

left tail area = prob. = .6

$q_{0.6} = 0.253$ $z = 0.253$

$\Phi \to$ CDF

$\Phi(z)$

Cumulative Probability (CDF)

$F(q_{0.6}) = 0.6$

$q_{0.6} = 0.253$
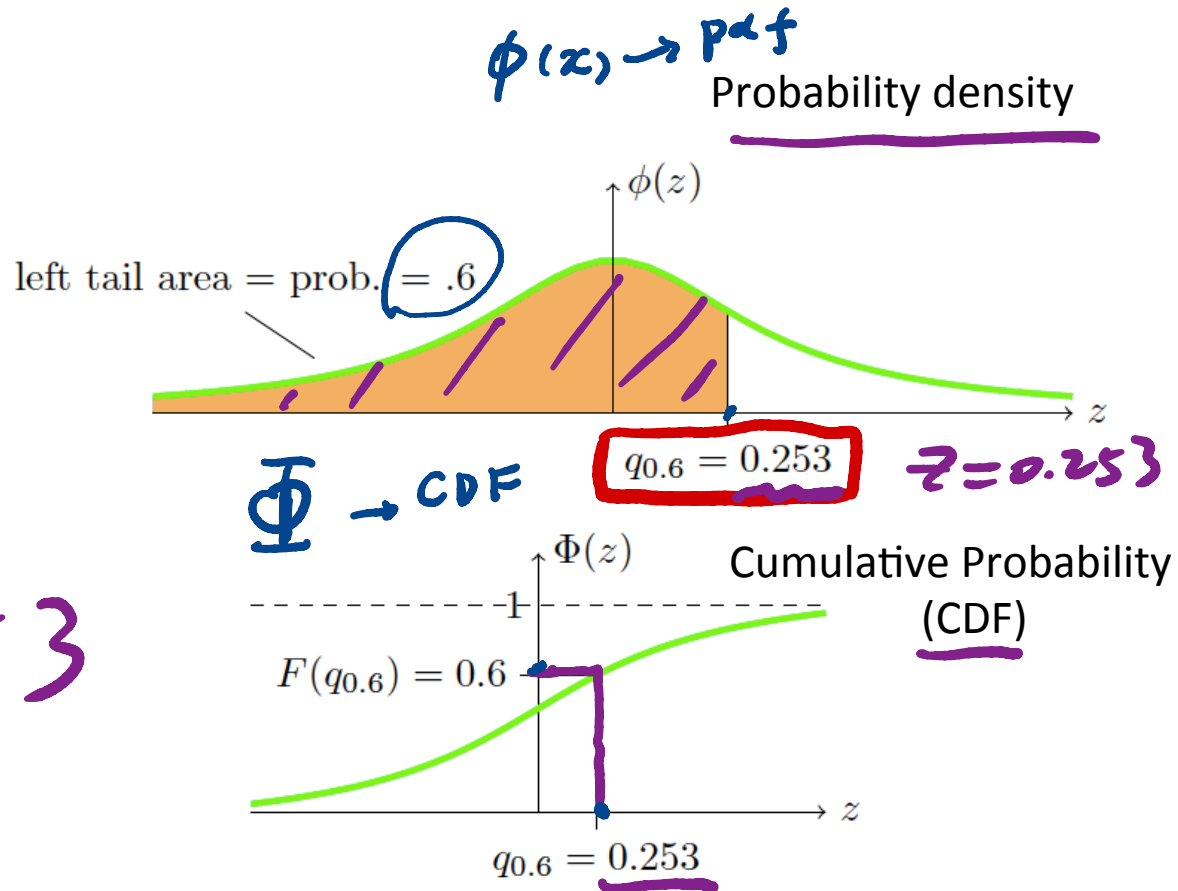
$q_{0.6}$: left tail area $= 0.6 \Leftrightarrow F(q_{0.6}) = 0.6$

Q.

✳ What is the value of <u>50% quantile in</u> a standard normal distribution?

