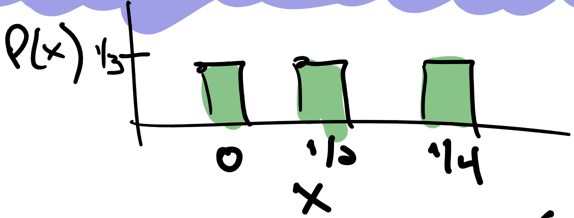
CS2810 Day 16

Mar 22

Admin:
Quiz2 grades release thurs (quiz2 mean .830 vs quiz 1 mean .832)

Content:
Continuous Distributions
Normal Distributions
Cumulative Distribution Functions
Central Limit Theorem

Discrete Distribution vs Continuous Distribution

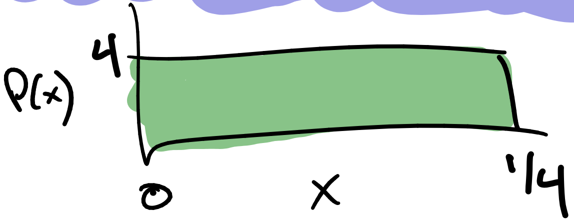


X has sample space $\{0, 1/2, 1/4\}$

$$\sum_i P(x_i) = 1$$

$$0 \le P(x_i) \le 1$$

Distribution is Prob Mass Function
PMF

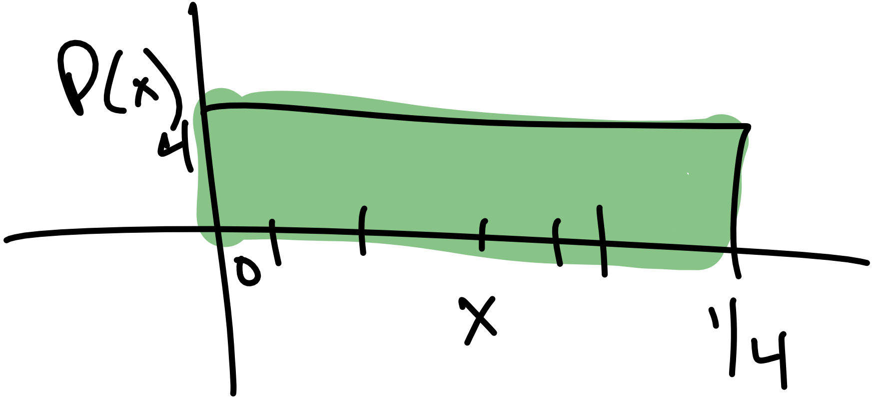


X takes any value from 0 to 1/4

$$\int_{-\infty}^{\infty} P(x) = 1$$

$$0 \le P(x_i)$$

Distribution is Prob Density Fnc
PDF

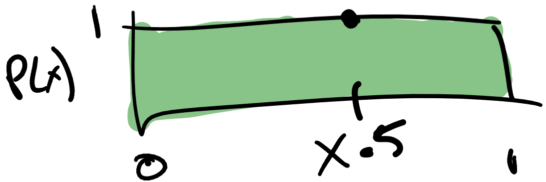


$$\int_{-\infty}^{\infty} P(x) = 1$$

THE $P(x) = x$ CONONDRUM

LET X BE A CONTINUOUS RANDOM VARIBLE ON $[0,1]$

(INCLUDES 0,1)



WHAT IS PROB THAT X IS EXACTLY .5?

$$P(x=.5) > 0$$

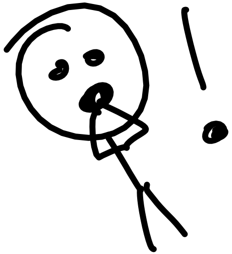
(won't work since there are infinitely many values in domain of x, "sum of (uncountably) infinite number of positive values is infinite")

$$P(x=.5) = 1$$

(this suggests x can't be anything else, no prob leftover)

