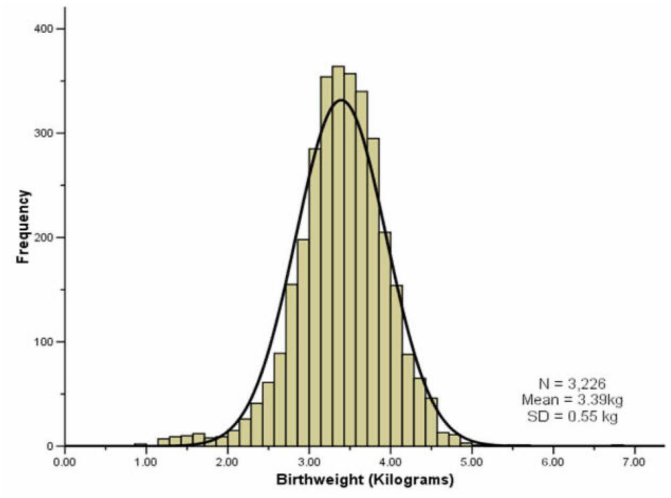
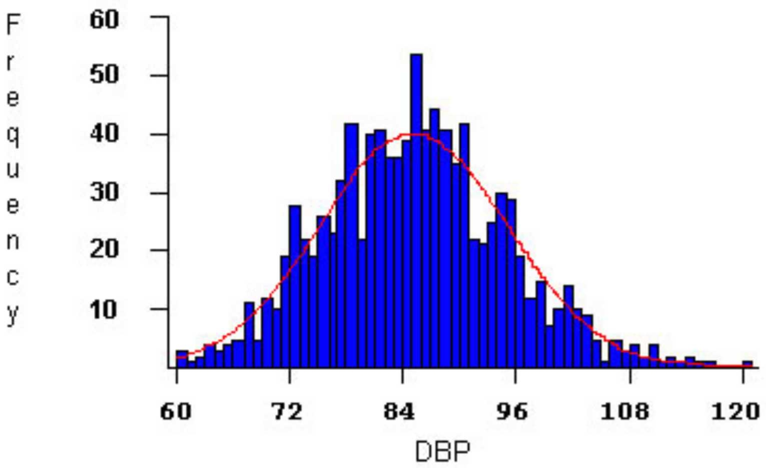
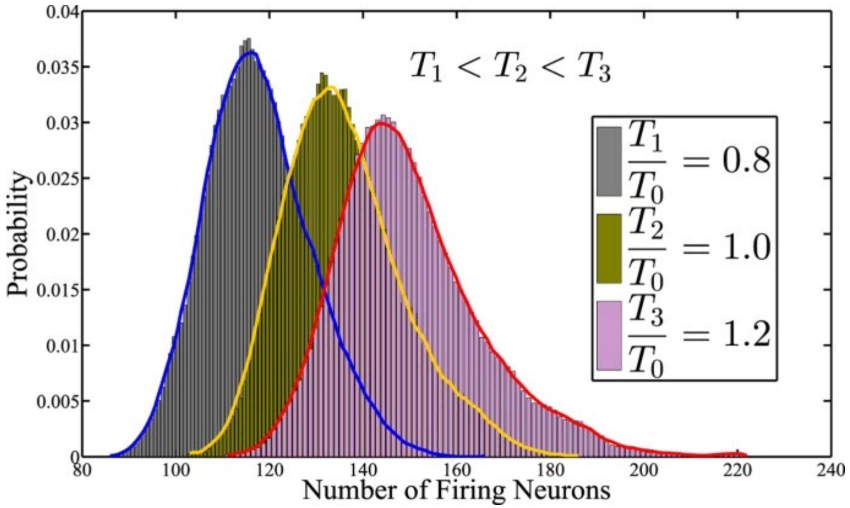
Central Limit Theorem

CSE 312 Su23
Lecture 14

We see normal distributions… everywhere!