A. -1

B. 0

C. 1
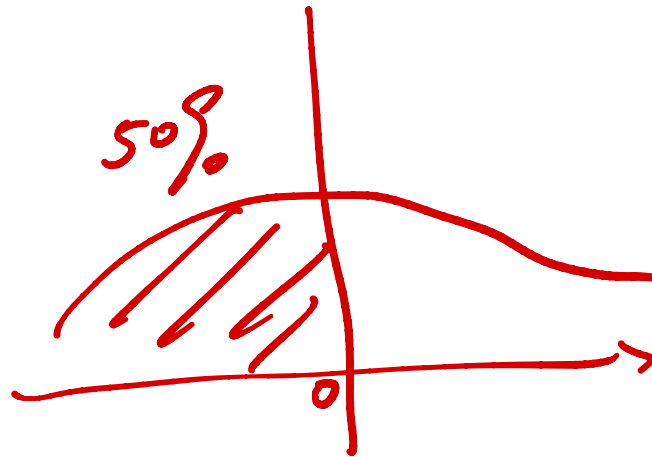
Spread of normal (Gaussian) distributed data
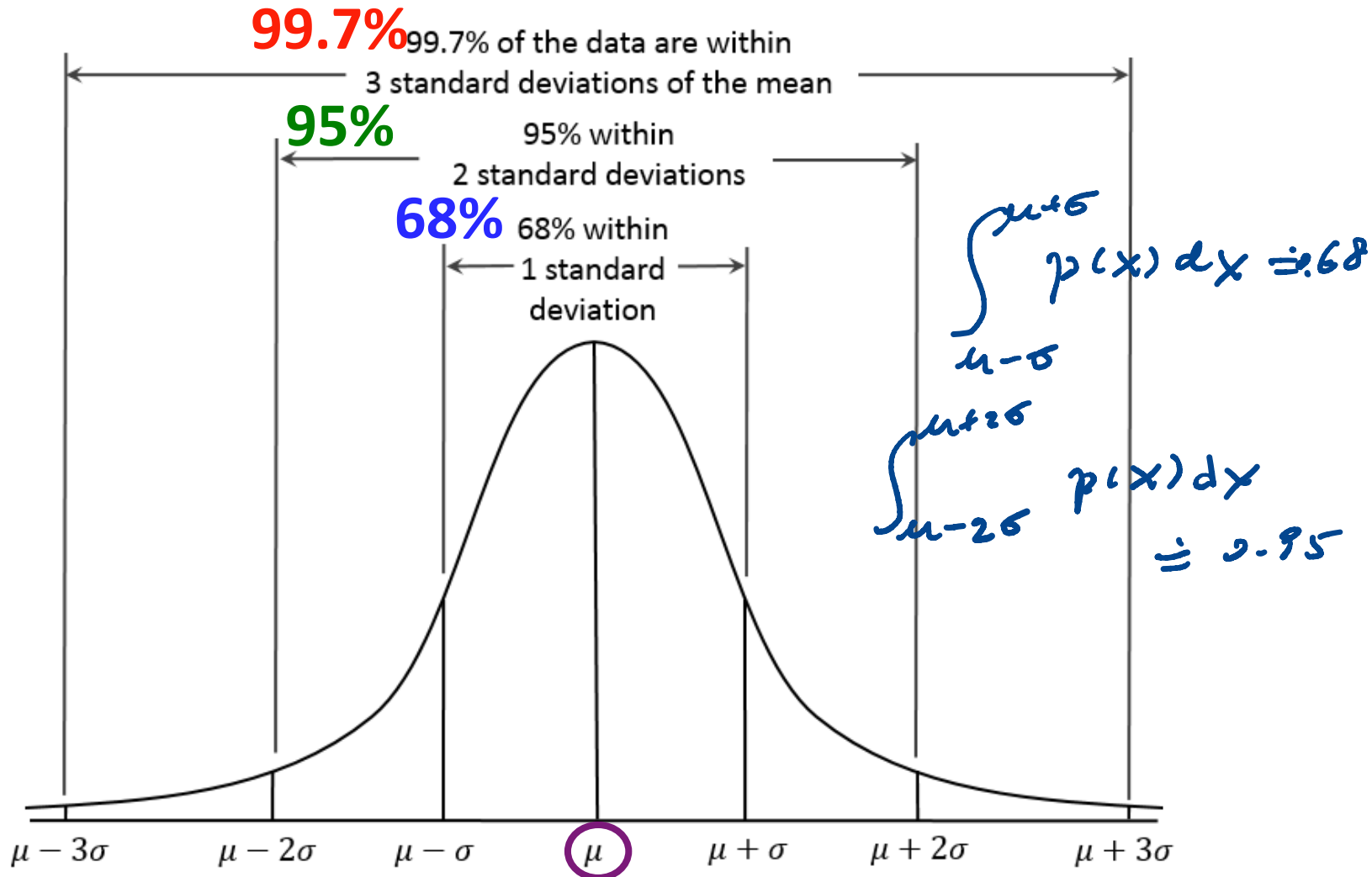
99.7% 99.7% of the data are within
3 standard deviations of the mean

