Confidence Intervals & Joint Distributions

CSE 312 Summer 25
Lecture 17

Why Learn Normals?

When we add together independent normal random variables, you get another normal random variable.

The sum of **any** independent random variables **approaches** a normal distribution.

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables, with mean μ and variance σ^2. Let $Y_n = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$

As $n \to \infty$, the CDF of Y_n converges to the CDF of $\mathcal{N}(0, 1)$

Approximating a continuous distribution

You buy lightbulbs that burn out according to an exponential distribution with parameter of $\lambda = 1.8$ lightbulbs per year.

You buy a 10 pack of (independent) light bulbs. What is the probability that your 10-pack lasts at least 5 years?

Let X_i be the time it takes for lightbulb i to burn out.

Let X be the total time. Estimate $\mathbb{P}(X \geq 5)$.

$$E[X_i] = \frac{1}{\lambda} = \frac{1}{1.8}$$

$$Var(X_i) = \frac{1}{\lambda^2} = \frac{1}{(1.8)^2}$$

$$X = \sum_{i=1}^{10} X_i \qquad E[X] = \frac{10}{1.8} \qquad Var(X) = \frac{10}{1.8^2}$$

Where's the continuity correction?

There's no correction to make – it was already continuous!!

$$\mathbb{P}(X \geq 5)$$

$$= \mathbb{P}\left(\frac{X - 10/1.8}{\sqrt{10/1.8^2}} \geq \frac{5 - 10/1.8}{\sqrt{10/1.8^2}}\right)$$

$$\approx \mathbb{P}\left(Y \geq \frac{5 - 10/1.8}{\sqrt{10/1.8^2}}\right) \text{ By CLT}$$

$$\approx \mathbb{P}(Y \geq -0.32)$$

$$= 1 - \Phi(-0.32) = \Phi(0.32)$$

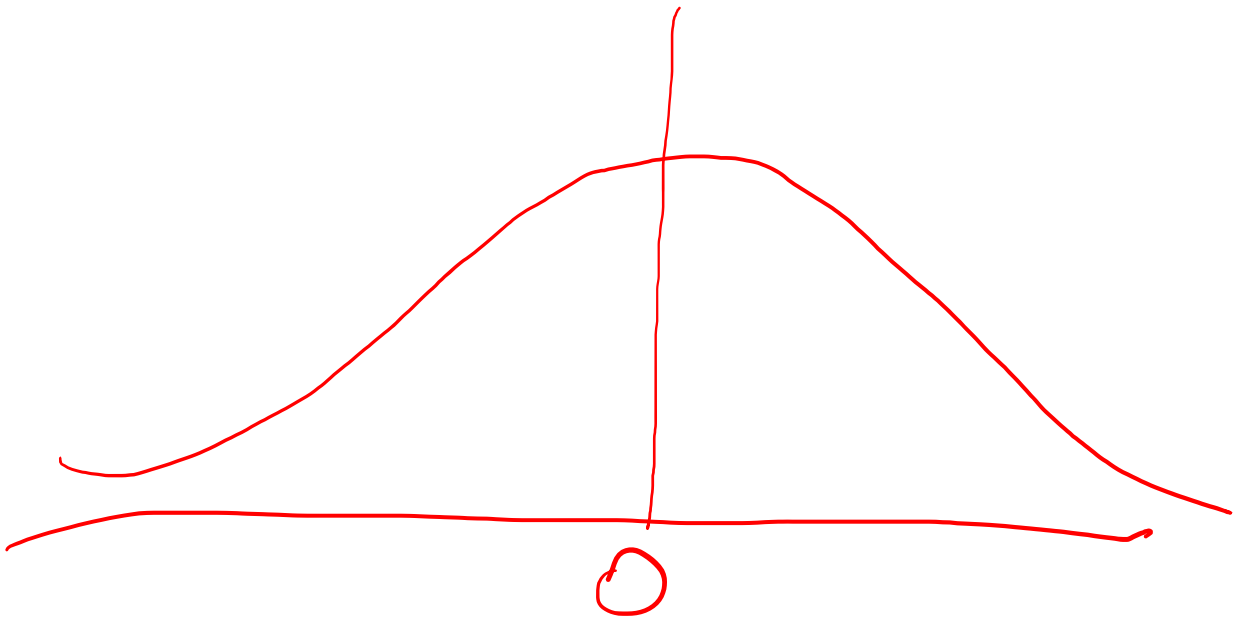
$$\approx .62552$$

$Y \sim N(0, 1)$

True value (needs a distribution not in our zoo) is ≈ 0.58741

Outline of CLT steps

1. Write event you are interested in, in terms of sum of random variables.
2. Apply continuity correction if RVs are discrete.
3. Standardize RV to have mean 0 and standard deviation 1.
4. Replace RV with $\mathcal{N}(0,1)$.
5. Write event in terms of Φ
6. Look up in table.