THE $P(x) = X$ CONONDRUM

$$P(x = 1/2) = 0$$



PROBABILITY OF OUTCOMES IN CONTINUOUS DISTRIBUTION CAN BE ZERO

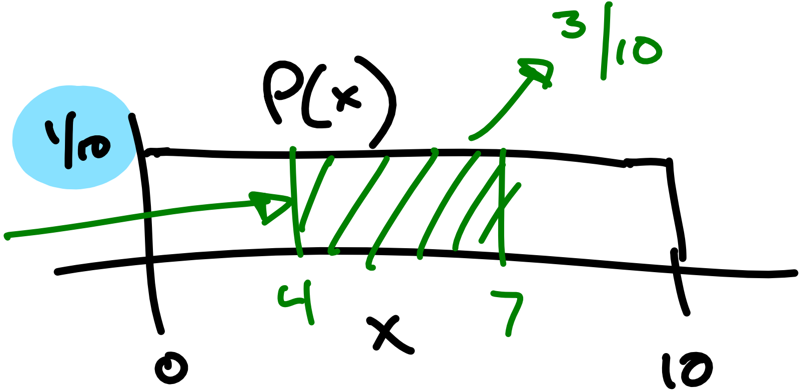
in a continuous distribution, the probability of an outcome (which happens) is zero

SO WHATS PDF GOOD FOR THEN?

CONTINUOUS DISTRIBUTION

Continuous Distribution's Prob Distribution Function gives the probability an outcome falls in some range via integration:

GIVEN X HAS PDF



COMPUTE

AREA UNDER PDF FROM 4 TO 7

$$P(4 < x < 7) = \int_4^7 \frac{1}{10}\,dx = \frac{x}{10}\Big|_4^7 = \frac{7}{10} - \frac{4}{10}$$