95% 95% within
2 standard deviations

68% 68% within
1 standard
deviation

$$\int_{\mu-\sigma}^{\mu+\sigma} p(x)\,dx \doteq .68$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} p(x)\,dx \doteq 0.95$$

Credit:
wikipedia

$\mu - 3\sigma$ $\mu - 2\sigma$ $\mu - \sigma$ μ $\mu + \sigma$ $\mu + 2\sigma$ $\mu + 3\sigma$

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.

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

✳ 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

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

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

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

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

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

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

$$\int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)\frac{1}{2}}\ dx$$

$$= \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\ dx$$

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

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

$$CDF(b)$$
$$- CDF(a)$$
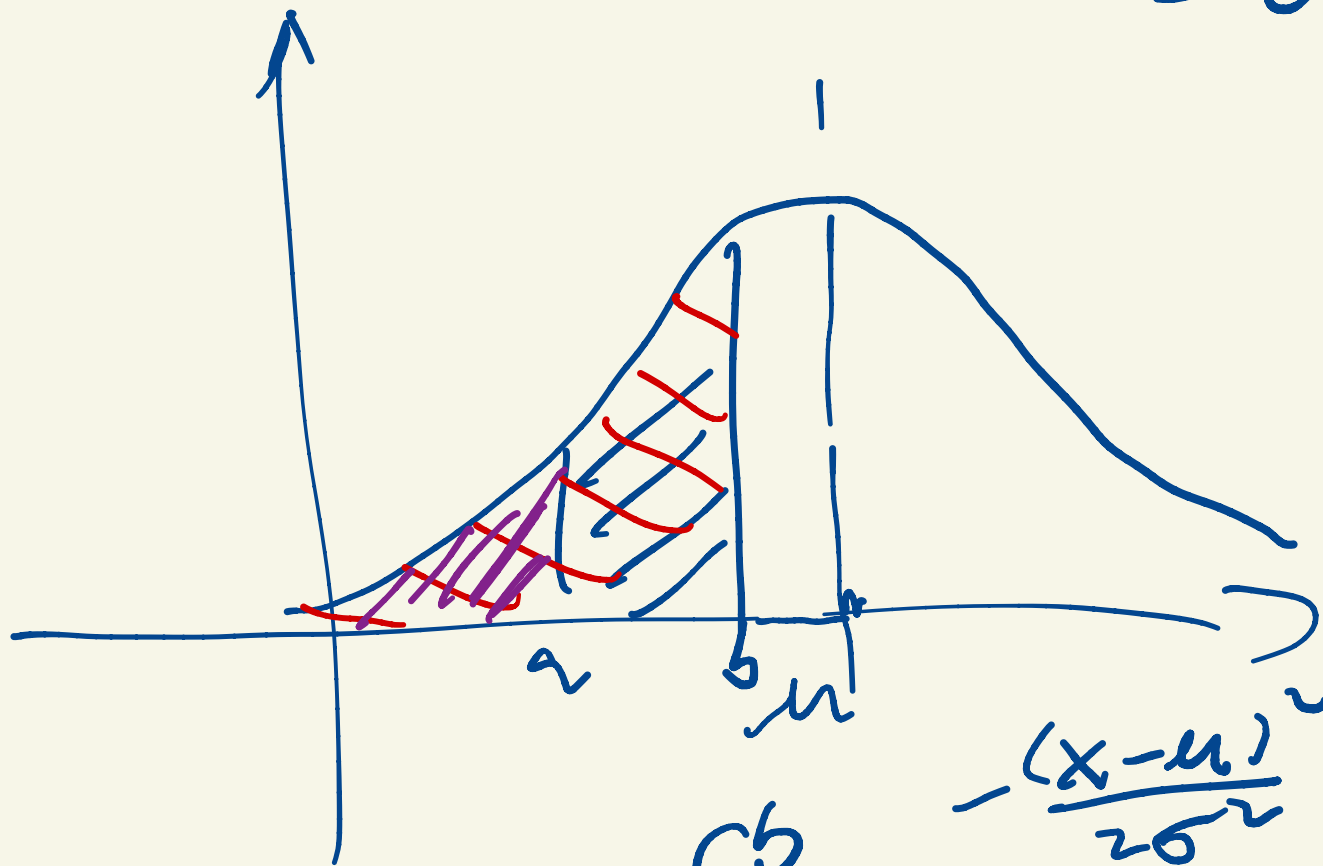
$$P(a \leqslant x \leqslant b) = \int_a^b \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma}\, dx$$