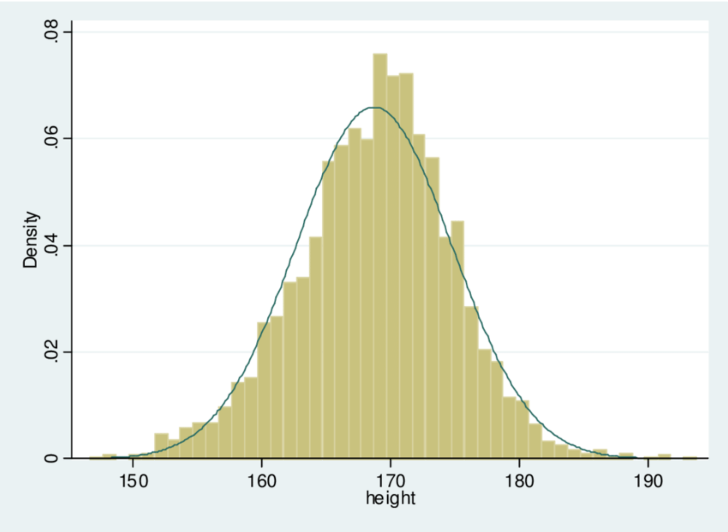
But… why?

As it turns out, this is because of the following idea:

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

Wait! What does the *'sum of independent random variables'* have to do with the real-world examples?

Real-world example explained: height



(cm)

R.A. Fisher (1918) suggested:

Height may be expressed as the sum of many independent, random factors

For e.g.,

How much milk you drank per day as a child

Which variant of GH1 (a growth hormone) you have

How much protein/calcium/vitamins/minerals you had as a child

How many hours of sleep you averaged

How many hours of physical activity you averaged

I.e., $H = H_1 + H_2 + \cdots + H_n, with\ idp\ H_i s, H \sim \mathcal{N}(.,.)$

More formally...

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

independent & identically distributed

Central Limit Theorem

Let $\underbrace{X_1, X_2, \ldots, X_n}$ 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 $\mathcal{N}(0,1)$

Independent and identically distributed (i.i.d)

For n variables $X_1, \ldots, X_n$ they are independent and identically distributed if:

- $X_1, X_2, \ldots, X_n$ are independent, and,
- all have the same PMF (if discrete) or PDF (if continuous)

  that is, they have the same mean and variance

i.i.d?

Are $X_1, X_2, \ldots, X_n$ i.i.d with the following distributions?

> $X_i \sim \boxed{Exp(\lambda)}, X_i$ *independent* ✓

> $X_i \sim Exp(\lambda_i), X_i$ *independent* $\lambda_1 \neq \lambda_2$ ✗

> $X_i \sim Exp(\lambda), X_1 = X_2 = \cdots = X_n$ ✗

> $X_i \sim Bin(n_i, p), X_i$ *independent* ✗ n_1 $Ber(p)$

Breaking down the theorem

$X_1 + \dots + X_n$

$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = n\mu$ $\quad$ $Var\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2$

$\sqrt{n}\sigma$

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

Caveat (for real-world height example)

The height problem has a few caveats:

- The factors are not necessarily completely independent of each other. But we can somewhat reasonably approximate them as being so…
- They are not identically distributed

It still works out: convergence of the mean to the normal distribution also occurs for non-identical distributions or for non-independent observations if they comply with certain conditions

(TLDR; other versions of CLT work when we relax the i.i.d assumption. These versions are outside the scope of this class.)

Proof of the CLT?

$\mathbb{E}[Y_n^1]$

$\mathbb{E}[Y_n^2]$

$\vdots$

We're not going to prove this.

moment

How is the proof done?