$$= 3/10$$

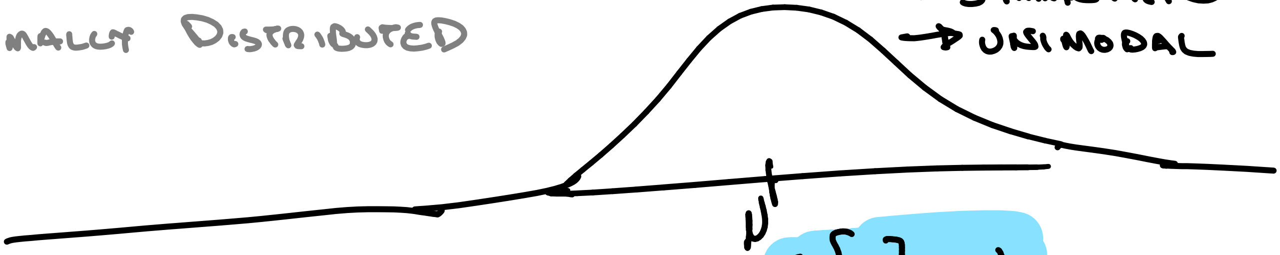
Normal | Gaussian Distribution

- → Continuous
- → Unbounded
- → Symmetric
- → Unimodal

X is R.V. Normally Distributed

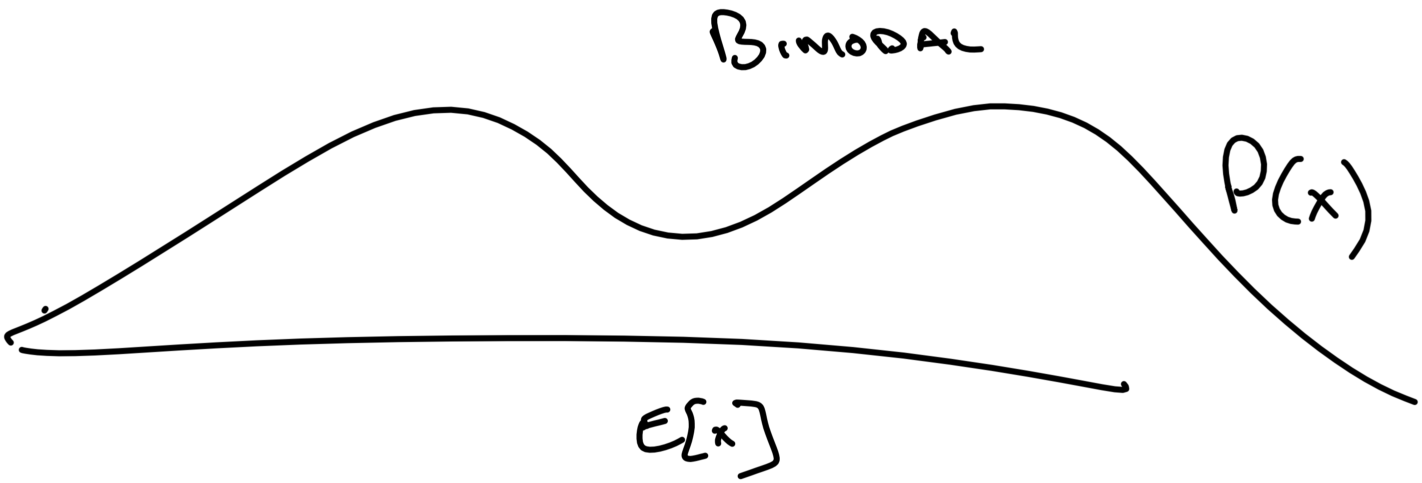
$$X \sim N(\mu, \sigma^2)$$

$$\Rightarrow P(X=x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$



$$E[x] = \mu$$

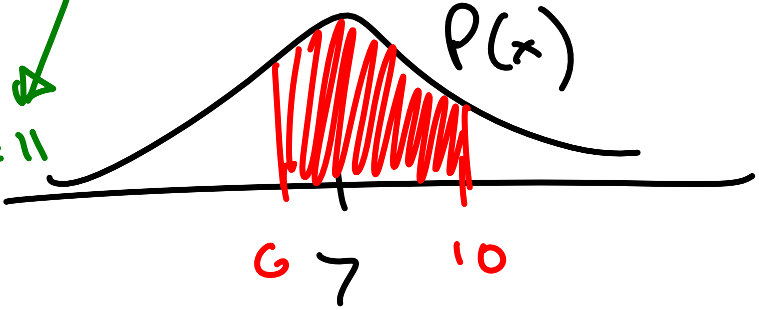
$$Var(x) = \sigma^2$$


BIMODAL
P(x)
E[x]

GIVEN $X \sim N(7, 11)$ COMPUTE PROB THAT

X IS BETWEEN 6 AND 10

$\mu = 7 \quad \sigma^2 = 11$

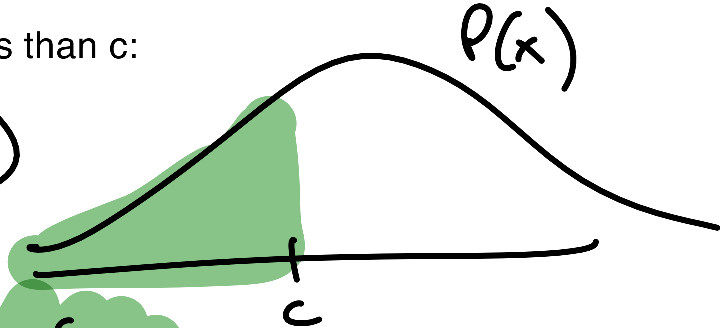


$$P(6 < x < 10) = \int_6^{10} \frac{1}{\sqrt{11}\sqrt{2\pi}} e^{-1/2 \frac{(x-7)^2}{11}} dx$$

Cumulative Distribution Functions

The CDF gives the probability that an outcome is less than c:

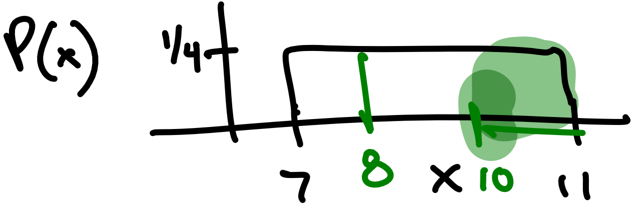
Given RV x with PDF $p(x)$



$$\mathrm{CDF}(c) = P(x < c) = \int_{-\infty}^{c} p(x)\,dx$$

CDF Demo

ICA



Given the PDF to the right, compute:

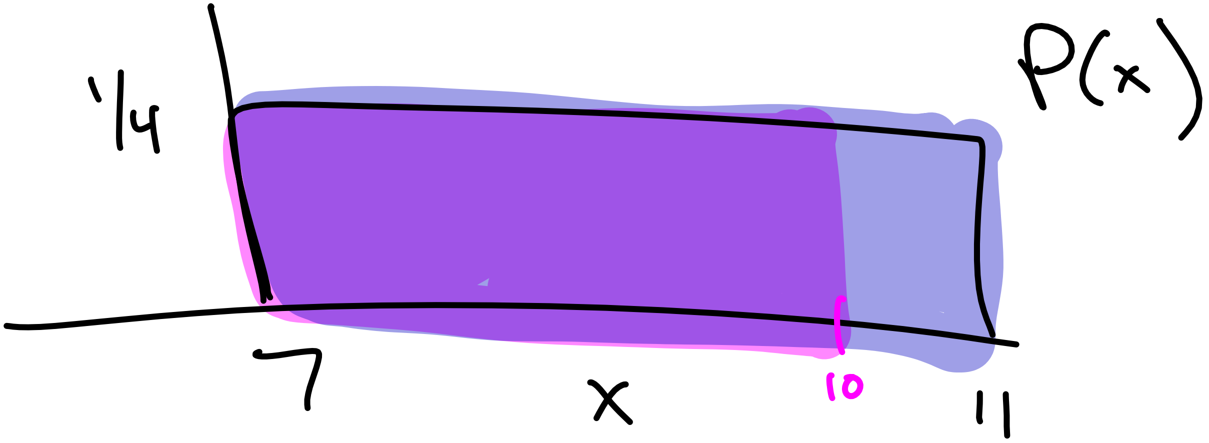
1. the probability that X is between 10 and 11
2. the probability X is less than 8
3. Express your answers for each of the above using only the CDF (not the PDF)

(+) As c gets smaller and smaller, describe the behavior of the CDF
(+) As c gets larger and larger, describe the behavior of the CDF

