CSCI 3022-002 Intro to Data Science

What is the **expected value** of the Bernoulli distribution with parameter p?

Announcements and Reminders

probably Thursday

- Practicum posted soon! No HW next Monday!

Last Time...: Expectation

Definition: *Expected Value:*
For a continuous random variable X with pdf $f(x)$, the *expected* value or *mean* value of X is denoted as $E(X)$ and is calculated as:

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

times their probabilities

For a discrete random variable with pmf f, this is

add up outcomes

$$E[X] = \sum_{x \in \Omega} x \cdot P(X = x)$$

f(x)

.

We interpret $E[X]$ as the sample average value of a hypothetical "infinite" sample of the population. Our goal in data science is often to use sample statistics from our limited in size samples to make inferences about underlying population characteristics.

Expected Value of a Function

Last time example R.V. X : (classes)

If a discrete r.v. X has a a density $P(X = x)$, then the expected value of any function $g(X)$ is computed as:

$E[X] = 4.37$

1. Continuous: $\int_{-\infty}^{\infty} g(x)\, f(x)\, dx$

   all outcomes

2. Discrete: $\sum_{x \in \Omega} g(x) \underbrace{f(x)}_{P(X=x)}$

Note that $E[g(X)]$ is computed in the same way that $E(X)$ itself is, except that $g(x)$ is substituted in place of x.

Expected Value of a Function

If a discrete r.v. X has a a density $P(X = x)$, then the expected value of any function $g(X)$ is computed as:

1. Continuous:

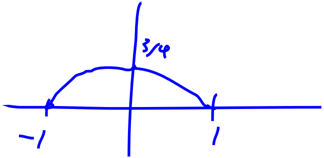
$$E[X] = \int_{\infty}^{\infty} g(x) f(x)\ dx$$

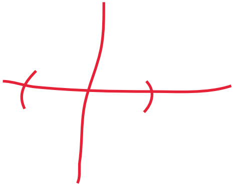
2. Discrete:

$$E[X] = \sum_{x} g(x) f(x)\ dx$$

Note that $E[g(X)]$ is computed in the same way that $E(X)$ itself is, except that $g(x)$ is substituted in place of x.

Expected Value of a Function





Example: A random variable X has pdf:

$$f(x) = \frac{3}{4}(1 - x^2); \quad -1 \leq X \leq 1$$

$3/4(1-x)(1+x)$

What is $E(X^3)$?

$$E(X^3) = \int_{-1}^{1} \underbrace{x^3}_{g(x)=x^3} \cdot \underbrace{\frac{3}{4}(1-x^2)}_{f(x)} \underbrace{dx}_{dx} = \int_{-1}^{1} \underbrace{\frac{3}{4}x^3 - \frac{3}{4}x^5}_{\text{odd integrand}} dx = 0$$

symmetric odd integrand

$$= \frac{3x^4}{4\cdot 4} - \frac{3x^6}{4\cdot 6}\Big|_{-1}^{1}$$

$$= \left(\frac{3}{16} - \frac{3}{24}\right) - \left(\frac{3}{16} - \frac{3}{24}\right) = 0$$

Review: What is $F(x)$?

$$F(x) = \int_{-1}^{x} f(t)\, dt = \frac{3t}{4} - \frac{3t^3}{12}\Big|_{-1}^{x}$$

left

Expected Value of a Function

Example: A random variable X has pdf:

$$f(x) = \frac{3}{4}(1 - x^2); \quad -1 \leq X \leq 1$$

What is $E(X^3)$?

$$E(X^3) = \int_{-1}^{1} x^3 \frac{3}{4}(1 - x^2)\, dx = \frac{3x^4}{16} - \frac{3x^6}{24}|_{-1}^{1} = 0$$

Review: What is $F(x)$?

$$F(x) = \int_{-1}^{x} f(t)\, dt = \frac{3t}{4} - \frac{3t^3}{12}|_{-1}^{x}$$

Expected Value of a Linear Function

multiply by constants
add constants

If $g(X)$ is a linear function of X (i.e., $g(X) = aX + b$) then $E[g(X)]$ can be easily computed from $E(X)$.