Process For Continuity Correction

Let X be the discrete random variable you are approximating with Y.

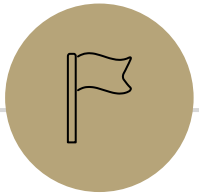
To do a continuity correction, find all real numbers that, when rounded to nearest value in Ω_X, would be part of the event.

For example, if $X \sim \text{Bin}(n, p)$, $\Omega_X = \{0, 1, \ldots, n\}$

To find event $\mathbb{P}(X \geq 6)$, 5.5 rounds to 6, which is ≥ 6. 5.4 rounds to 5 not ≥ 6. Want $\mathbb{P}(X \geq 5.5)$

To find event $\mathbb{P}(X > 6)$ 5.5 rounds to 6, which is not >6, 6.1 rounds to 6 which is not > 6, 6.5 rounds to 7; Want $\mathbb{P}(X \geq 6.5)$

To find event $\mathbb{P}(X = 5)$, 4.5 rounds to 5, 5.4 rounds to 5, 4.4 rounds to 4. Want $\mathbb{P}(4.5 \leq X < 5.5)$



Confidence Intervals

Confidence Intervals

A "confidence interval" tells you the probability (how confident you should be) that your random variable fell in a certain range (interval)

Usually "close to its expected value"

$$\mathbb{P}(|X - \mu| > \varepsilon) \leq \delta$$

If your RV has expectation equal to the value you're searching for (like our polling example) you get a probability of being "close enough" to the target value.

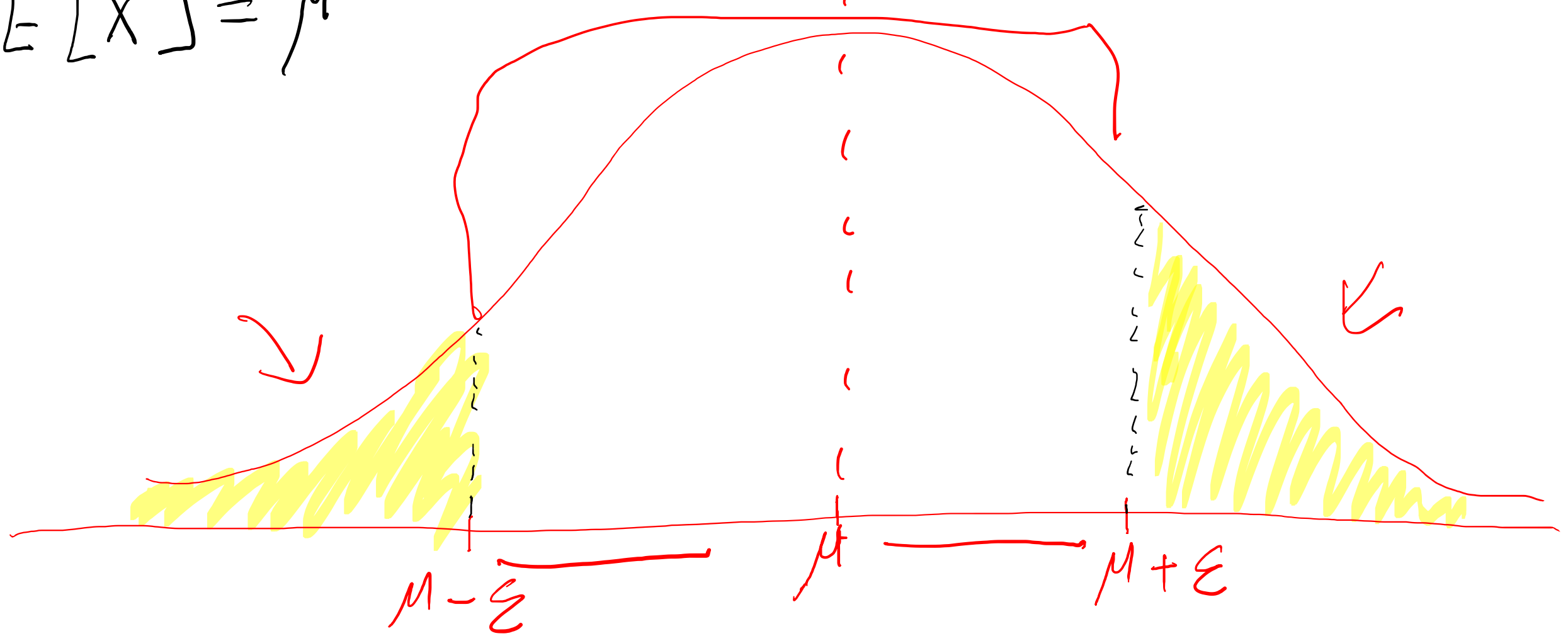
$$\mathbb{P}(|X - \mu| > \varepsilon) \leq \delta$$