★ Step 1: Prove that for all positive integers k, $\mathbb{E}[(Y_n)^k] \to \mathbb{E}[Z^k]$

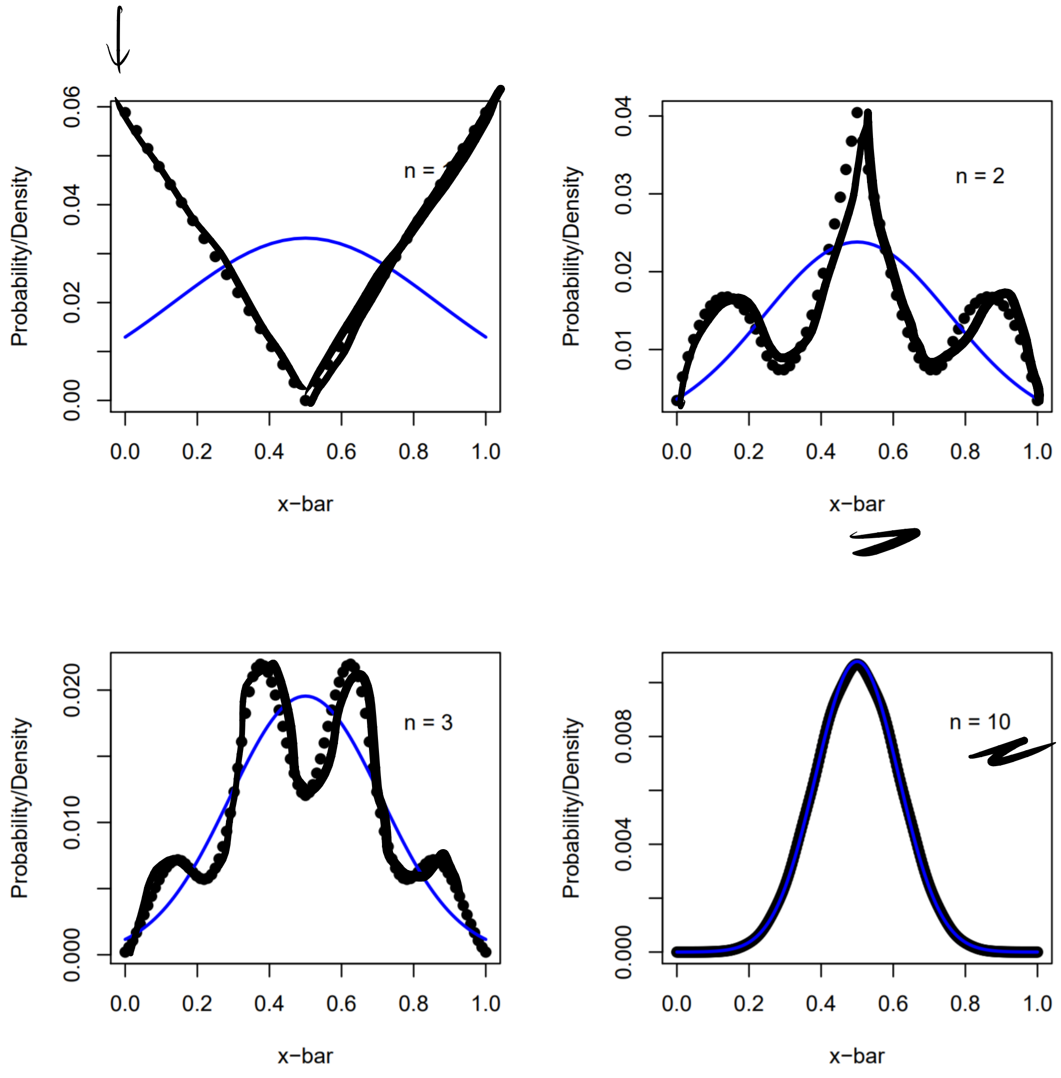
$\to 1$

$N(0,1)$

Step 2: Prove that if $\mathbb{E}[(Y_n)^k] = \mathbb{E}[Z^k]$ for all k then $F_{Y_n}(z) = F_Z(z)$

$$\text{Var} = \text{2nd moment} - (\text{1st moment})^2$$

“Proof by example”