$\int (x+2)\,dx \neq \int x\,dx + 2$

Theorem:
Let $a, b \in \mathbb{R}$ and X be a random variable with density f. Then:

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

Proof:

$E[aX+b] = \int_{\text{all values}} (ax+b) f(x)\,dx = \int ax f(x)\,dx + \int b f(x)\,dx$

$= a \underbrace{\int x f(x)\,dx}_{\text{defn } E[X]} + b \underbrace{\int f(x)\,dx}_{\text{integrates to } 1!\ f \text{ is a pdf, } \int \text{over all outcomes}}$

Note: This works for continuous and discrete random variables.

Expected Value of a Linear Function

If $g(X)$ is a linear function of X (i.e., $g(X) = aX + b$) then $E[g(X)]$ can be easily computed from $E(X)$.

Theorem:
Let $a, b \in \mathbb{R}$ and X be a random variable with density f. Then:

$$E[g(X)] = g(E[X])$$

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

Proof:
$E[aX + b] = \int (ax + b) f(x)\, dx = a \int x f(x)\, dx + b \underbrace{\int f(x)\, dx} = aE[X] + b$, since integration is also linear! (and also f is a pdf)
Note: This works for continuous and discrete random variables.

Linear Expectation

Example:

Consider a university having 15,000 students and let X equal the number of courses for which a randomly selected student is registered.

The pdf of X is given to you as follows:

x	1	2	3	4	5	6	7
$f(x) = P(X = x)$	.01,	.03	.13	.25	.39	.17	.02

classes/student

← P(takes that many)

Earlier, we calculated that $E(X)$ was 4.57. If students pay \$500 per course plus a \$100 per-semester registration fee, what is the average amount of money the university can expect a student to pay each a semester?

E[money] where money: 500 · classes + 100

X: classes! money: $g(X) = 500x + 100$

$E[g(x)] = 500 \cdot E[X] + 100 = 500 \cdot 4.57 + 100$

Linear Expectation

Example:
Consider a university having 15,000 students and let X equal the number of courses for which a randomly selected student is registered.
The pdf of X is given to you as follows:

x	1	2	3	4	5	6	7
$f(x) = P(X = x)$	.01,	.03	.13	.25	.39	.17	.02

Earlier, we calculated that $E(X)$ was 4.57. If students pay \$500 per course plus a \$100 per-semester registration fee, what is the average amount of money the university can expect a student to pay each a semester?
$Money = 500 \cdot Courses + 100 = 500X + 100 = g(X)$. Then,

$$E[g(X)] = g(E[X]) = 500 \cdot 4.57 + 100 = 2385.$$

Opening sol:

The pmf of the Bernoulli is given by

$P(\text{outcome})$ outcome

$$f(x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases}$$

$$E[X] = \sum_x x f(x)$$

$$= \underbrace{0 \cdot P(X=0)}_{0} + \underbrace{1 \cdot P(X=1)}_{1}$$

$$= 0(1-p) + 1 \cdot p$$

$$= p.$$

Opening sol:

The pmf of the Bernoulli is given by

$$f(x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases}$$

We now must sum over both outcomes while multiplying by the probability of those outcomes:

$$\sum_{x \in \{0,1\}} x\, f(x) = \sum_{x \in \{0,1\}} x\, P(X = x)$$

$$= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = 0 \cdot (1 - p) + 1 \cdot p = \boxed{p}$$

... which makes perfect sense, since p is how often we "expect" a heads.

Announcements and Reminders

- Practicum 1 tentatively posted tonight?
- Midterm 1 upcoming: takehome!

