CS 109 Quiz 2 Review Session

Anand Shankar and Grace Hu

February 22, 2021

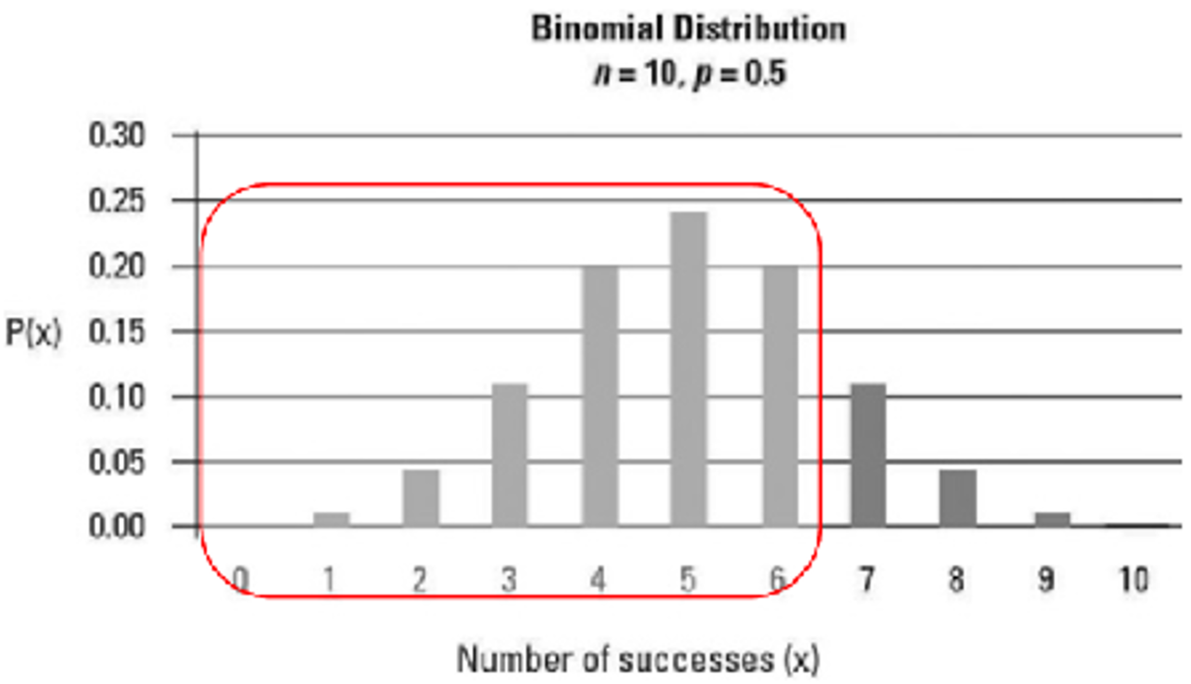
Review Time!

- **Random Variables**
- **Joint Distributions**
- **Joint RV Statistics**
- **Conditional Distribution**