CDF

① $P(10 \le x \le 11) = 1/4 = CDF(11) - CDF(10)$

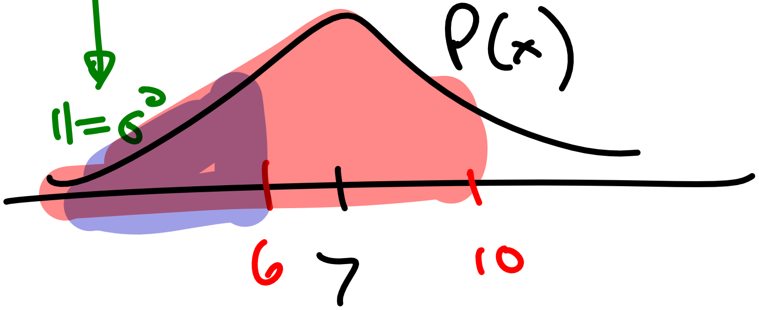
② $P(x \le 8) = 1/4 = CDF(8) - CDF(7)$



$$P(10 \le x \le 11) = 1/4 = CDF(11) - CDF(10)$$

Given $X \sim N(7, 11)$ compute prob that X is between 6 and 10

$7 = \mu$ $\quad$ $11 = \sigma^2$

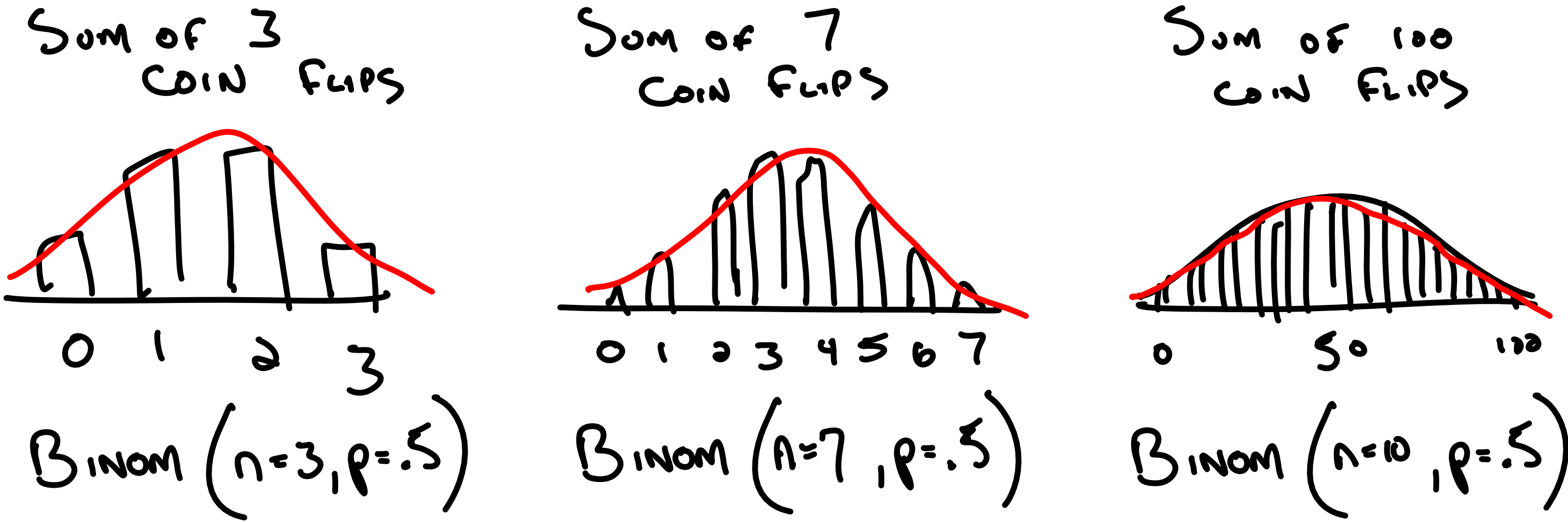


Demo: Updated Prob / Stats Calculator

$CDF(10) - CDF(6) \hat{=} .435$

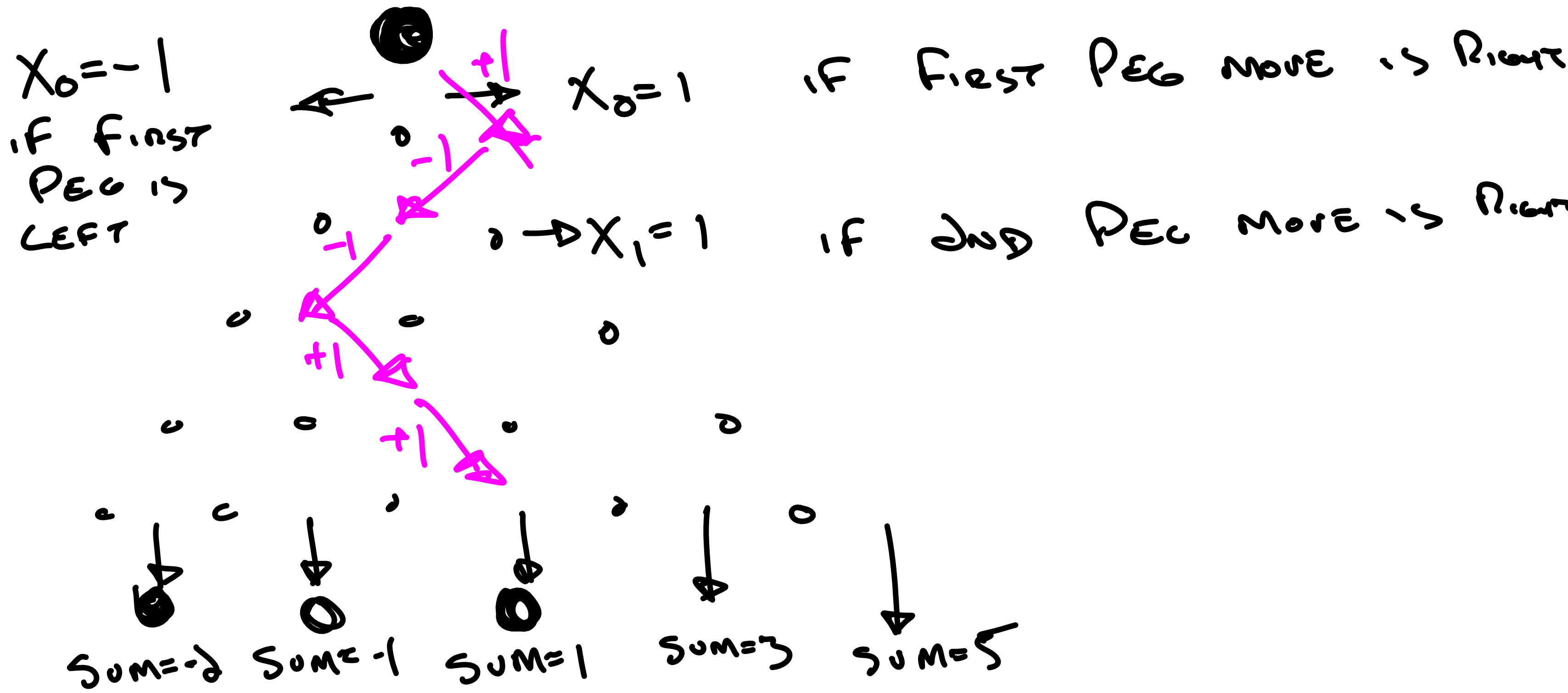
Central Limit Theorem

As we add more independent random variables together, the resulting sum gets closer and closer to normally distributed