EV Recap

1. **Expected Value:** The average value for X coming from a distribution (not a sample!). Denoted $E[X]$ or μ or μ_X.
   Discrete: $\sum_{x \in \Omega} x f(x)$; Continuous: $\int_{x \in \Omega} x \cdot f(x)\, dx$

   μ := population average

   how does $\bar{X}$ describe μ_X?
2. Expected value of a function $g(X)$ of X is:
   $\sum_{x \in \Omega} g(x) f(x)$; $\int_{x \in \Omega} g(x) \cdot f(x)\, dx$
3. $Y = g(X)$ is a *change of variables*.
4. Expectation is **linear:** $E[aX + b] = aE[X] + b$ Proof: (2 slides ago.)

EV Recap

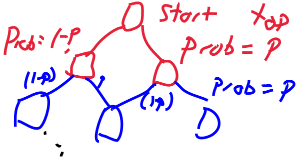
1. **Expected Value:** The average value for X coming from a distribution (not a sample!). Denoted $E[X]$ or μ or μ_X.
   Discrete: $\sum_{x\in\Omega} xf(x)$; Continuous: $\int_{x\in\Omega} x \cdot f(x)\,dx$
2. Expected value of a function $g(X)$ of X is:
   $\sum_{x\in\Omega} g(x)f(x)$; $\int_{x\in\Omega} g(x) \cdot f(x)\,dx$
3. $Y = g(X)$ is a *change of variables*.
4. Expectation is **linear:** $E[aX + b] = aE[X] + b$ Proof:
   $E[aX + b] = \int (ax + b)f(x)\,dx = a\int xf(x)\,dx + b\int f(x)\,dx = aE[X] + b$, since integration is also linear!

Let's play Plinko!!

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.)

https://www.youtube.com/watch?v=naUppHrHJpI

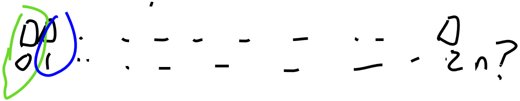
Let's play Plinko!!



Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What *distribution* does X follow?

n

Neg binom



P(end in 2nd from left)
= P(moved R 1 time
& moved L n-1 times)

P(end in bin zero)
= P(every single peg/juncture
we go left)

Let's play Plinko!!

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What *distribution* does X follow?

> Each row results in a move right with probability p, and a move left with probability $1 - p$. We have n rows... or *trials*

Let's play Plinko!!

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What *distribution* does X follow?

Each row results in a move right with probability p, and a move left with probability $1-p$. We have n rows... or *trials*

Each row is a *Bernoulli*, and our ending bucket is the total number of right-hand moves over the entire experiment, or the sum of n Bernoullis!

independence ... Sure, ok

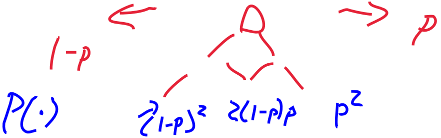
Let's play Plinko!!

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What *distribution* does X follow?

Each row results in a move right with probability p, and a move left with probability $1-p$. We have n rows... or *trials*

For $Y_i \stackrel{iid}{\sim} Bern(p)$, we have $X = \sum Y_i$. So X is a **binomial** with parameters n and p.

Let's play Plinko!!



Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What is the *expected value* of X?

$$E[X] = \underbrace{\sum}_{\text{outcomes}: \{0,1,\dots n\}} x f(x)$$

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\sum x \binom{n}{x} p^x (1-p)^{n-x}$$

intuition: np

$$E_X[n=2]$$

$$0 \cdot P(\cancel{X=0}) + 1\, P(X=1) + 2 \cdot P(X=2)$$

$$\underbrace{2(1-p)p} + 2p^2$$

$$2p - 2p^2 + 2p^2$$

$$= 2p$$:)

Let's play Plinko!!

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What is the *expected value* of X?