$N(0, 1)$

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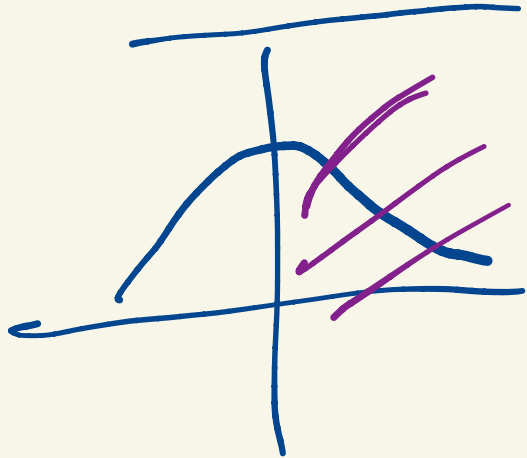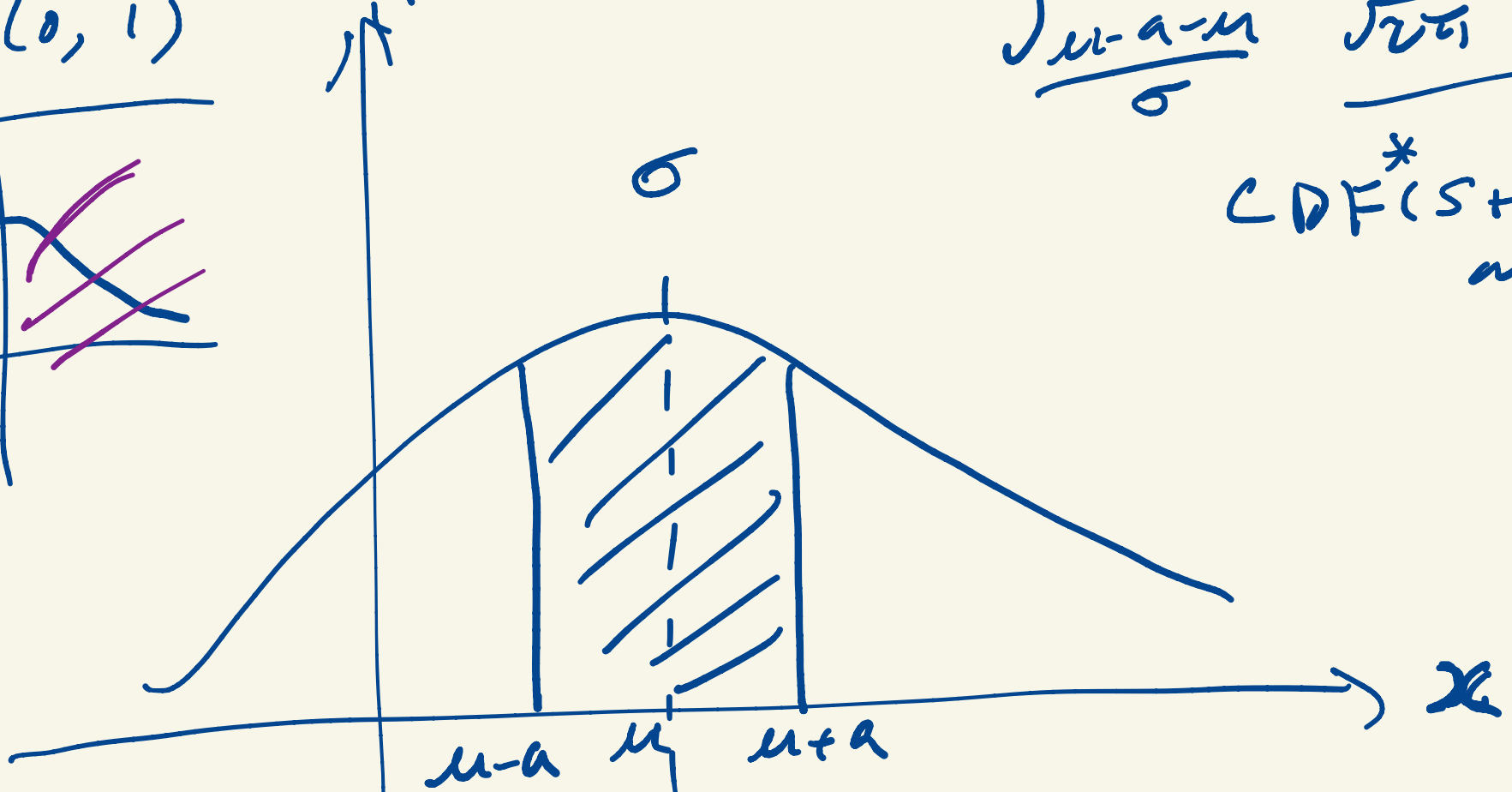
$$\int_{\frac{\mu-a-\mu}{\sigma}}^{\frac{\mu+a-\mu}{\sigma}} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx$$