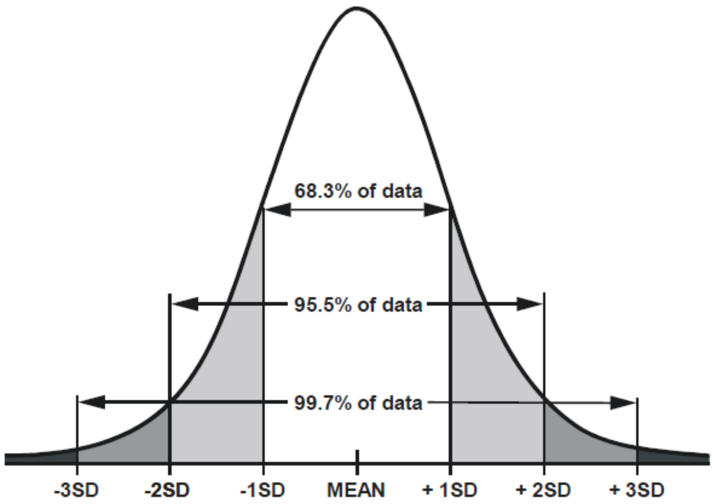
Demo: CLT

Galton Board on YouTube



$X_0 = -1$
if first peg is left
$X_0 = 1$
if first peg move is right
$X_1 = 1$
if 2nd peg move is right
+1
-1
-1
+1
+1
Sum=-2
Sum=-1
Sum=1
Sum=3
Sum=5

NORMAL DISTRIBUTIONS USEFUL LANDMARKS



95% OF DATA LIES IN 1.96 STD DEVIATIONS

Why is the normal distribution so popular?

1. We often sum independent random variables

   CLT tells us this sum is (roughly) normally distributed

2. Linear Functions of Normal Random Variables are also Normal!

   And the linearity of expectation formulae can tell us the mean and variance of the sum!

