CSCI 3022 Intro to Data Science
Variance

We begin lecture with a an overarching question. *What are the odds*? Take a peek at this video, in which all of a man's dreams come true.

At some point we come back down from our euphoric high and decide to ask a mathier question:

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

Announcements and Reminders

- HW 4 extended one day (due tomorrow, not tonight)
- Exam posted tomorrow
- Practicum posted laster this week!

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:

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

For a discrete random variable with pmf f, this is

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

.

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

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:

2. Discrete:

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:

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

2. Discrete:

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

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.

Expectation and Spread

The idea of **Expected value** can be extended to describe all kind of notions of "what should happen if we have a (arbitrarily large) sample.

> Suppose we wish to know the variance or standard deviation of the population. For a *sample*, recall that
>
> $$s = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n - 1}$$

Expectation and Spread

The idea of **Expected value** can be extended to describe all kind of notions of "what should happen if we have a (arbitrarily large) sample.

Suppose we wish to know the variance or standard deviation of the population. For a *sample*, recall that

$$s = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n-1}$$

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

Expectation and Spread

The idea of **Expected value** can be extended to describe all kind of notions of "what should happen if we have a (arbitrarily large) sample.

Suppose we wish to know the variance or standard deviation of the population. For a *sample*, recall that

$$s = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n-1}$$

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

We might ask: what is the *expected* value of how spread out x-value are?

Population variance is this idea expressed as an expectation:

Expectation and Spread

The idea of **Expected value** can be extended to describe all kind of notions of "what should happen if we have a (arbitrarily large) sample.

Suppose we wish to know the variance or standard deviation of the population. For a *sample*, recall that

$$s = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n-1}$$

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

Population variance is this idea expressed as an expectation:

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

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

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: $\int_{x \in \Omega} x \cdot f(x)\, dx$
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:
   $E[aX + b] = \int (ax + b) f(x)\, dx = a \int x f(x)\, dx + b \int f(x)\, dx = aE[X] + b$, 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?

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!

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?

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.

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.

$E[X] = E[Y_1 + Y_2 + Y_3 + \dots Y_n]$, 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.

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 $\dfrac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}$

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?

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:

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

Variance of a Random Variable

Definition: *Variance:*

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

1. Continuous:

2. Discrete:

The standard deviation (SD) of X is:

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 $Var[aX + b]$?

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

Calculating Variance

When tasked with computing Variance sums/integrals, it is often a little tedious to compute

$$Var[x] = \sum_x (x - E[x]) f(x) \quad \text{or} \quad \sum_x \int (x - E[x]) f(x) \, dx$$

.

Calculating Variance

When tasked with computing Variance sums/integrals, it is often a little tedious to compute

$$Var[x] = \quad \sum_x (x - E[x])f(x) \quad \text{or} \quad \sum_x \int (x - E[x])f(x)\,dx$$

.

Important Formula: $Var[X] = E[X^2] - E[X]^2$
Proof:

Calculating Variance

When tasked with computing Variance sums/integrals, it is often a little tedious to compute

$$Var[x] = \sum_x (x - E[x])f(x) \quad \text{or} \quad \sum_x \int (x - E[x])f(x)\,dx$$

.

Important Formula: $Var[X] = E[X^2] - E[X]^2$
Proof:

$$Var[X] = E[(X - E[X])^2] \overbrace{=}^{foil} E[X^2 - 2XE[X] + E[X]^2]$$

$$\overbrace{=}^{linear} E[X^2] - E[2XE[X]] + E[E[X]^2]$$

$$\overbrace{=}^{non-random} E[X^2] - 2E[X]E[X] + E[X]^2 \overbrace{=}^{simplify} E[X^2] - E[X]^2$$

Calculating Variance

This can help a lot! Note that

$$E[X^2] = \sum x^2 f(x) \quad \text{and} \quad \sum_x \int x^2 f(x)\, dx$$

look like a very similar mechanical computations to

$$E[X] = \sum x f(x) \quad \text{and} \quad \sum_x \int x^2 f(x)\, dx$$

, so we can reuse a lot of work, as we'll always compute $E[x]$ before $Var[X]$ either way!
Important Formula: $Var[X] = E[X^2] - E[X]^2$

Really non-linear Variance

What if we want to know what happens to two events that *aren't* independent? For example, what's the variance of $Z = X + Y$?

Really non-linear Variance

What if we want to know what happens to two events that *aren't* independent? For example, what's the variance of $Z = X + Y$?

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

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

Really non-linear Variance

What if we want to know what happens to two events that *aren't* independent? For example, what's the variance of $Z = X + Y$?

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

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

If we expand this out, we have to deal with a bunch of $XY, XE[Y]$, etc. terms. It matters if X and Y move *together*. It helps to define this concept. What does it mean for X and Y to move *together*?

Example: what if $Y = -X$? Then the variance of Z is zero!

Covariance

When two random variables X and Y are not independent, it is frequently of interest to assess how strongly they are related to one another.