CDF^* (Standard normal)

σ

$\mu-a \quad \mu \quad \mu+a$

x

$2\left(CDF^*\left(\frac{a}{\sigma}\right) - 0.5\right)$

$CDF^*\left(\frac{\mu+a-\mu}{\sigma}\right) - CDF^*\left(\frac{\mu-a-\mu}{\sigma}\right)$

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

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

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

Another way to check the spread of normal distributed data

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

$1 = \sigma$

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

✳ Fraction of **normal** data within **k** standard deviations from the mean.
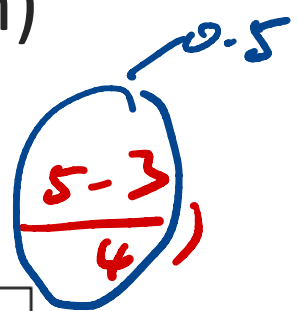
$$\frac{1}{\sqrt{2\pi}} \int_{-k}^{k} exp(-\frac{x^2}{2})dx$$

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

※ If $X \sim N$ (μ=3, σ² =16) (normal distribution)

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

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

$$= P\left(Z \leq \frac{5-3}{4}\right)$$ ← 0.5

$N(0,1)$ ← for

= CDF*
(0.5)

	.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

✳ If $X \sim \boldsymbol{N}$ (μ=3, σ² =16) (normal distribution)

$$P(X \leq 5) =?$$ A . 0.5199 B. 0.5987 C. 0.6915

	.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.