5%

2%

$$P(\quad) \leq \delta$$

$$E[X] = \mu$$



Confidence Intervals

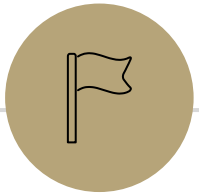
Using the CLT, we estimated the probability of "missing low"

There's a few drawbacks though

1. Using the CLT we get an estimate, not a guarantee---what if the CLT estimate is underestimating the probability of failure?

2. We needed to know the true value to do that computation---if we knew the true value, we wouldn't run the poll!

Some algebra tricks can handle problem 2, but 1 really asks for a new tool; we'll see concentration inequalities later this week.



Application: Idealized Polling

Polling

Our end goal is to answer the question “how many people do I need to poll to get an accurate sense of how the population is going to vote?”

That’s a weird question (it’ll require “going backwards” in the algebra) so first we’ll “go forwards” (given the poll size how accurate will we be?) to see what’s happening more clearly.

Polling

Suppose you know that 60% of CSE students support you in your run for SAC. If you draw a sample of 30 students, what is the probability that you don't get a majority of their votes.

How are you sampling?

Method 1: Get a uniformly random subset of size 30.

Method 2: Independently draw 30 people with replacement.

Which do we use?

Polling

$X_i = \begin{cases} 1 & \text{if support} \\ 0 & \text{if not support} \end{cases}$

Let X_i be the indicator for "person i in the sample supports you."

$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{30}$ is the fraction who support you.

We're interested in the event $\mathbb{P}(\bar{X} \leq .5)$.

What is $\mathbb{E}[\bar{X}]$? What is $\mathbf{Var}(\bar{X})$?

$E[\bar{X}] = E\left[\frac{\sum_{i=1}^{n} X_i}{30}\right]$

Polling

Let X_i be the indicator for "person i in the sample supports you."

$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{30}$ is the fraction who support you.

We're interested in the event $\mathbb{P}(\bar{X} \leq .5)$.

What is $\mathbb{E}[\bar{X}]$? What is $\mathrm{Var}(\bar{X})$?

$$\mathbb{E}[\bar{X}] = \frac{1}{30}\mathbb{E}[\sum X_i] = \frac{.6 \cdot 30}{30} = \frac{3}{5}$$

$$\mathrm{Var}(\bar{X}) = \frac{1}{30^2}\mathrm{Var}(\sum X_i) = \frac{1}{30} \cdot .6 \cdot .4 = \frac{1}{125}$$

Using the CLT

$$\mathbb{P}(\bar{X} \leq .5)$$

$$= \mathbb{P}\left(\frac{\bar{X}-.6}{1/\sqrt{125}} \leq \frac{.5-.6}{1/\sqrt{125}}\right)$$

$$\approx \mathbb{P}\left(Y \leq \frac{.5-.6}{1/\sqrt{125}}\right) \text{ where } Y \sim \mathcal{N}(0,1)$$

CLT

$$\approx \mathbb{P}(Y \leq -1.12)$$

$$= \Phi(-1.12) = 1 - \Phi(1.12) \approx 1 - 0.86864 = 0.13136$$

Hey! Where's the continuity correction?

If this were just a question about $n = 30$, we would have used one. But for preparing for the next calculation it made sense to skip it.

What is $\bar{X}$?

It's the *average* of a bunch of indicators.

So the support is:

$\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \dots, \frac{n-1}{n}, \frac{n}{n}$.

Instead of $.5$, we'd use $.5 + \frac{1}{2n}$. Which makes the algebra much worse.

And for real polling applications, n is going to be quite big anyway where $\frac{1}{2n}$ is not going to make a substantial difference.