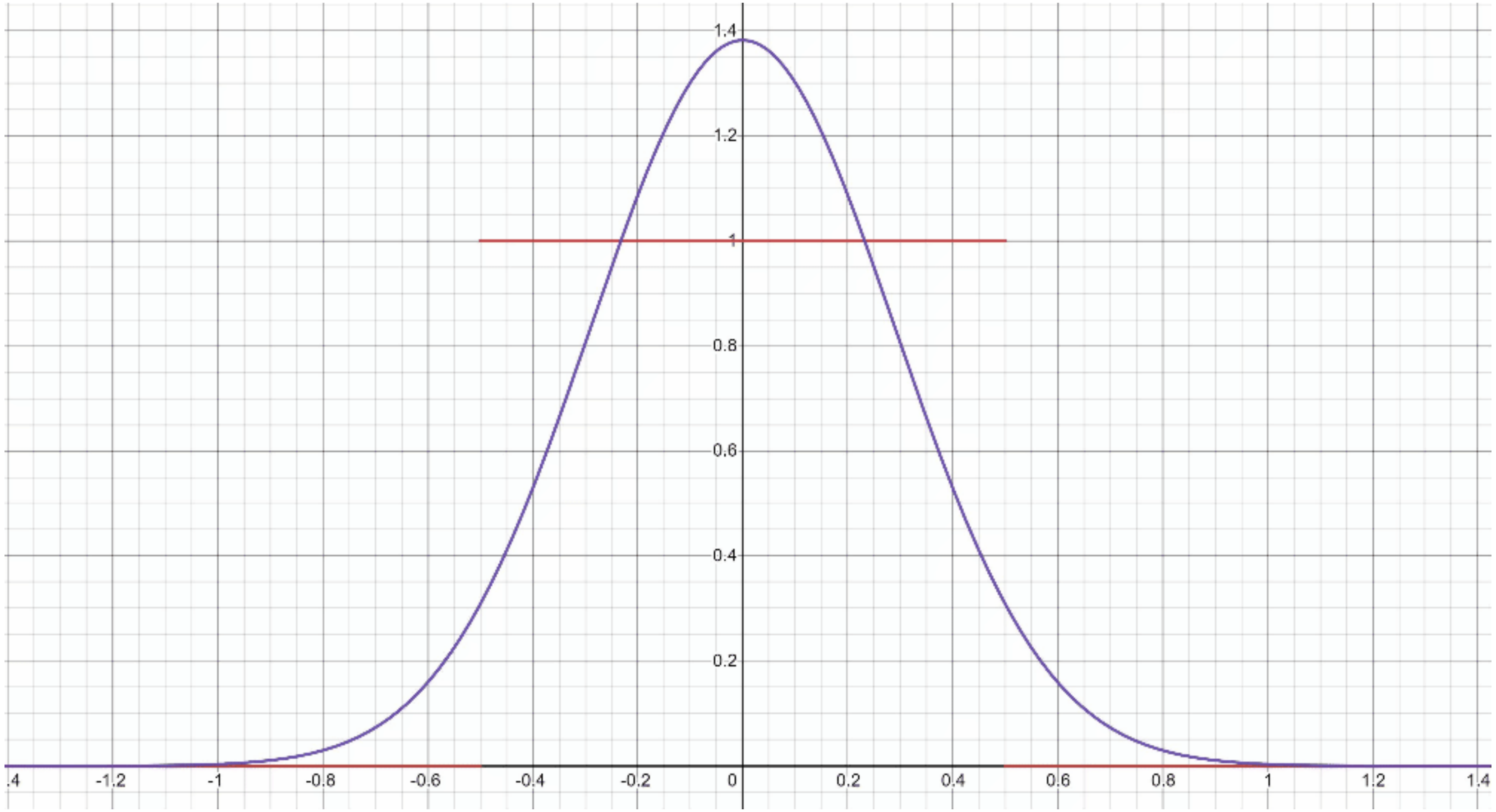


The dotted lines show an “empirical PMF” – a PMF estimated by running the experiment a large number of times.

The blue line is the normal rv that the CLT predicts.