|	.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

$$S = X_1 + X_2 + \cdots + X_N$$

$$N \to \infty$$

$$S \to$$

$$\Phi = CDF$$

$$\lim_{N \to \infty} \Phi(S_N) \to \Phi(\text{Normal})$$
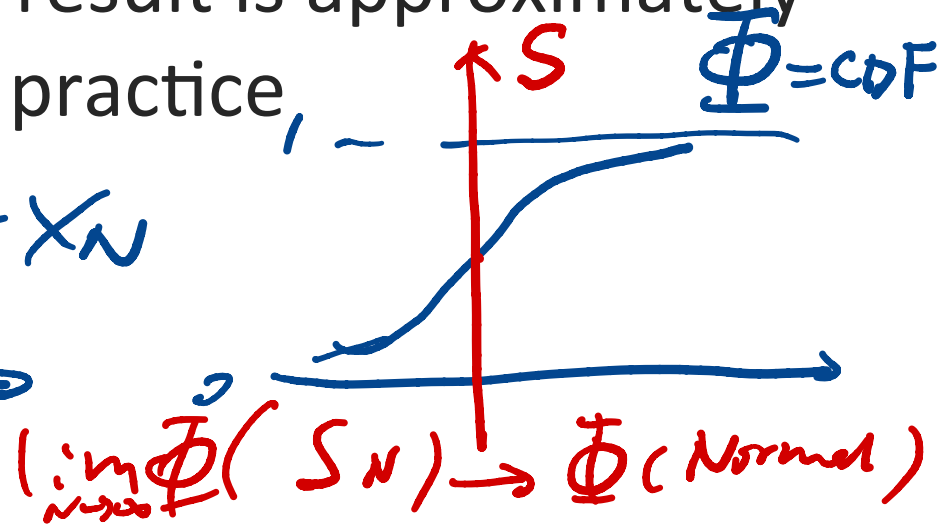
Central limit theorem (CLT)