The Reverse Question

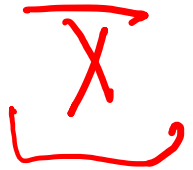
Polls are made by sampling n people from a population. They are then reported with "52% of likely voters would vote in favor of proposal if held today (margin of error +/- 3%)"

You are going to run your own poll. And you want a better "margin of error" – you want 2% how many people do you need to poll?

Let's think about idealized polling – pretend we're really getting a uniformly random person.

Margin of Error

Wait…what's a "margin of error"



The result of the poll is a random variable – it has a distribution.

You'd like to know something about its variance (Did you poll everyone in the entire country? Just 3 people? How much variance is there in the poll?)

A "margin of error" is an intuitive measurement of the variance of the poll. "If I performed this poll repeatedly, 95% of the time, we're within true +/- the margin of error."

Our Goal

Set a target – I want my margin of error to be 2%. That is, at least 95% of the time, your poll's estimate of the fraction of people in favor will be within 2 percentage points of the true value.

So…how many people are you going to need to interview?

Poll Setup

Let X_i be the indicator that the i^{th} person you interview supports the proposal.

Your random variable is $\hat{p}$: $\sum X_i / n$

Let p be the true fraction of people who support the proposal.

What is the

$\mathbb{E}[\hat{p}] =$ $E\left[\frac{\sum X_i}{n}\right] = \frac{1}{n} E\left[\sum X_i\right] = p$

$\text{Var}(\hat{p}) =$

Poll Setup

Let X_i be the indicator that the i^{th} person you interview supports the proposal.

Your random variable is $\hat{p}$: $\sum X_i / n$

Let p be the true fraction of people who support the proposal.

What is the

$$\mathbb{E}[\hat{p}] = \frac{1}{n} \cdot \mathbb{E}[\sum X_i] = \frac{pn}{n} = p$$

$$\text{Var}(\hat{p}) = \frac{1}{n^2} \text{Var}(X_i) = \frac{p(1-p)}{n}$$

Using the CLT

What are we looking for? Well we have a margin of error:

$$\mathbb{P}(p - .02 \leq \hat{p} \leq p + .02) \geq .95$$

That says we're within the 2% margin of error at least 95% of the time.

What is that probability? Well let's setup to use the CLT. Subtract the expectation and divide by the standard deviation.

$$\mathbb{P}\left(\frac{p-.02-p}{\sqrt{p(1-p)/n}} \leq \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq \frac{p+.02-p}{\sqrt{p(1-p)/n}}\right) \geq .95$$

Apply the CLT

$$\mathbb{P}\left(\frac{p-.02-p}{\sqrt{p(1-p)/n}} \leq \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq \frac{p+.02-p}{\sqrt{p(1-p)/n}}\right) \geq .95$$

Is well approximated by $\mathbb{P}\left(\frac{-\sqrt{n}\cdot.02}{\sqrt{p(1-p)}} \leq Z \leq \frac{\sqrt{n}\cdot.02}{\sqrt{p(1-p)}}\right) \geq .95$ for $Z \sim \mathcal{N}(0,1)$

So as n changes, the probability changes. So choose the smallest n for which the probability is at least $.95$

WAIT, what's $\sqrt{p(1-p)}$? We don't know p. That's *why* we're doing the poll in the first place.

Handling $\sqrt{p(1-p)}$

Justification 1: If we make a mistake, we want it to be making n bigger. (since we're trying to say "take n at least this big, and you'll be safe").

The bigger the standard deviation, the bigger n will need to be to control it. So assume the biggest possible standard deviation.

Justification 2:

As $\sqrt{p(1-p)}$ gets bigger, the interval gets smaller (it's in the denominator), so assuming the biggest value of $\sqrt{p(1-p)}$ gives us the most restricted interval. So no matter what the true interval is we have a subset of it. And if our probability is at least $.95$ then the true probability is at least $.95$.

What's the maximum of $\sqrt{p(1-p)}$?

Worst value of p

Calculus time!