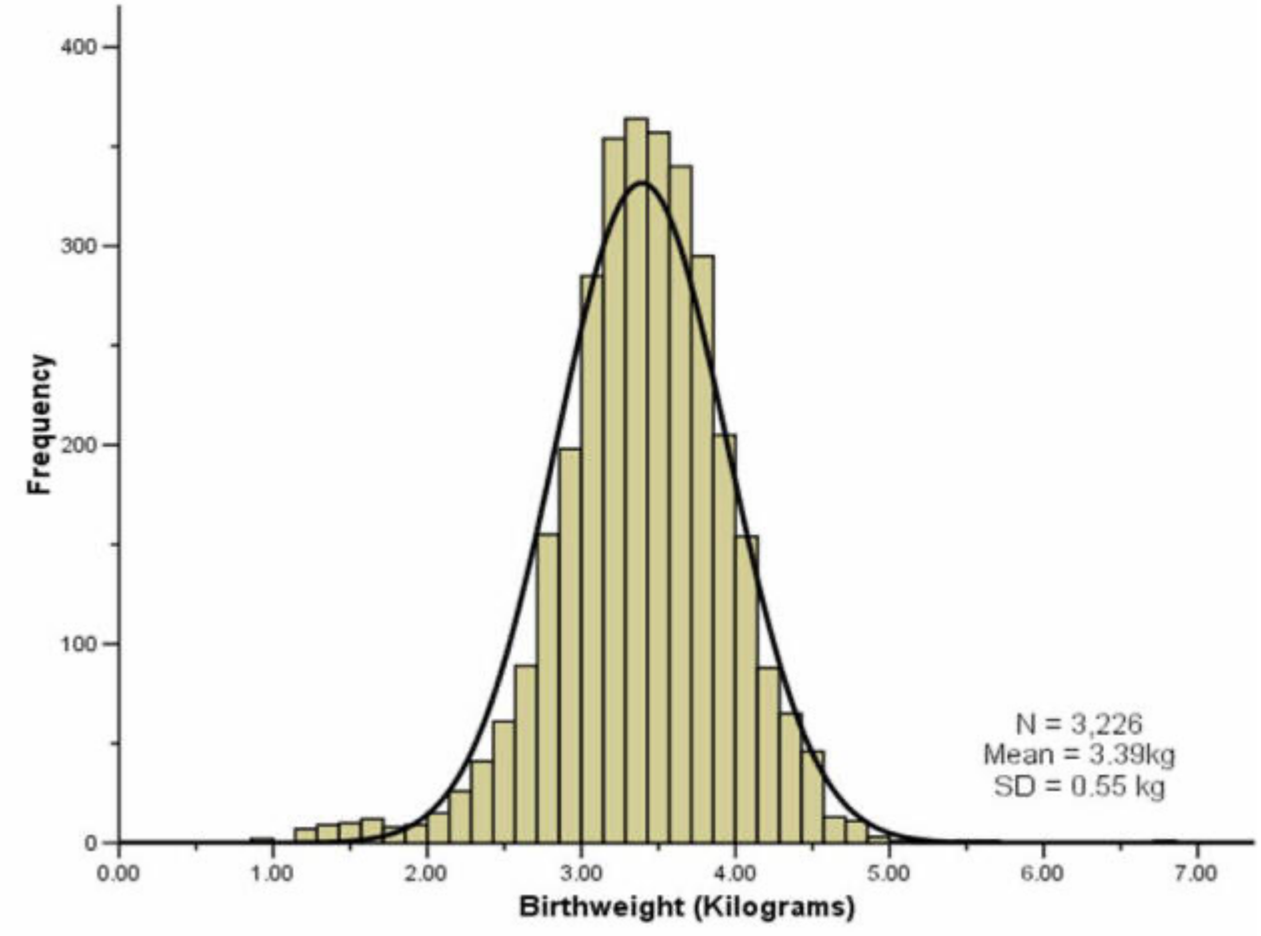
Shown are $n = 1,2,3,10$

“Proof by example” -- uniform



https://www.desmos.com/calculator/2n2m05a9km

“Proof by real-world”



birthweight

A lot of real-world bell-curves can be explained as:

1. The random variable comes from a combination of independent factors.

2. The CLT says the distribution will become like a bell curve.

Theory vs. Practice

The formal theorem statement is “in the limit” $n \to \infty$

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)

X : # of non-~~def~~ widget

What is the probability?

