Central Limit Theorem

CSE 312 Summer 21
Lecture 16

Announcements

Review Summary 2 due this Friday.

Programming question in HW 6.

Table of Standard Normal CDF

The way we'll evaluate the CDF of a normal is to:

1. convert to a standard normal
2. Round the "z-score" to the hundredths place.
3. Look up the value in the table.

It's 2021, we're using a table?

The table makes sure we have consistent rounding rules (makes it easier for us to debug with you).

You can't evaluate this by hand – the "z-score" can give you intuition right away.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999

More Practice

Let $X \sim \mathcal{N}(3, 2)$

What is the probability of $1 \leq X \leq 4$?

$$\mathbb{P}(1 \leq X \leq 4)$$

$$= \mathbb{P}\left(\frac{1-3}{\sqrt{2}} \leq \frac{X-3}{\sqrt{2}} \leq \frac{4-3}{\sqrt{2}}\right)$$

$$= \mathbb{P}\left(-1.41 \leq \frac{X-3}{\sqrt{2}} \leq 0.71\right)$$

$$= \Phi(0.71) - \Phi(-1.41)$$

$$= \Phi(0.71) - \left(1 - \Phi(1.41)\right) = 0.76115 - (1 - 0.92073) = 0.68188$$

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

Breaking down the theorem

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

Proof of the CLT?

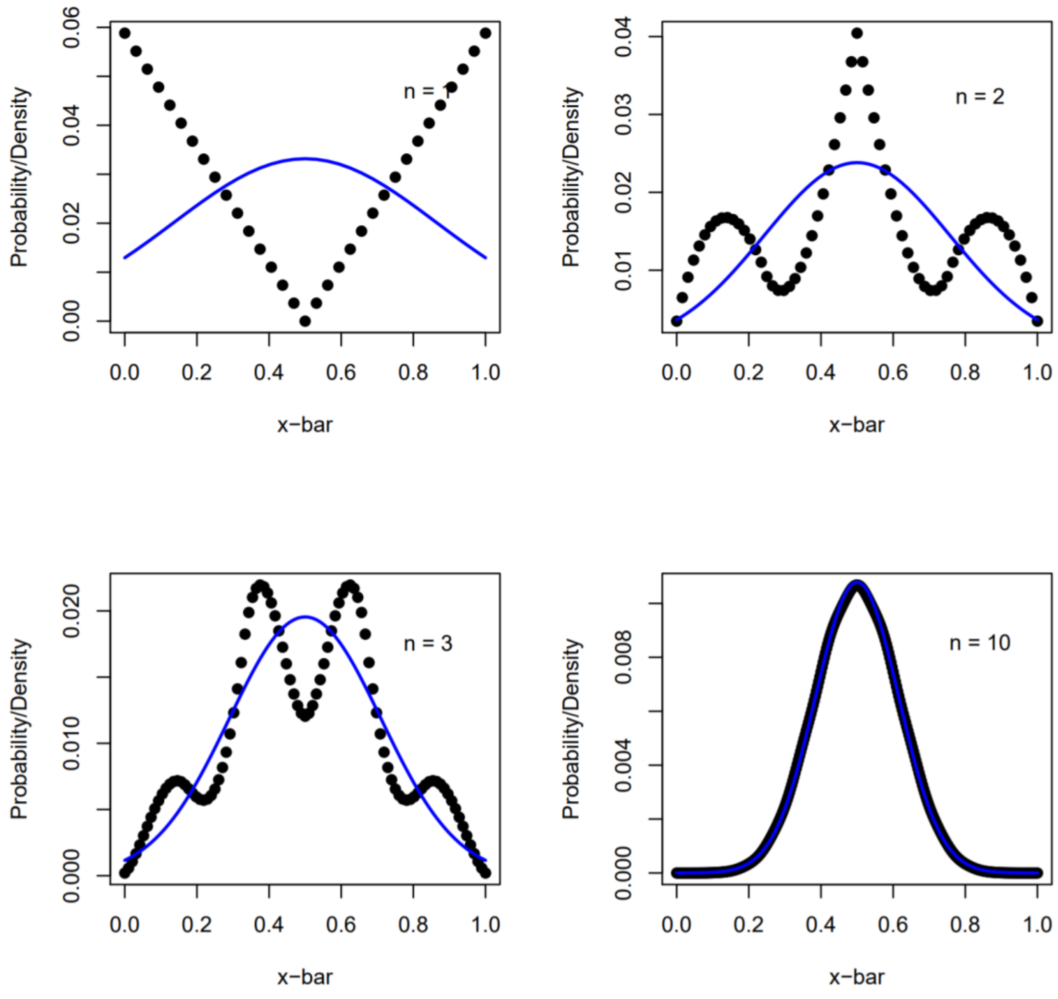
No.

How is the proof done?