Probability Distributions

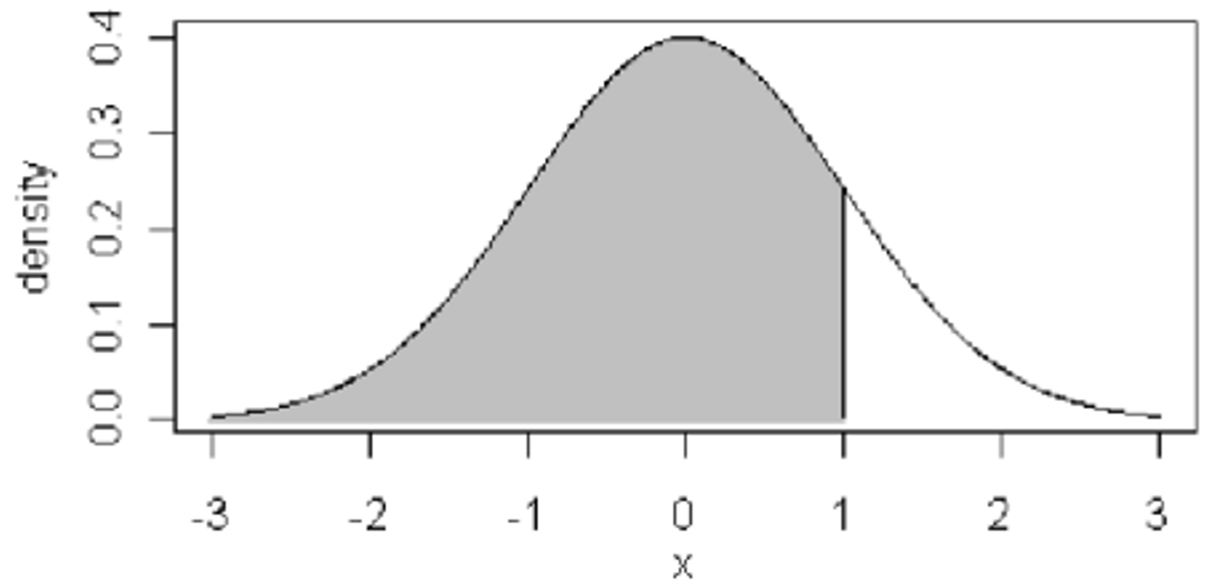
Discrete

CDF:



Continuous

CDF:



Expectation & Variance

Discrete definition

$$E[X] = \sum_{x:p(x)>0} p(x) \cdot x$$

Continuous definition

$$E[X] = \int_a^b x \cdot f(x)dx$$

Properties of Expectation

$$E[X + Y] = E[X] + E[Y]$$

$$E[aX + b] = aE[X] + b$$

$$E[g(X)] = \sum_x g(x)p(x)$$

Properties of Variance

$$Var(X) = E[(X - \mu)^2]$$

$$Var(X) = E[X^2] - E[X]^2$$

$$Var(aX + b) = a^2 Var(X)$$

Discrete Random Variables

Ber(p)	Bin(n, p)	Poi(λ)	Geo(p)	NegBin(r,p)
P(X) = p	$\binom{n}{k} p^k (1-p)^{n-k}$	$\frac{\lambda^k e^{-\lambda}}{k!}$	$(1-p)^{k-1}p$	$\binom{k-1}{r-1} p^r (1-p)^{k-r}$
E[X] = p	E[X] = np	E[X] = λ	E[X] = 1 / p	E[X] = r / p
Var(X) = p(1-p)	Var(X) = np(1-p)	Var(X) = λ	$\frac{1-p}{p^2}$	$\frac{r(1-p)}{p^2}$
1 experiment with prob p of success	n independent trials with prob p of success	Number of success over experiment duration, λ rate of success	Number of independent trials until first success	Number of independent trials until r successes

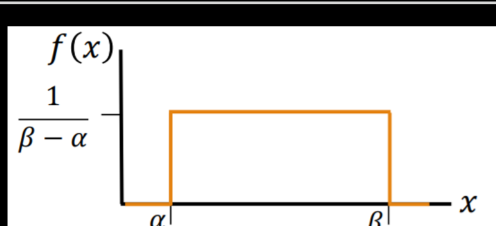
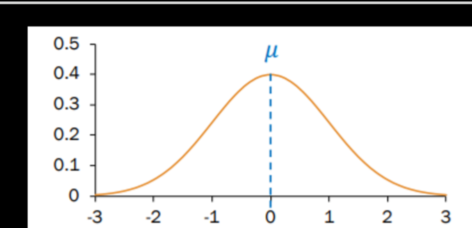
Continuous Random Variables - PDF

For continuous RVs, we need to calculate the PDF, instead of the PMF

PDF for RV X
$f(x) \geq 0$ such that $-\infty < x < \infty$

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$

Continuous Random Variables

Uni(α, β)	Exp(λ)	N(μ, σ ^{2})
$\boldsymbol{f(x) = \frac{1}{\beta - \alpha}}$	$f(x) = \lambda e^{-\lambda x}$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
$P(a \leq X \leq b) = \frac{b-a}{\beta-\alpha}$	$F(x) = 1 - e^{-\lambda x}$	$\boldsymbol{F(x) = \phi\left(\frac{x-\mu}{\sigma}\right)}$
$\mathbf{E}[\boldsymbol{X}] = \boldsymbol{\frac{\alpha + \beta}{2}}$	E[X] = 1 / λ	E[X] = μ
$Var(x) = \frac{(\beta-\alpha)^2}{12}$	$Var(x) = \frac{1}{\lambda^2}$	$Var(x) = \sigma^2$
	Duration of time until success occurs. λ is rate of success	

Approximations

When can we approximate a binomial?