ICA 2

$X_0 + X_1 + X_2 + X_3 + \dots + X_{399}$ ← WHOLE SUM IS (ROUGHLY) NORMALLY DISTRIBUTED (CLT)

The following distribution gives the prices of ice creams bought at a shop.

x	\$1	\$2	\$10
$P(x)$	.5	.4	.1

What is the probability that the shop sells more than \$1000 of ice cream to 400 customers? (make any assumptions you deem necessary)

→ EACH CUSTOMER BUYS ICE CREAM INDEPENDENTLY

X	$1	$2	$10
P(X)	.5	.4	.1

$$E[X] = 1 \cdot .5 + 2 \cdot .4 + 10 \cdot .1$$

$$= .5 + .8 + 1 = 2.3$$

$$E[X^2] = 1^2 \cdot .5 + 2^2 \cdot .4 + 10^2 \cdot .1$$

$$= .5 + 1.6 + 10 = 12.1$$

$$\mathrm{VAR}(X) = E[X^2] - E[X]^2 = 12.1 - 2.3^2 = 6.81$$

$$E[X_0 + X_1 + X_2 + X_3 + \ldots + X_{399}] =$$

$$E[X_0] + E[X_1] + \ldots + E(X_{399}) = 400\,E[X]$$

$$= 400 \cdot 2.3$$

$$= 920$$

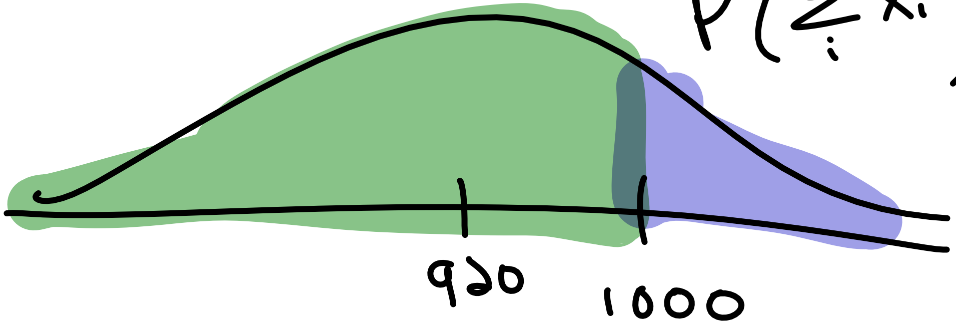
$$VAR(X_0 + X_1 + X_2 + X_3 + \ldots + X_{399})$$

$$= VAR(X_0) + VAR(X_1) + \ldots + VAR(X_{399}) = 400 \cdot 6.81$$

$$\simeq 2724$$

INDEP

$$P\left(\sum_i x_i\right) \sim N(920, 2724)$$

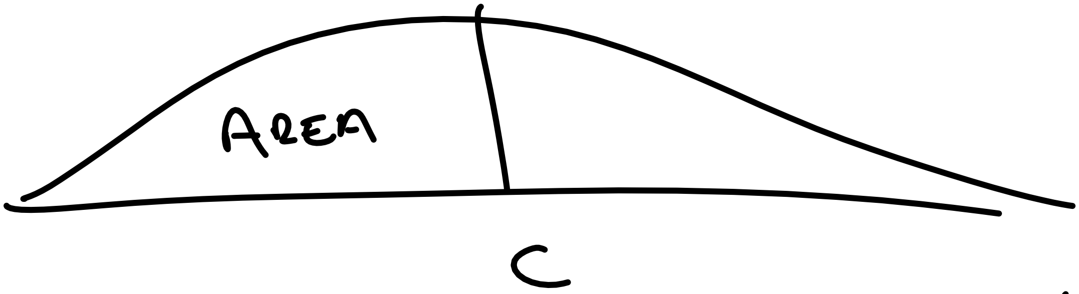


$$1 - CDF(1000)$$

$$E[X_0 + X_1] = E[X_0] + E[X_1]$$

$$VAR(X_0 + X_1) = VAR(X_0) + VAR(X_1)$$

INDEP



AREA = CDF(C)

PPF(AREA) = C

INVERSE OF CDF = PPF

PERCENT POINT FUNCTION

$VAR(cX) = c^2 \, VAR(X)$ $VAR(X_0 + X_1) = VAR(X_1) + VAR(X_0)$

$$400X \neq X_0 + X_1 + \ldots + X_{399}$$

ONE CUSTOMER'S BILL $\times$ 400

SUM OF 400 INDEP