Step 1: Prove that for all positive integers $k\ and\ Z \sim \mathcal{N}(0,1), \mathbb{E}\left[(Y_n)^k\right] \to \mathbb{E}[Z^k]$

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

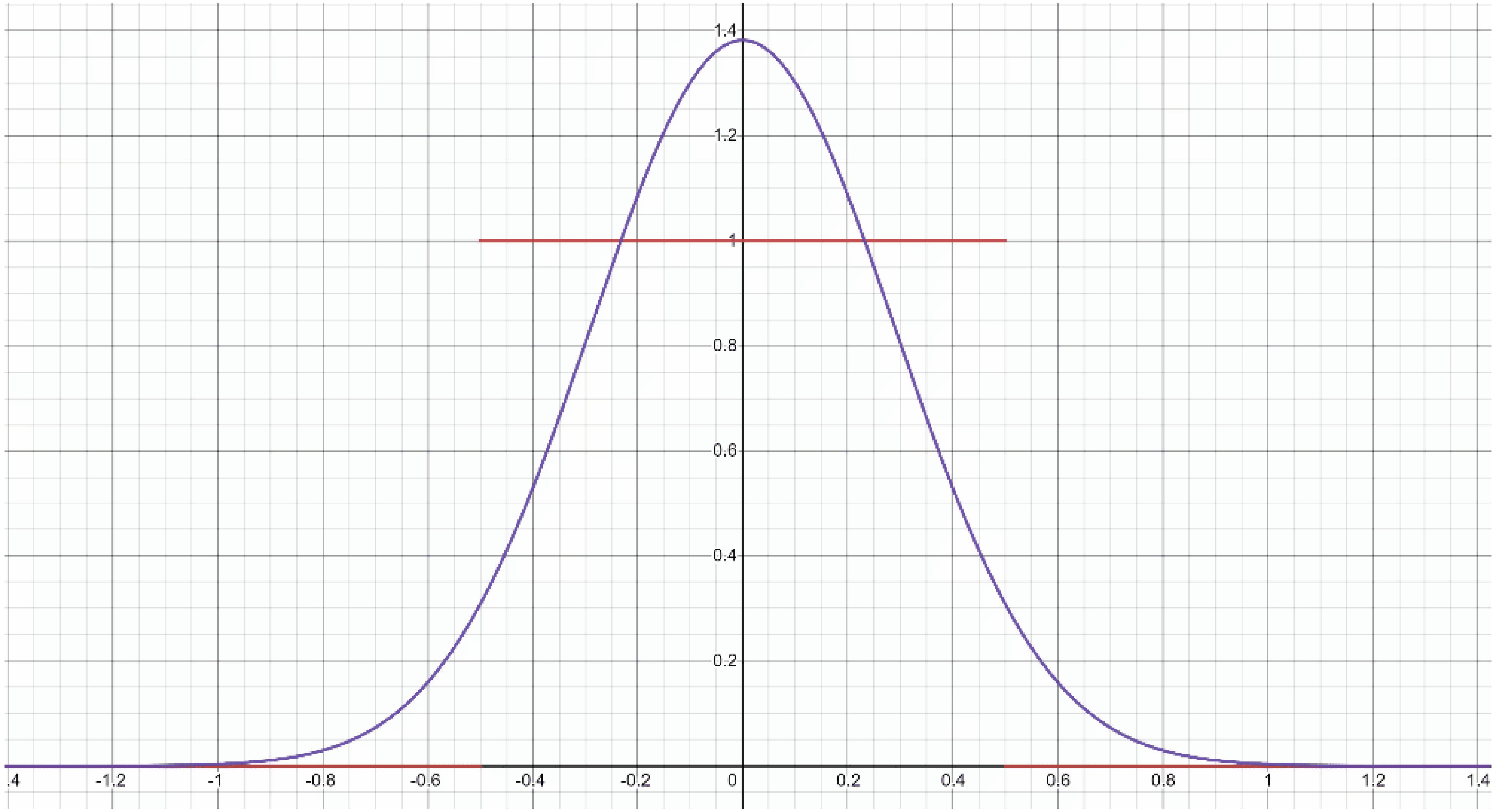
“Proof by example”



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.