Set $\frac{d}{dp}\sqrt{p-p^2}=0$

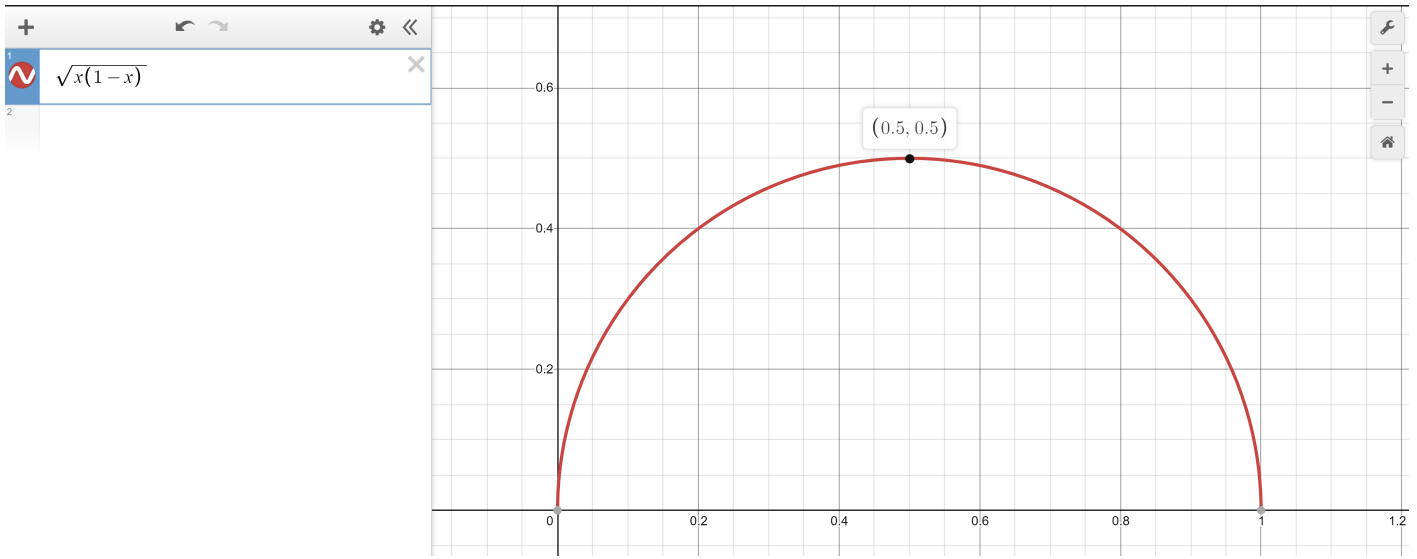
$\frac{1}{\sqrt{p-p^2}}(1-2p)=0$

$1-2p=0 \rightarrow p=1/2$

Second derivative test will confirm $p=\frac{1}{2}$ is a maximizer

Or just plot it.

$\sqrt{\frac{1}{2}\left(1-\frac{1}{2}\right)}=\sqrt{1/4}.$



Doing the algebra

$$\mathbb{P}\left(\frac{p-.02-p}{\sqrt{p(1-p)/n}} \le \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \le \frac{p+.02-p}{\sqrt{p(1-p)/n}}\right)$$

$$\approx \mathbb{P}\left(\frac{-\sqrt{n}\cdot.02}{\sqrt{p(1-p)}} \le Z \le \frac{\sqrt{n}\cdot.02}{\sqrt{p(1-p)}}\right) \text{ by CLT; } Z \sim \mathcal{N}(0,1)$$

$$\ge \mathbb{P}\left(\frac{-\sqrt{n}\cdot.02}{\sqrt{1/4}} \le Z \le \frac{\sqrt{n}\cdot.02}{\sqrt{1/4}}\right)$$

$$= \mathbb{P}(-.04\sqrt{n} \le Z \le .04\sqrt{n})$$

$$= \Phi(.04\sqrt{n}) - \left(1 - \Phi(.04\sqrt{n})\right) = 2\Phi(.04\sqrt{n}) - 1$$

$$2\Phi(.04\sqrt{n}) - 1 \ge .95 \rightarrow \Phi(.04\sqrt{n}) \ge \frac{1.95}{2}$$

Using the Φ-table

$$\Phi(.04\sqrt{n}) \geq .975$$

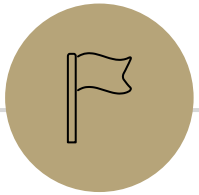
Φ-table says:

$$.04\sqrt{n} \geq 1.96$$

$$\sqrt{n} \geq 49$$

$n \geq 2401$. gives 95% confidence interval of +/- 2%.

I.e. 95% of the time, our poll gets a value within 2% of the true value.



Multiple Random Variables

This part of lecture and next lecture

Somewhat out-of-place content.

When we introduced multiple random variables, we've always had them be independent.

Because it's hard to deal with non-independent random variables.

Today and Wednesday are a crash-course in the toolkit for when you have multiple random variables and they aren't independent.

Going to focus on discrete RVs, we'll talk about continuous at the end.