We can use our knowledge about Bernoullis! X is a sum of Bernoulli r.v.s that each have mean or expected value of p. $E[Y] = p$ if $Y \sim \mathrm{bern}(p)$.

Let $Y_1, Y_2, Y_3, \ldots, Y_n$ be independent & identical Bernoulli R.V.s w/ probability (parameter) p.

GOAL: $E[Y_1 + Y_2 + Y_3 + \cdots + Y_n]$ but $E[\cdot]$ is linear

$= E[Y_1] + E[Y_2] + E[Y_3] + \cdots E[Y_n].$

Let's play Plinko!!

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What is the *expected value* of X?

We can use our knowledge about Bernoullis! X is a sum of Bernoulli r.v.s that each have mean or expected value of p.

$E[X] = E[Y_1 + Y_2 + Y_3 + \dots Y_n]$, then use linearity:

Let's play Plinko!!

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What is the *expected value* of X?

We can use our knowledge about Bernoullis! X is a sum of Bernoulli r.v.s that each have mean or expected value of p.

$E[X] = E[Y_1 + Y_2 + Y_3 + \ldots Y_n]$, then use linearity:
$E[X] = E[Y_1] + E[Y_2] + E[Y_3] + \cdots + E[Y_n]$. This works even though each Y_i is also a random variable!

Let's play Plinko!!

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What is the *expected value* of X?

We can use our knowledge about Bernoullis! X is a sum of Bernoulli r.v.s that each have mean or expected value of p.

$E[X] = E[Y_1 + Y_2 + Y_3 + \ldots Y_n]$, then use linearity:
$E[X] = E[Y_1] + E[Y_2] + E[Y_3] + \cdots + E[Y_n]$. This works even though each Y_i is also a random variable!

$E[X] = p + p + p + \cdots + p = np$, since each Y is identical.
... which again makes perfect sense, since it's n tries that have a per-try expected value of p.

$$\sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = np.$$

Plinko... is random?

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What is the *Variance* of X?

Plinko... is random?

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What is the *Variance* of X?

Recall: Sample *Variance* is $\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}$

Squared deviations

~~add & average.~~

now: Expected

Plinko... is random?

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What is the *Variance* of X?

Another way: sample variance is $\underbrace{\frac{1}{n-1}\sum_{i=1}^{n}}_{\text{averaged out}} \underbrace{\left(X_i - \bar{X}\right)^2}_{\text{squared deviations}}$

Population variance is this idea expressed as an expectation:

Plinko... is random?

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What is the *Variance* of X?

$\bar{X}$: sample mean

$X_i - \bar{X}$

Another way: sample variance is $\underbrace{\frac{1}{n-1}\sum_{i=1}^{n}}_{\text{averaged out}} \underbrace{(X_i - \bar{X})^2}_{\text{squared deviations}}$

$X - E[X]$

Population variance is this idea expressed as an expectation:

$$Var[X] = E[\underbrace{(X - E[X])^2}_{\text{squared deviations}}] = E[(X - \mu_X)^2]$$

Variance of a Random Variable

recall: Sample var: s^2
" std: s

Greek Signal

σ^2

Definition: *Variance:*

For a discrete random variable X with pdf $f(x)$, the *variance* of X is denoted as σ^2 and is calculated as:

both: $\sigma^2 = E[\underbrace{(X - E[X])^2}_{g(x)}]$

1. Continuous: $\int (X - \underbrace{E[X]}_{\mu_X})^2 f(x)\,dx$

'outcomes

2. Discrete: $\sum (X - \underbrace{E[X]}_{\mu_X})^2 f(x)$

Note: is $E[X]$ random? $\boxed{NO}$

The standard deviation (SD) of X is: $\sqrt{\sigma^2} = \sigma$

Variance of a Random Variable

Definition: *Variance:*

For a discrete random variable X with pdf $f(x)$, the *variance* of X is denoted as $\underline{Var[X] = \sigma^2}$ and is calculated as:

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

1. Continuous:

$$Var[X] = \int_{x \in \Omega} (x - \mu_x)^2 \cdot f(x)\, dx$$

2. Discrete:

$$Var[X] = \sum_{x \in \Omega} (x - \mu_x)^2 f(x)$$

The standard deviation (SD) of X is: $\sigma = \sqrt{\sigma^2}$

Variance Calculated

We want more Plinko! Let's find the variance of a Bernoulli so we can build on it.

