Quiz 3 info posted on webpage

↳ fill out conflict form, if needed

Don't forget HW6 is out

↳ more problems than usual are autograded

↳ due Wed.

Wrap CLT

CSE 312 Winter 26
Lecture 18

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 $X_1, X_2, \dots, X_n$ 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 $\mathcal{N}(0,1)$

Theory vs. Practice

The formal theorem statement is “in the limit”

You might not get exactly a normal distribution for any finite n (e.g. if you sum indicators, your random variable is always discrete and will be discontinuous for every finite n.

In practice, the approximations get very accurate very quickly (at least with a few tricks we’ll see soon).

They won’t be exact (unless the X_i are normals) but it’s close enough to use even with relatively small n.

Using the Central Limit Theorem

Suppose you are managing a factory, that produces widgets. Each widget produced is defective (independently) with probability 5%.

Your factory will produce 1000 (possibly defective) widgets. You want to know what the chances are of having a "very bad day" where "very bad" means producing at most 940 non-defective widgets.
(In expectation, you produce 950 non-defective widgets)

What is the probability?

True Answer

Let $X \sim \text{Bin}(1000, .95)$

What is $\mathbb{P}(X \leq 940)$?

The cdf is ugly…and that's a big summation.

$\sum_{k=0}^{940} \binom{1000}{k} (.95)^k \cdot (.05)^{1000-k} \approx .08673$

What does the CLT give?

CLT setup

Bin(1000,.95) is the sum of a bunch of independent random variables (the indicators/Bernoullis we summed to get the binomial)

$\sigma\sqrt{n} = \sqrt{\sigma^2 \cdot n}$

So, let's use the CLT instead

$\mathbb{E}[X_i] = p = .95.$ μ

$\mathrm{Var}(X_i) = p(1-p) = .0475$ σ^2

$Y_{1000} = \frac{\sum_{i=1}^{1000} X_i - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}$ is approximately $\mathcal{N}(0,1)$.

With the CLT.

The event we're interested in is $\mathbb{P}(X \leq 940)$

$\mathbb{P}(X \leq 940)$

$= \mathbb{P}\left(\frac{X - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}} \leq \frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right)$

$= \mathbb{P}\left(Y_{1000} \leq \frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right)$

$Y \sim \mathcal{N}(0,1)$

$\approx \mathbb{P}\left(Y \leq \frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right)$ *by CLT (where Y is a standard normal)*

$= \Phi\left(\frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right)$

$\approx \Phi(-1.45) = 1 - \Phi(1.45)$

$\approx 1 - .92647 = .07353.$

It’s an approximation!

The true probability is

$$1 - \sum_{k=941}^{1000} \binom{1000}{k} (.95)^k \cdot (.05)^{1000-k} \approx .08673$$

The CLT estimate is off by about 1.3 percentage points.

We can get a better estimate if we fix a subtle issue with this approximation.

A problem

What's the probability that $\mathrm{X} = 950$? (exactly)

True value, we can get with binomial:

$\binom{1000}{950} \cdot (.95)^{950} \cdot (.05)^{50} \approx .05779$

What does the CLT say?

$\mathbb{E}[X_i] = .95$

$Var(X_i) = (.95)(.05) = .0475$

A problem

What's the probability that $X = 950$? (exactly)

True value, we can get with binomial:

$\binom{1000}{950} \cdot (.95)^{950} \cdot (.05)^{50} \approx .05779$

What does the CLT say?

$= \mathbb{P}\left(\frac{X-1000\cdot.95}{\sqrt{1000\cdot.0475}} = \frac{950-1000\cdot.95}{\sqrt{1000\cdot.0475}}\right)$

$\approx \mathbb{P}(Y = 0)$ $Y \sim \mathcal{N}(0,1)$

$= 0$

Uh oh.

Continuity Correction

The binomial distribution is discrete, but the normal is continuous.

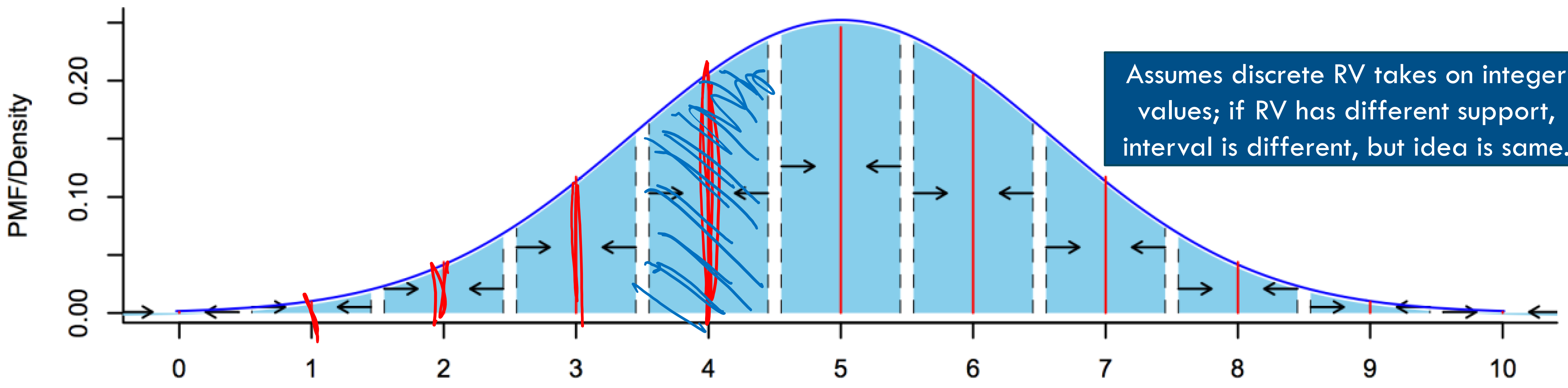
Let's correct for that (called a "continuity correction")