✳ 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

The Binomial distributed beads of the Galton Board

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

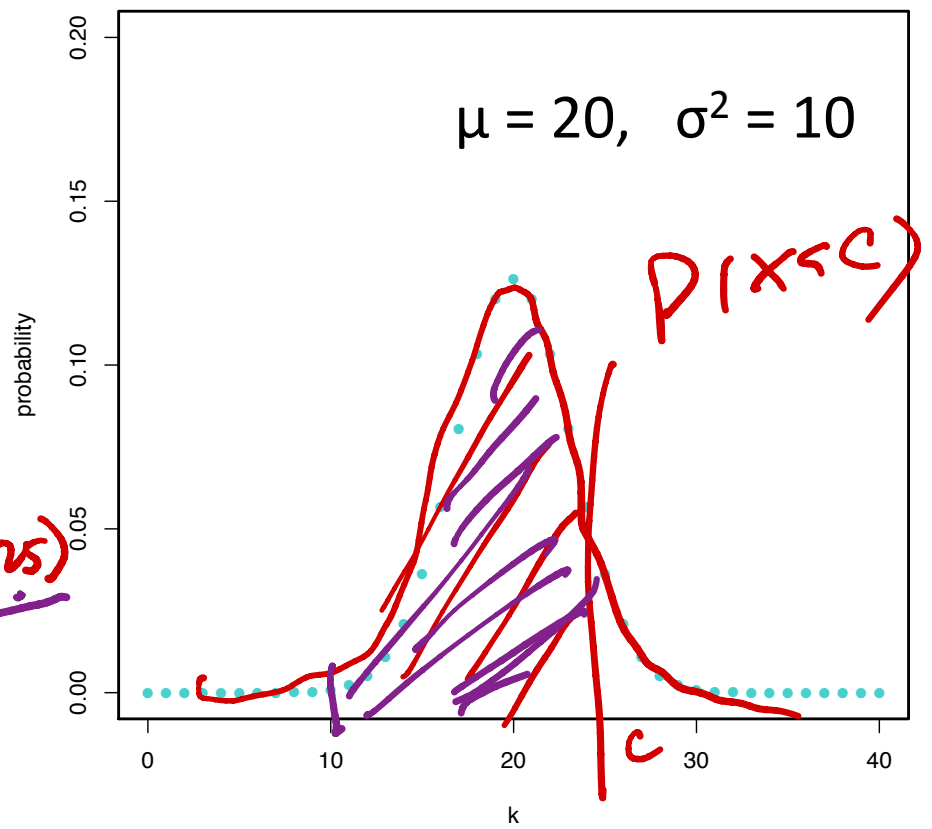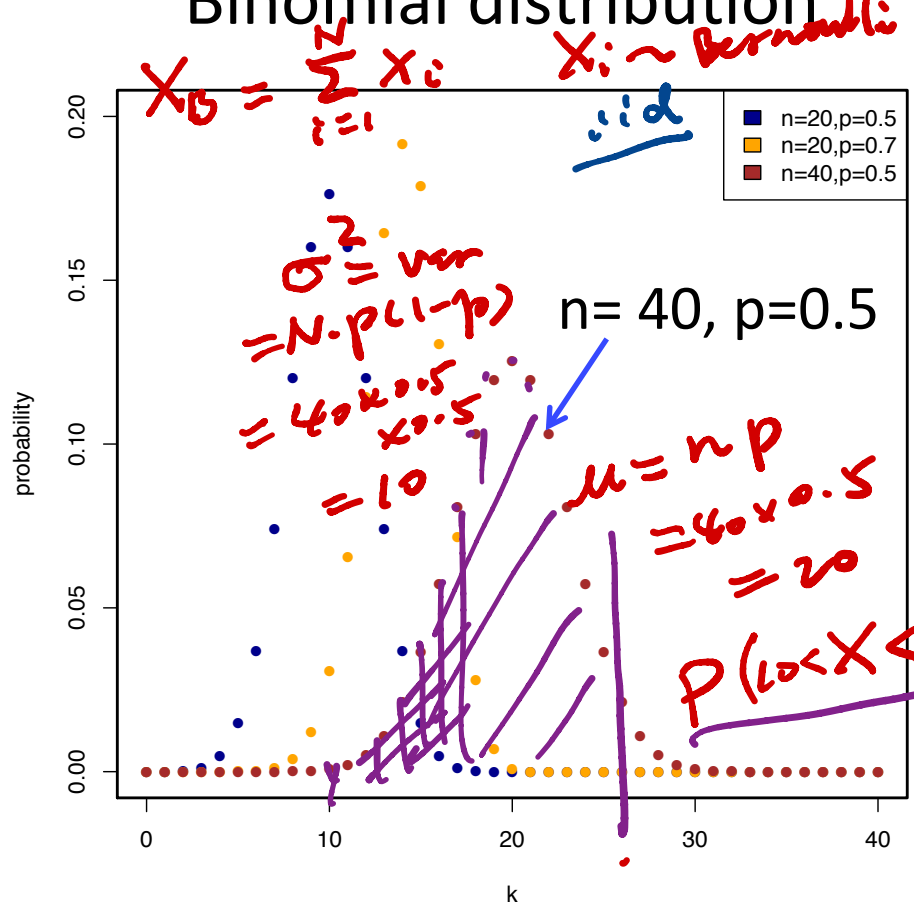# Binomial approximation with Normal

$N = 100$

$$P(X=k) = \binom{N}{k} p^k (1-p)^{N-k}$$

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

Binomial distribution

$X_i \sim$ Bernoulli

$X_B = \sum_{i=1}^{N} X_i$

i.i.d

n=20, p=0.5
n=20, p=0.7
n=40, p=0.5

$\sigma^2 = $ var
$= N \cdot p(1-p)$
$= 40 \times 0.5 \times 0.5$
$= 10$

n= 40, p=0.5

$\mu = n p$
$= 40 \times 0.5$
$= 20$

$P(10 < X < 25)$

Approximation with Normal

$\mu = 20, \quad \sigma^2 = 10$

$P(X < C)$

C

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

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

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

$$E[k] = np = 40 \cdot 0.5 = 20 \qquad std[k] = \sqrt{np(1-p)}$$
$$= \sqrt{40 \cdot 0.5 \cdot 0.5} = \sqrt{10}$$