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

95%

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

Ber(0.95)

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

So, let's use the CLT instead

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

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

$$\frac{\Sigma - n \cdot \mu}{\sqrt{n} \cdot \sigma}$$

$n = 1000$

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

CLT

With the CLT.

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

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

With the CLT.

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

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

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

$N(0,1)$.

With the CLT.

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

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

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

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

$Y_{1000} \sim \mathcal{N}(0,1)$

With the CLT.

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

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

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

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

$$\approx \mathbb{P}\left(Y \leq \frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right) \textit{ by CLT}$$

With the CLT.

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

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

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

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

$$\approx \mathbb{P}\left(Y \le \frac{940 - 1000\cdot.95}{\sqrt{1000\cdot.0475}}\right) \textit{ by CLT}$$

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

$\Phi = f_Z(z)$

With the CLT.

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

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

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

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

$$\approx \mathbb{P}\left(Y \leq \frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right) \text{ by CLT}$$

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

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

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

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?

A problem

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

True value, we can get with binomial:

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

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

What does the CLT say?

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

$\frac{-n\mu}{\sqrt{n}\sigma}$

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

$= 0$

Continuous.

Uh oh.

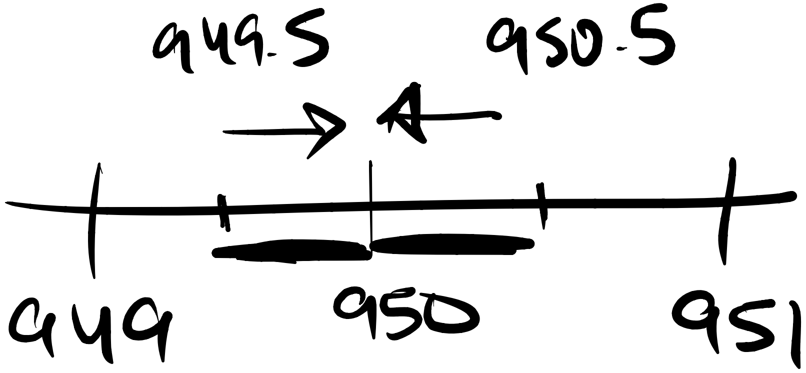
Continuity Correction

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

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

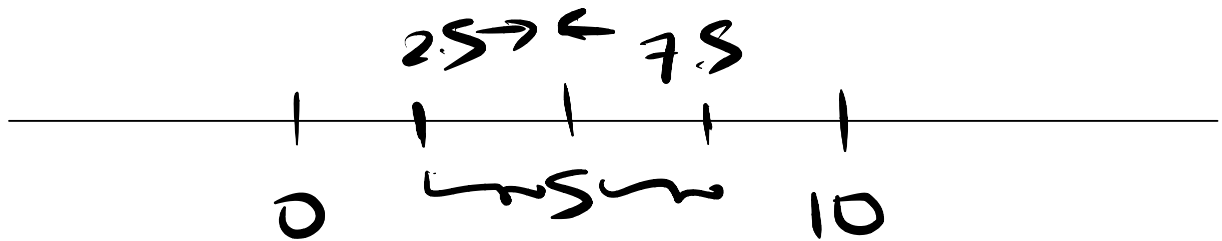
Before we switch from the binomial to the normal, ask "what values of a continuous random variable would round to this event?"

$X = 950$



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

$$\approx \mathbb{P}(949.5 \le Z \le 950.5)$$



½ interval size.