Recall: The pmf of the Bernoulli is given by

$$f(x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases}$$

and we know that $E[X] = p$.

Variance Calculated

We want more Plinko! Let's find the variance of a Bernoulli so we can build on it.

Recall: The pmf of the Bernoulli is given by

$$f(x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases}$$

and we know that $E[X] = p$. We now must sum over both outcomes' deviations from the mean while multiplying by those probabilities

$$E[(X - E[X])^2] = \sum_{x \in \{0,1\}} (x-p)^2 f(x) = \sum_{x \in \{0,1\}} (x-p)^2 P(X = x)$$

$$= (0-p)^2 \cdot P(X=0) + (1-p)^2 \cdot P(X=1) = (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p$$

$$= (p)(1-p)(p+1-p) = \boxed{p(1-p)}$$

Let's play Plinko!!

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What is the *variance* of X follow?

Need to know: if two random variables are **independent**,

$$\boxed{Var[X+Y] = Var[X] + Var[Y]}$$

Let's play Plinko!!

Let X be the random variable describing the result in each round of Plinko with n rows and probability p of moving to the right off of each peg. (Ignoring the edges for now.) What is the *variance* of X follow?

Need to know: if two random variables are **independent**,

$\boxed{Var[X+Y] = Var[X] + Var[Y]}$

So for Plinko, where $X = Y_1 + Y_2 + \cdots + Y_i$ but the $Y_i's$ are all independent,

$$Var[X] = Var[\sum Y_i] \underbrace{=}_{\text{indep}} \sum Var[Y] \underbrace{=}_{\text{ident}} nVar[Y_i] = \boxed{np(1-p)}$$

Sanity Check! Should variance be smaller if $p \approx 1$ or $p \approx 0$?

Let's talk Variance

For a random variable X and constants a and b, if we define $Y = aX + b$...
$E[Y] = aE[X] + b$ because Expectation $E[\cdot]$ is **linear**. Is $Var[\cdot]$?

1. What is $Var[X + b]$?

2. What is $Var[aX]$?

Let's talk Variance

For a random variable X and constants a and b, if we define $Y = aX + b$...
$E[Y] = aE[X] + b$ because Expectation $E[\cdot]$ is **linear**. Is $Var[\cdot]$?

1. What is $Var[X + b]$?
   Intuition: moving X doesn't change its spread!

2. What is $Var[aX]$?
   Intuition: multiplying X should change its spread!

Non-linear Variance

For a random variable X and constants a and b, if we define $Y = aX + b$...

What is $Var[aX + b]$?

Non-linear Variance

For a random variable X and constants a and b, if we define $Y = aX + b$...

What is $Var[aX + b]$?

$$\begin{aligned}
Var[aX + b] &= \sum_{x \in \Omega} (aX + b - E[aX + b])^2 f(x) \\
&= \sum_{x \in \Omega} (aX + b - aE[X] - b)^2 f(x) \\
&= \sum_{x \in \Omega} (aX - aE[X])^2 f(x) \\
&= \sum_{x \in \Omega} a^2 (X - E[X])^2 f(x) \\
&= a^2 \sum_{x \in \Omega} (X - E[X])^2 f(x) \\
&= \boxed{a^2 Var[X]}
\end{aligned}$$

Daily Recap

Today we learned

1. Variance

Moving forward:

- nb day Friday!

Next time in lecture:

- Wrap-up and some more examples on populations.