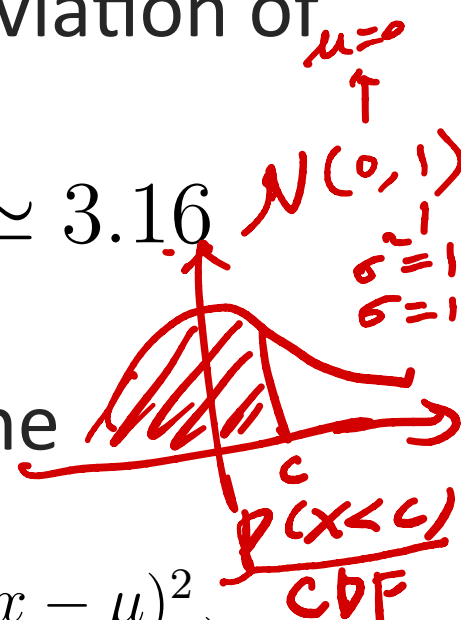# Binomial approximation with Normal

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

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

✳ Then standardize the normal to do the calculation

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

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

$$\simeq 0.94$$

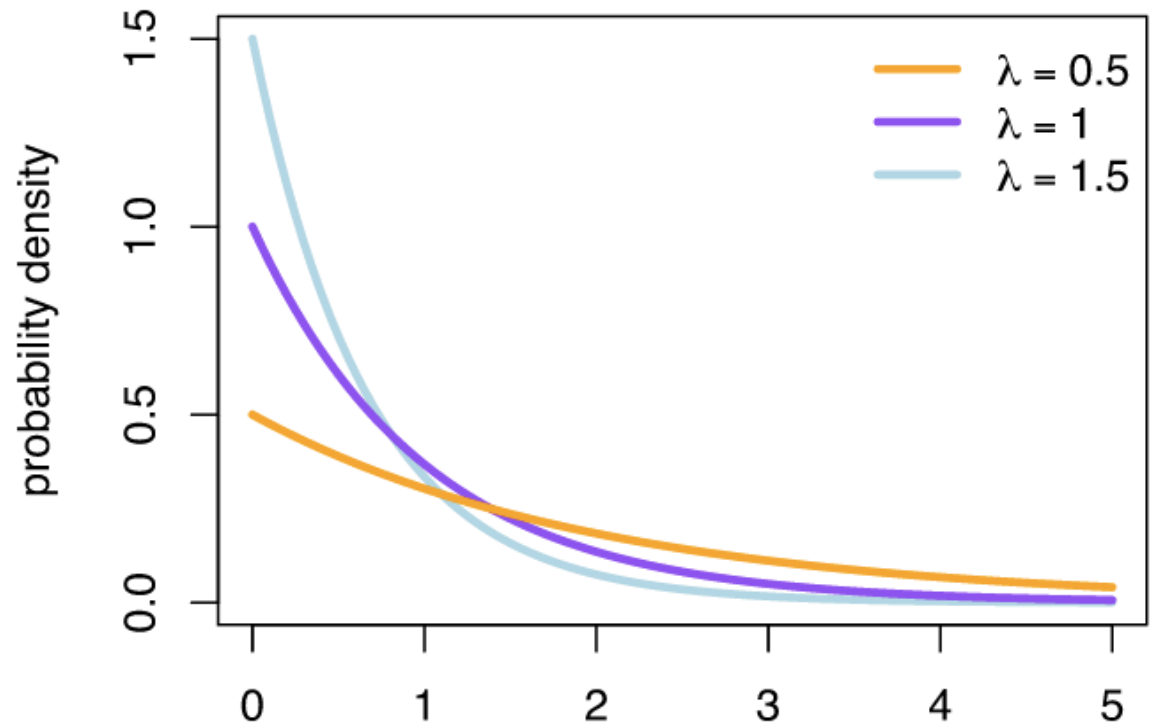*(handwritten annotations:)* $\mu=0$, $N(0,1)$, $\sigma^2=1$, $\sigma=1$, c, $P(X<c)$, CDF, $\hat{X} = \frac{X-\mu}{\sigma}$

Exponential distribution

* Common Model for waiting time

* Associated with the Poisson distribution with the same **λ**

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

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

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**.

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

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

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?

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?

    A. Yes   B. No

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!