Before we switch from the binomial to the normal, ask "what values of a continuous random variable would round to this event?"

Solution – Continuity Correction

Slide from 24AU 312 (Rao/Tessaro)

Probability estimate for i: Probability for all x that round to i!



Assumes discrete RV takes on integer values; if RV has different support, interval is different, but idea is same.

To estimate probability that discrete RV lands in (integer) interval $\{a, \ldots, b\}$, compute probability continuous approximation lands in interval $[a - \frac{1}{2}, b + \frac{1}{2}]$

Applying the continuity correction

$$\mathbb{P}(X = 950)$$

$$= \mathbb{P}(949.5 \leq X < 950.5)$$

$$= \mathbb{P}\left(\frac{949.5-950}{\sqrt{1000\cdot.0475}} \leq \frac{X-950}{\sqrt{1000\cdot.0475}} < \frac{950.5-950}{\sqrt{1000\cdot.0475}}\right)$$

$$\approx \mathbb{P}\left(\frac{949.5-950}{\sqrt{1000\cdot.0475}} \leq Y < \frac{950.5-950}{\sqrt{1000\cdot.0475}}\right)$$ By CLT ($Y \sim \mathcal{N}(0,1)$)

$$= \Phi\left(\frac{950.5-950}{\sqrt{1000\cdot.0475}}\right) - \Phi\left(\frac{949.5-950}{\sqrt{1000\cdot.0475}}\right)$$

$$\approx \Phi(0.07) - \Phi(-0.07) = \Phi(0.07) - (1 - \Phi(0.07))$$

$$\approx 0.5279 - (1 - 0.5279) = 0.0558$$

Continuity correction.
This step really is an "exactly equal to"
The discrete rv X can't equal 950.2.

Still an Approximation

$\binom{1000}{950} \cdot (.95)^{950} \cdot (.05)^{50} \approx .05779$ is the true value

The CLT approximates: 0.0558

Very close! But still not perfect.

Let's fix that other one

Question was "what's the probability of seeing at most 940 non-defective widgets?"

$$\mathbb{P}(X \leq 940)$$

With the CLT.

The event we're interested in is $\mathbb{P}(X \leq 940)$

$\mathbb{P}(X<940.5)$

$$\mathbb{P}(X \leq 940)$$
$$= \mathbb{P}\left(\frac{X-1000\cdot.95}{\sqrt{1000\cdot.0475}} \leq \frac{940-1000\cdot.95}{\sqrt{1000\cdot.0475}}\right)$$
$$\approx \mathbb{P}\left(Y \leq \frac{940-1000\cdot.95}{\sqrt{1000\cdot.0475}}\right) \textit{ By CLT}$$
$$= \Phi\left(\frac{940-1000\cdot.95}{\sqrt{1000\cdot.0475}}\right)$$
$$\approx \Phi(-1.45) = 1 - \Phi(1.45)$$
$$\approx 1 - .92647 = .07353.$$

$$\mathbb{P}(X \leq 940.5)$$
$$= \mathbb{P}\left(\frac{X-1000\cdot.95}{\sqrt{1000\cdot.0475}} \leq \frac{940.5-1000\cdot.95}{\sqrt{1000\cdot.0475}}\right)$$
$$\approx \mathbb{P}\left(Y \leq \frac{940.5-1000\cdot.95}{\sqrt{1000\cdot.0475}}\right) \textit{ By CLT}$$
$$= \Phi\left(\frac{940.5-1000\cdot.95}{\sqrt{1000\cdot.0475}}\right)$$
$$\approx \Phi(-1.38) = 1 - \Phi(1.38)$$
$$\approx 1 - .91621 = .08379.$$

True answer: .08673

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

Where's the continuity correction?

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

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

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

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

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

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

$\approx .62552$

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 $X \sim \mathrm{Bin}(n, p)$, $\Omega_X = \{0,1, \dots, n\}$

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.

Confidence Interval (visualized)