Definition: *Covariance:*

The covariance between two rv's X and Y is defined as:

$$E[\ \underbrace{(X-\mu_X)}_{\text{X versus its mean}}\ \overbrace{(Y-\mu_Y)}^{\text{Y versus its mean}}\ ]$$

If both variables tend to deviate in the same direction (both go above their means or below their means at the same time), then the covariance will be positive.

If the opposite is true, the covariance will be negative.

If X and Y are not strongly related, the covariance will be near 0.

Correlation

Definition: *Correlation*
The *correlation* coefficient of X and Y, denoted by __________ or just _, is the *unitless* measure of covariance defined by:

It represents a "scaled" covariance: correlation ranges between -1 and 1.

Correlation

Definition: *Correlation*
The *correlation* coefficient of X and Y, denoted by $\underline{Cov[X,Y]}$ or just $\underline{\rho}$, is the *unitless* measure of covariance defined by:

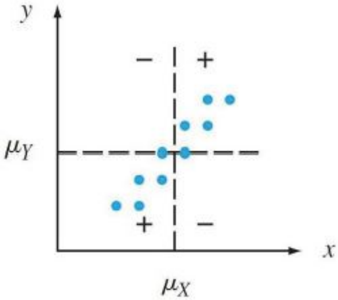
$$\rho = \frac{Cov[X,Y]}{\sigma_X \sigma_Y}$$

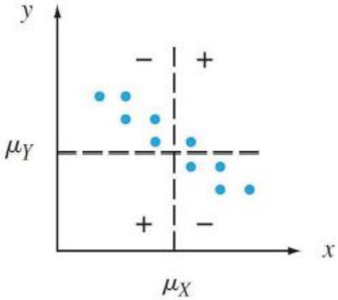
It represents a "scaled" covariance: correlation ranges between -1 and 1.

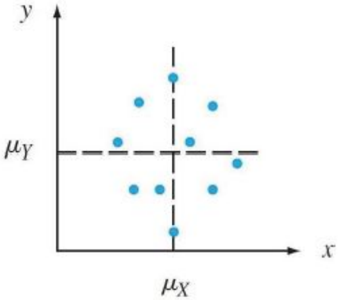
Covariance Pictured

The covariance depends on both the set of possible pairs and the probabilities of those pairs.

Below are examples of 3 types of "co-varying":







Interpreting Correlation

If X and Y are independent, then _____, but _____ does not imply independence.

The correlation coefficient is a measure of the *linear relationship* between X and Y, and only when the two variables are perfectly related in a *linear* manner will be as positive or negative as it can be.

Two variables could be uncorrelated yet highly dependent because there is a strong nonlinear relationship, so be careful not to conclude too much from low correlation. (e.g. $y = x^2$)

We return to covariance in a few weeks...

Interpreting Correlation

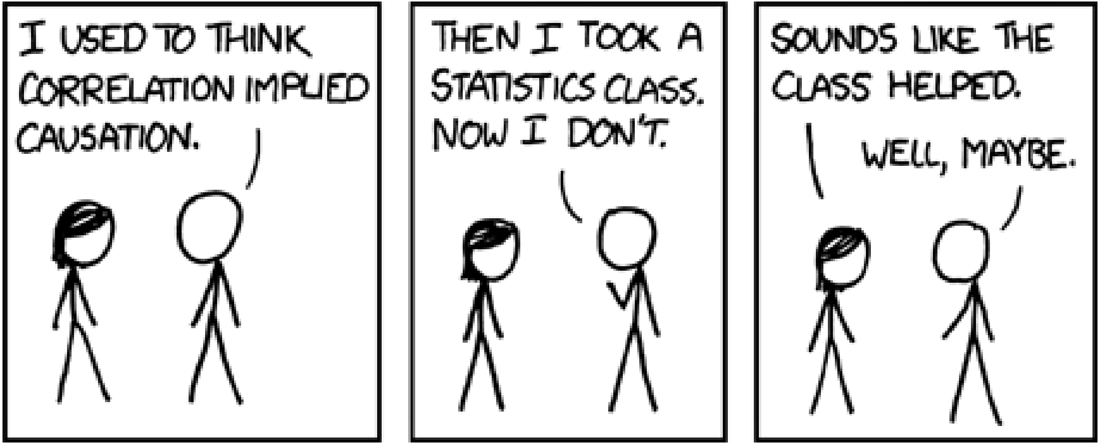
If X and Y are independent, then $\underline{\rho = 0}$, but $\underline{\rho = 0}$ does not imply independence.

The correlation coefficient is a measure of the *linear relationship* between X and Y, and only when the two variables are perfectly related in a *linear* manner will be as positive or negative as it can be.

Two variables could be uncorrelated yet highly dependent because there is a strong nonlinear relationship, so be careful not to conclude too much from low correlation. (e.g. $y = x^2$)

We return to covariance in a few weeks...

Interpreting Correlation



Daily Recap

Today we learned

1. Variance

Moving forward:

- nb day Friday!

Next time in lecture:

- Wrap-up and some more examples on populations.