Poisson

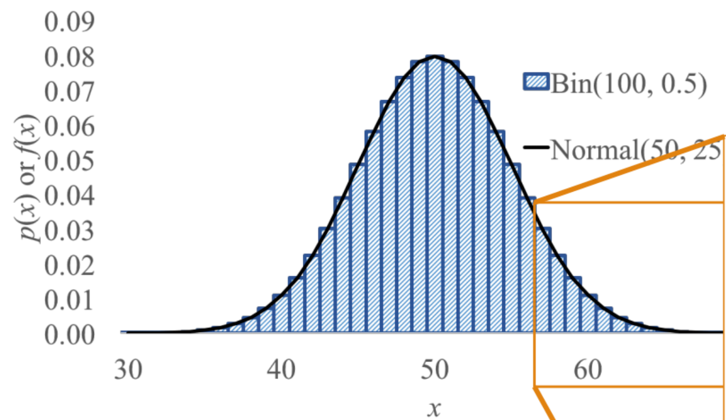
- $n > 20$
- p is small
- $\lambda = np$ is moderate
  - $n > 20$ and $p < 0.05$
  - $n > 100$ and $p < 0.1$
- Slight dependence ok

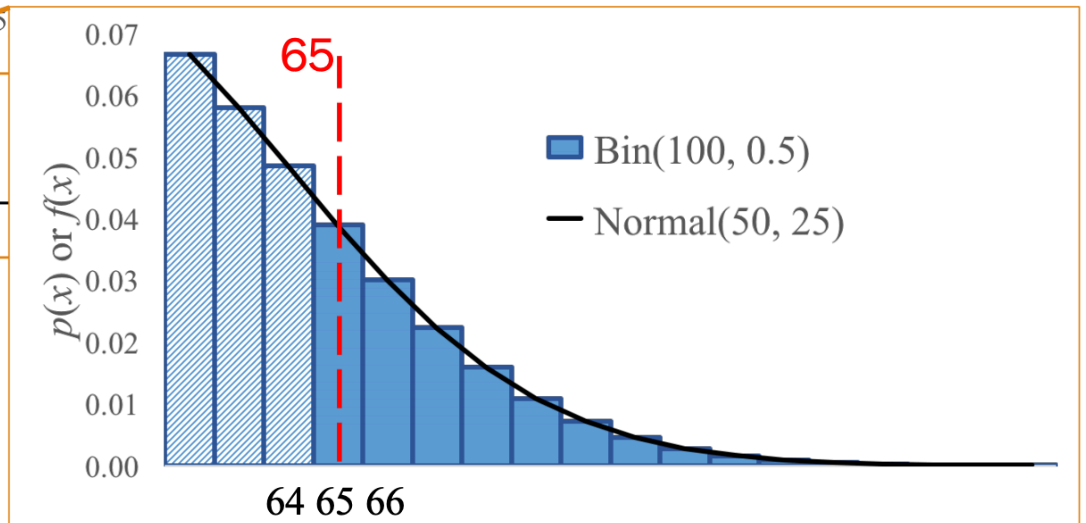
Normal

- $n > 20$
- p is moderate
  - $np(1-p) > 10$
- Independent trials

Continuity Correction

In our website testing, $Y \sim \mathcal{N}(50, 25)$ approximates $X \sim \text{Bin}(100, 0.5)$.