Applying the continuity correction

$X \sim \text{Bin}$

$\mathbb{P}(X = 949.5) = 0$

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

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

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

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

$$\approx \mathbb{P}\left(\frac{949.5-950}{\sqrt{1000\cdot.0475}} \le Y < \frac{950.5-950}{\sqrt{1000\cdot.0475}}\right) \quad \text{By CLT}$$

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

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

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

Still an Approximation

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

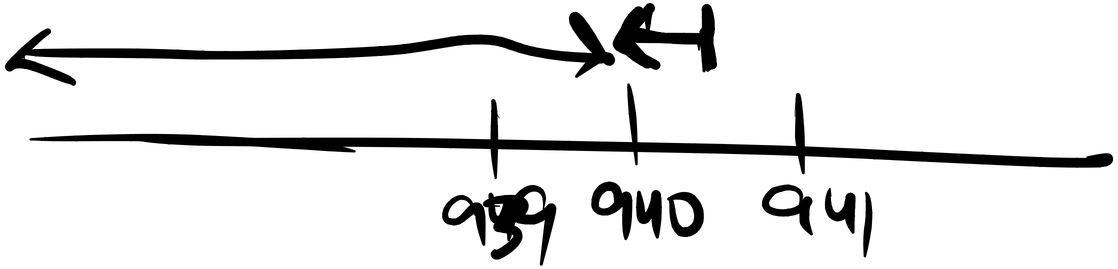
The CLT approximates: 0.0558

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?”

With the CLT.



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

$\mathbb{P}(X \le 940)$ x-discrete

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

$$\approx \mathbb{P}\left(Y \le \frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right) \textit{ By CLT}$$

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

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

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

$\mathbb{P}(X \le 940.5)$ continuity correction

$$= \mathbb{P}\left(\frac{X - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}} \le \frac{940.5 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right)$$

$$\approx \mathbb{P}\left(Y \le \frac{940.5 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right) \textit{ By CLT}$$

$$= \Phi\left(\frac{940.5 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right)$$

$$\approx \Phi(-1.38) = 1 - \Phi(1.38)$$

$$\approx 1 - .91621 = .08379.$$

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

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

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

$\approx .62552$

$X = \sum_{i=1}^{10} X_i$

$\underline{\underline{n = 10}}$

$-n \cdot \mu$

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

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. Normalize 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.

more effective when n is small.

$\lim \quad n \to \infty$

Other Versions of CLT

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 $Y_n = X_1 + X_2 + \cdots + X_n$

As $n \to \infty$, the CDF of Y_n converges to the CDF of $\mathcal{N}(n \cdot \mu, n \cdot \sigma^2)$

A Note on Approximating Binomials

Poisson Approximation

When deriving the Poisson PMF, we showed that a Poisson is a Binomial in the limit as $n \to \infty$ and $p = \frac{\lambda}{n}$.

Similar line of thought: Poisson is a good approximation for Binomials when n is large, p is small and $\lambda = np$ is "moderate" (typical ranges: $n > 20$ and $p < 0.05$ or $n > 100$ and $p < 0.1$.

Works well, even if trials are not fully independent.

Normal Approximation

This is the use of CLT with the continuity correction to approximate a Binomial.

Typically, a better approximation if n is large and p is moderate. p is fixed with this approximation.

Especially if variance is > 10.

If both work, use the Normal!

Sample Statistics

Sample Mean

The sample mean is the average of n independent and identically distributed random variables.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\frac{x_1 + x_2 + \cdots + x_n}{n}$$

Distribution of Sample Mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Let $X_1, X_2, \dots, X_n$ be i.i.d with mean μ and variance σ^2. As $\boxed{n \to \infty}$:

$Y = \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$ [CLT, as $n \to \infty$]

$\bar{X} = \frac{1}{n} Y$

$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ [Scaling a normal]

$\frac{1}{n} n\mu = \mu$

1) Normally distributed.

$\left(\frac{1}{n}\right)^2 n\sigma^2 = \frac{\sigma^2}{n}$

Key Consequences

If we have a sample of size n whose mean μ and variance σ^2 are known, we can compute probabilities on the sample mean $\bar{X}$, using the CLT to approximate it as a normal.

sampling
bootstrapping

Real world:

We often don't know μ or σ^2 of our original distribution.

We can collect data on a sample of our population.

How do we estimate μ and σ^2 from our sample? And what can we conclude using those estimates?

CLT

Next time: polling, and a brief glance at the above questions :D