Joint PMF, support

For two (discrete) random variables X, Y their joint pmf

$$p_{X,Y}(x, y) = \mathbb{P}(X = x \cap Y = y)$$

When X, Y are independent then $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.

Examples

Roll a blue die and a red die. Each die is 4-sided. Let X be the blue die's result and Y be the red die's result.
Each die (individually) is fair. But not all results are equally likely when looking at them both together.

$p_{X,Y}(1,2) = 3/16.$

$p_{X,Y}$	X=1	X=2	X=3	X=4
Y=1	1/16	1/16	1/16	1/16
Y=2	3/16	0	0	1/16
Y=3	0	2/16	0	2/16
Y=4	0	1/16	3/16	0
	4/16	4/16	4/16	4/16

Marginals

What if I just want to talk about X?

Well, use the law of total probability:

$\mathbb{P}(X = k) = \sum_{\text{partition } \{E_i\}} \mathbb{P}(X = k | E_i)\mathbb{P}(E_i)$

and use E_i to be possible outcomes for Y For the dice example

$\mathbb{P}(X = k) = \sum_{\ell=1}^{4} \mathbb{P}(X = k \,|Y = \ell)\mathbb{P}(Y = \ell)$

$= \sum_{\ell=1}^{4} \mathbb{P}(X = k \cap Y = \ell)$

$p_X(k) = \sum_{\ell=1}^{4} p_{X,Y}(k, \ell)$

$p_X(k)$ is called the "marginal" distribution for X (we "marginalized out" Y) it's the same pmf we've always used; the name comes from being in the margin of the paper when people printed these on paper.

Marginals

$$p_X(k) = \sum_{\ell=1}^{4} p_{X,Y}(k,\ell)$$

So

$$p_X(2) = \frac{1}{16} + 0 + \frac{2}{16} + \frac{1}{16} = \frac{4}{16}$$

$p_{X,Y}$	X=1	X=2	X=3	X=4
Y=1	1/16	1/16	1/16	1/16
Y=2	3/16	0	0	1/16
Y=3	0	2/16	0	2/16
Y=4	0	1/16	3/16	0

Different dice

(1, 2) $U = 1$, $V = 2$

Roll two fair dice independently. Let U be the minimum of the two rolls and V be the maximum

Are U and V independent?

Write the joint distribution in the table

What's $p_U(z)$? (the marginal for U)

$p_{U,V}$	U=1	U=2	U=3	U=4
V=1	1/16	0	0	0
V=2	2/16	1/16	0	0
V=3			1/16	0
V=4				1/16

Different dice

Roll two fair dice independently.
Let U be the minimum of the two rolls and V be the maximum

$$p_U(z) = \begin{cases} \frac{7}{16} & \text{if } z = 1 \\ \frac{5}{16} & \text{if } z = 2 \\ \frac{3}{16} & \text{if } z = 3 \\ \frac{1}{16} & \text{if } z = 4 \\ 0 & \text{otherwise} \end{cases}$$

$p_{U,V}$	U=1	U=2	U=3	U=4
V=1	1/16	0	0	0
V=2	2/16	1/16	0	0
V=3	2/16	2/16	1/16	0
V=4	2/16	2/16	2/16	1/16

Joint Expectation

Expectations of joint functions

For a function $g(X, Y)$, the expectation can be written in terms of the joint pmf.

$$\mathbb{E}[g(X,Y)] = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} g(x,y) \cdot p_{X,Y}(x,y)$$

This definition hopefully isn't surprising at this point (it's the value of g times the probability g takes on that value), but it's good to see.

Expectation of a function of two RVs

What's $\mathbb{E}[UV]$ for U, V from the last slide?

Expectation of a function of two RVs

What's $\mathbb{E}[UV]$ for U, V from the last slide?

$$\sum_{u\in\Omega_U}\sum_{v\in\Omega_V} uv \cdot p_{U,V}(u,v)$$

$$= 1\cdot 1\cdot\frac{1}{16} + 1\cdot 2\cdot\frac{2}{16} + 1\cdot 3\cdot\frac{2}{16} + 2\cdot 2\cdot\frac{1}{16} + 2\cdot 3\cdot\frac{2}{16} + 2\cdot 4\cdot\frac{2}{16} +$$

$$3\cdot 3\cdot\frac{1}{16} + 3\cdot 4\cdot\frac{2}{16} + 4\cdot 4\cdot\frac{1}{16}$$

$$= \frac{92}{16} = \frac{23}{4} = 5.75.$$