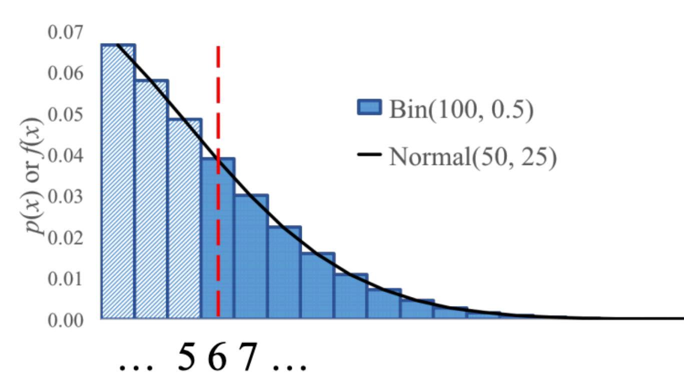
$P(X \geq 65)$ Binomial

$\approx P(Y \geq 64.5)$ Normal

≈ 0.0018 ✅ the better Approach 2

You must perform a **continuity correction** when approximating a Binomial RV with a Normal RV.

Discrete (e.g., Binomial) probability question	Continuous (Normal) probability question
$P(X = 6)$	$P(5.5 \leq Y \leq 6.5)$
$P(X \geq 6)$	$P(Y \geq 5.5)$
$P(X > 6)$	$P(Y \geq 6.5)$
$P(X < 6)$	$P(Y \leq 5.5)$
$P(X \leq 6)$	$P(Y \leq 6.5)$



Joint Distributions – Discrete

For two discrete joint random variables X and Y,
the joint probability mass function is defined as:

$$p_{X,Y}(a,b) = P(X = a, Y = b)$$

The marginal distributions of the joint PMF are defined as:

$$p_X(a) = P(X = a) = \sum_y p_{X,Y}(a,y)$$

$$p_Y(b) = P(Y = b) = \sum_x p_{X,Y}(x,b)$$

Use marginal distributions to get a 1-D RV from a joint PMF.

Multinomial RVs

Consider an experiment of n independent trials:

- Each trial results in one of m outcomes. $P(\text{outcome } i) = p_i$, $\sum_{i=1}^{m} p_i = 1$
- Let X_i= # trials with outcome i

Joint PMF

$$P(X_1 = c_1, X_2 = c_2, \dots, X_m = c_m) = \binom{n}{c_1, c_2, \dots, c_m} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}$$

where $\sum_{i=1}^{m} c_i = n$ and $\sum_{i=1}^{m} p_i = 1$

Example:

- Rolling 2 twos, 3 threes, and 5 fives on 10 rolls of a fair-sided die
- Generating a random 5-word phrase with 1 “the”, 2 “bacon”, 1 “put”, 1 “on”

Independent Discrete RVs

Two discrete random variables X and Y are independent if for all x, y:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Sum of Independent RVs

Sum of independent Binomials

- If $X \sim Bin(n_1, p)$ and $Y \sim Bin(n_2, p)$ and X and Y are independent, then $X + Y \sim Bin(n_1 + n_2, p)$

Sum of independent Poisson RVs

- If $X \sim Poi(\lambda_1)$ and $Y \sim Poi(\lambda_2)$ and X and Y are independent, then $X + Y \sim Poi(\lambda_1 + \lambda_2)$

Sum of Independent Normal RVs

- If $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$and X and Y are independent, then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Variance

For any random variables X and Y,

$$Var(X+Y) = Var(X) + 2 \cdot Cov(X,Y) + Var(Y)$$

When X and Y are independent,

$$Var(X+Y) = Var(X) + Var(Y)$$

Covariance

$$
\begin{aligned}
Cov(X, Y) &= E[(X - E[X])(Y - E[Y])] \\
&= E[XY] - E[X]E[Y]
\end{aligned}
$$

Note: if X and Y are independent, then Cov(X, Y) = 0. We can’t make the same assumption in the reverse direction.

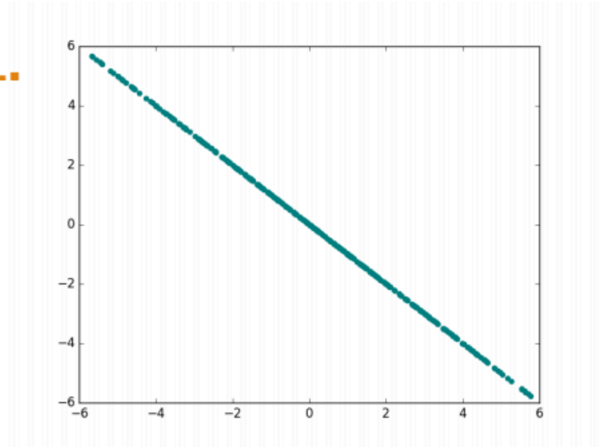
Correlation

Correlation reps

What is the correlation coefficient $\rho(X,Y)$?

A. $\rho(X,Y) = 1$
B. $\rho(X,Y) = -1$
C. $\rho(X,Y) = 0$
D. Other

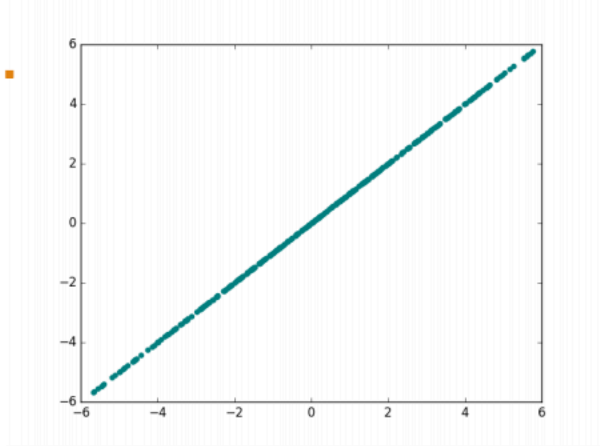
1.



B. $\rho(X,Y) = -1$

$$Y = -aX + b$$
$$a > 0$$

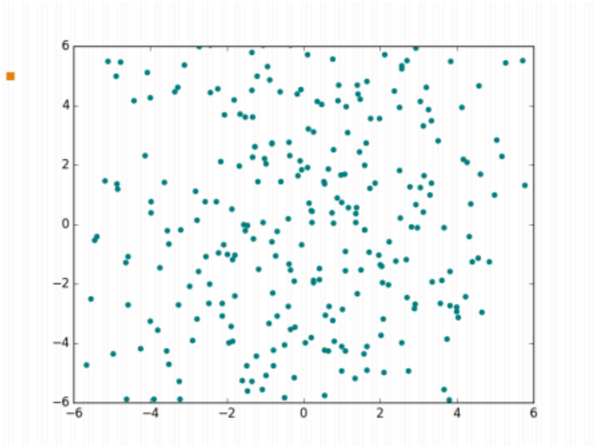
2.



A. $\rho(X,Y) = 1$

$$Y = aX + b$$
$$a > 0$$

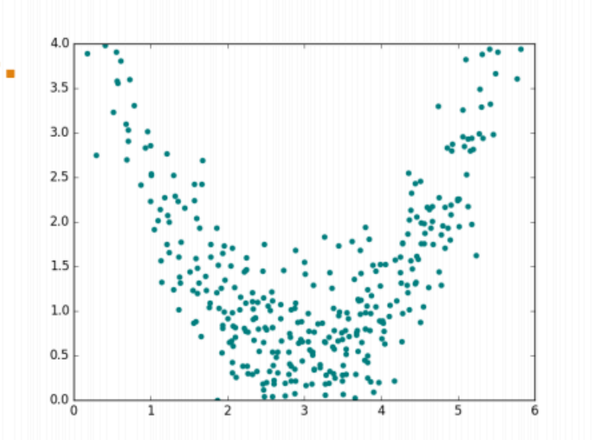
3.



C. $\rho(X,Y) = 0$

"uncorrelated"

4.



C. $\rho(X,Y) = 0$

$$Y = X^2$$

X and Y can be nonlinearly related even if $\rho(X,Y) = 0$.

Correlation

Correlation of X and Y

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$\sigma_X^2 = \text{Var}(X)$,
$\sigma_Y^2 = \text{Var}(Y)$

Note: $-1 \leq \rho(X,Y) \leq 1$
Measures the linear relationship between X and Y

$\rho(X,Y) = 1 \quad \Rightarrow Y = aX + b$, where $a = \sigma_Y / \sigma_X$
$\rho(X,Y) = -1 \quad \Rightarrow Y = aX + b$, where $a = -\sigma_Y / \sigma_X$
$\rho(X,Y) = 0 \quad \Rightarrow$ "uncorrelated" (absence of linear relationship)

Conditional Distribution

Conditional PMF for discrete X given Y

$$P(X = x|Y = y) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

Conditional Expectation

$$E[X|Y = y] = \sum_x xP(X = x|Y = y)$$

Law of Total Expectation

$$E[E[X|Y]] = \sum_{y} P(Y = y)E[X|Y = y] = E[X]$$

If we only have a conditional PMF of X on some discrete variable Y, we can compute $E[X]$ as follows:

1. Compute expectation of X given some value of $Y = y$
2. Repeat step 1 for all values of Y
3. Compute a weighted sum (where weights are $P(Y = y)$)

```
def recurse():
  if (random.random() < 0.5):
      return 3
  else: return (2 + recurse())
```

Useful for analyzing recursive code!!

Practice Problems!

Practice Problems

The Huffmeister flood plane in Houston has historically been estimated to flood at an average rate of 1 flood for every 500 years.

- What is the probability of observing at least 3 floods in 500 years?
- What is the probability that a flood will occur within the next 100 years?
- What is the expected number of years until the next flood?

Practice Problems

- What is the probability of observing at least 3 floods in 500 years?
  - Let X denote the number of floods we observe in 500 years.
  - $X \sim Poi(1)$; we observe 1 flood every 500 years
  - $\mathrm{P}(X \geq 3) = 1 - \mathrm{P}(X < 3) = 1 - \sum_{i=0}^{2} P(X = i) = 1 - \frac{5}{2e}$
- What is the probability that a flood will occur within the next 100 years?
  - Let Y denote the number of years until the next flood.
  - $Y \sim \mathrm{Exp}(1/500)$
  - $P(Y \leq 100) = F(100) = 1 - e^{-0.2} = 0.181$
- What is the expected number of years until the next flood?
  - $E[Y] = \frac{1}{\lambda} = \frac{1}{\frac{1}{500}} = 500$

Practice Problems

```
int Near() {
   int b = randomInteger(1, 4); // equally likely to be 1, 2, 3 or 4
   if (b == 1) return 2;
   else if (b == 2) return 4;
   else if (b == 3) return (6 + Near());
   else return (8 + Near());
}

int Far() {
   int a = randomInteger(1, 3); // equally likely to be 1, 2 or 3
   if (a == 1) return 2;
   else if (a == 2) return (2 + Near());
   else return (4 + Far());
}
```

Let Y = the value returned by `Far()`.

Let X = the value returned by `Near()`.

What is E[Y]?

Practice Problems

Return value for `Far()` is based on `Near()`, so calculate E[X] first.

```
int Near() {
   int b = randomInteger(1, 4); // equally likely to be 1, 2, 3 or 4
   if (b == 1) return 2;
   else if (b == 2) return 4;
   else if (b == 3) return (6 + Near());
   else return (8 + Near());
}
```

$$
\begin{aligned}
E[X] &= \textstyle\sum_{i=1}^{4} P(B = i)E[X|B = i] \\
&= \tfrac{1}{4}\,(2 + 4 + E[6 + X] + E[8 + X]) \\
&= \tfrac{1}{4}\,(2 + 4 + 6 + E[X] + 8 + E[X]) \\
&= \tfrac{1}{4}\,(20 + 2E[X]) \\
&= 5 + \tfrac{1}{2}\,E[X]
\end{aligned}
$$

We have an equation solely in terms of E[X].

$\tfrac{1}{2}$ E[X] = 5, so E[X] = 10.

Practice Problems

Now, we are ready to calculate E[Y].

```
int Far() {
   int a = randomInteger(1, 3); // equally likely to be 1, 2 or 3
   if (a == 1) return 2;
   else if (a == 2) return (2 + Near());
   else return (4 + Far());
}
```

E[Y] = $\sum_{i=1}^{3} P(A = i) E[Y|A = i]$

= ⅓ (2 + E[2 + X] + E[4 + Y])

= ⅓ (2 + 2 + E[X] + 4 + E[Y])

= ⅓ (8 + E[X] + E[Y])

= ⅓ (8 + 10 + E[Y])

= 6 + ⅓E[Y]

Therefore, E[Y] = 9

Practice Problems

```
int Near() {
   int b = randomInteger(1, 4); // equally likely to be 1, 2, 3 or 4
   if (b == 1) return 2;
   else if (b == 2) return 4;
   else if (b == 3) return (6 + Near());
   else return (8 + Near());
}

int Far() {
   int a = randomInteger(1, 3); // equally likely to be 1, 2 or 3
   if (a == 1) return 2;
   else if (a == 2) return (2 + Near());
   else return (4 + Far());
}
```

Let Y = the value returned by `Far()`.

Let X = the value returned by `Near()`.

What is Var[Y]?

Practice Problems

```
int Near() {
   int b = randomInteger(1, 4); // equally likely to be 1, 2, 3 or 4
   if (b == 1) return 2;
   else if (b == 2) return 4;
   else if (b == 3) return (6 + Near());
   else return (8 + Near());
}

int Far() {
   int a = randomInteger(1, 3); // equally likely to be 1, 2 or 3
   if (a == 1) return 2;
   else if (a == 2) return (2 + Near());
   else return (4 + Far());
}
```

Let Y = the value returned by `Far()`.

Let X = the value returned by `Near()`.

What is Var[Y]?

Note: $Var[Y] = E[Y^2] – E[Y]^2$. We know $E[Y]^2$, but we need to solve for $E[Y^2]$.

To find $E[Y^2]$, we need to find $E[X^2]$, as before.

Practice Problems

Let's find $E[X^2]$.

```
int Near() {
   int b = randomInteger(1, 4); // equally likely to be 1, 2, 3 or 4
   if (b == 1) return 2;
   else if (b == 2) return 4;
   else if (b == 3) return (6 + Near());
   else return (8 + Near());
}
```

$E[X^2] = \sum_{i=1}^{4} P(B = i)E[X^2|B = i]$

$= \frac{1}{4} (2^2 + 4^2 + E[(6 + X)^2] + E[(8 + X)^2]$

$= \frac{1}{4} (4 + 16 + 36 + 12E[X] + E[X^2] + 64 + 16E[X] + E[X^2])$

$= \frac{1}{4} (120 + 28E[X] + 2E[X^2])$

$= \frac{1}{4} (120 + 28(10) + 2E[X^2])$

$= \frac{1}{4} (400 + 2E[X^2])$

$= 100 + \frac{1}{2} E[X^2]$

$E[X^2] = 2(100) = 200$

Practice Problems

Now, we find $E[Y^2]$.

```
int Far() {
   int a = randomInteger(1, 3); // equally likely to be 1, 2 or 3
   if (a == 1) return 2;
   else if (a == 2) return (2 + Near());
   else return (4 + Far());
}
```

$E[Y^2] = \sum_{i=1}^{3} P(A = i)E[Y^2|A = i]$

$= ⅓(2^2 + E[(2 + X)^2] + E[(4 + Y)^2]$

$= ⅓(4 + 4 + 4E[X] + E[X^2] + 16 + 8E[Y] + E[Y^2])$

$= ⅓(24 + 40 + E[X^2] + 8(9) + E[Y^2])$

$= ⅓(136 + 200 + E[Y^2])$

$= ⅓(336 + E[Y^2])$

$E[Y^2] = 336/2 = 168$

Practice Problems

Now that we have $E[X^2]$ and $E[Y^2]$, we are ready to calculate $Var[Y]$

$$Var[Y] = E[Y^2] - E[Y]^2 = 168 - (9)^2 = 168 - 81 = 87$$