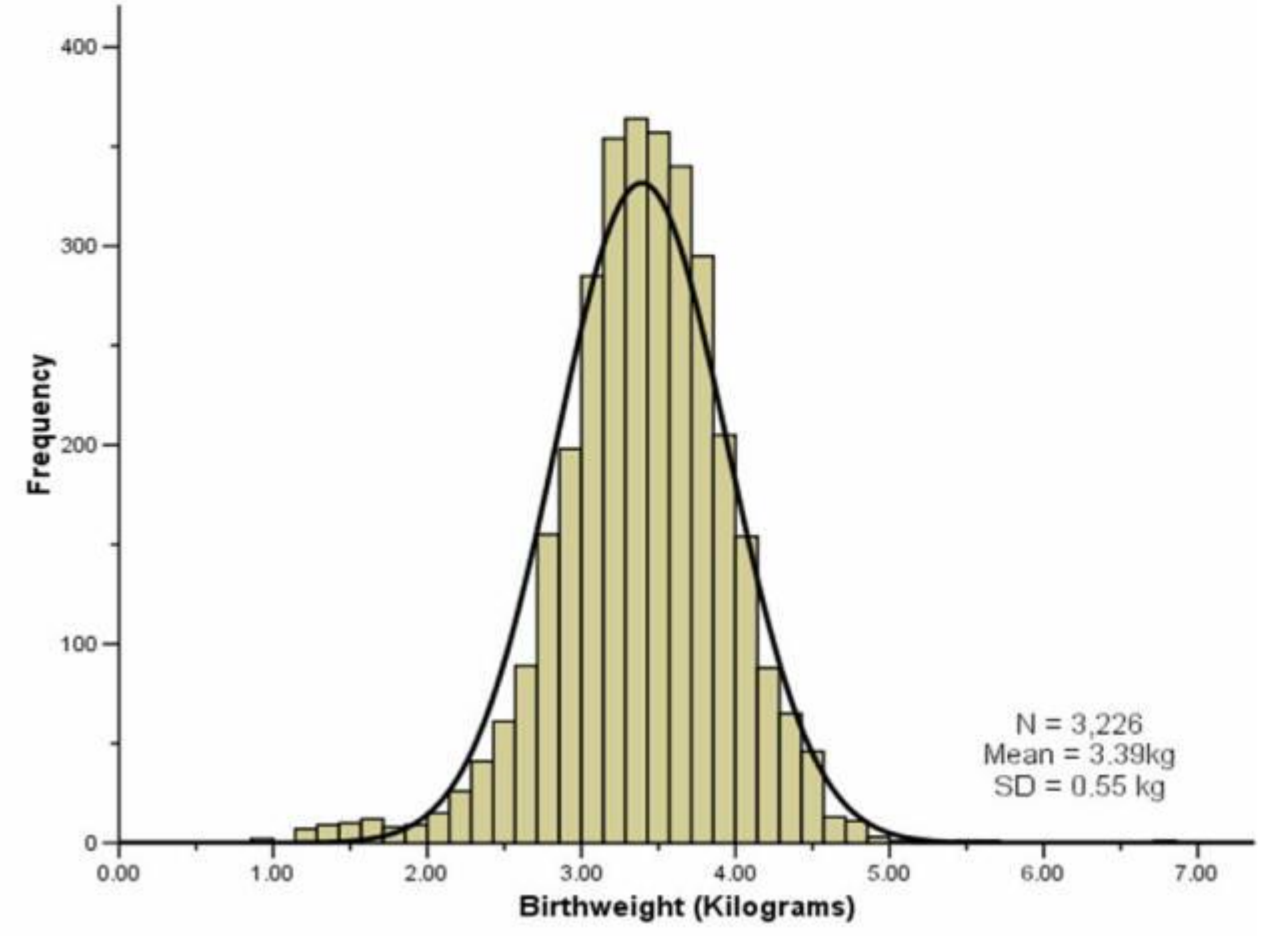
Shown are $n = 1,2,3,10$

“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"

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

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

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

$\sum_{k=0}^{940} \binom{1000}{k} (0.95)^k \cdot (0.05)^{1000-k} \approx 0.08673$

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)

So, let's use the CLT instead

$\mathbb{E}[X_i] = p = 0.95.$

$\text{Var}(X_i) = p(1-p) = 0.0475$

$Y_{1000} = \frac{\sum_{i=1}^{1000} X_i - 1000 \cdot 0.95}{\sqrt{1000 \cdot 0.0475}}$ is approximately $\mathcal{N}(0,1)$.

With the CLT.

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

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

$$= \mathbb{P}\left(\frac{X-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}} \leq \frac{940-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}}\right)$$

$$= \mathbb{P}(Y_{1000} \leq \frac{940-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}})$$

$$\approx \mathbb{P}\left(Y \leq \frac{940-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}}\right) \textit{ by CLT}$$

$$= \Phi\left(\frac{940-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}}\right)$$

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

$$\approx 1 - 0.92647 = 0.07353.$$

It’s an approximation!

The true probability is

$$1 - \sum_{k=941}^{1000} \binom{1000}{k} (0.95)^k \cdot (0.05)^{1000-k} \approx 0.08673$$

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:

$\binom{1000}{950} \cdot (0.95)^{950} \cdot (0.05)^{50} \approx 0.05779$

What does the CLT say?

$= \mathbb{P}\left(\frac{X-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}} = \frac{950-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}}\right)$

$\approx \mathbb{P}(Y = 0)$

$= 0$

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?”

Applying the continuity correction

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

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

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

$= \mathbb{P}\left(\frac{949.5-950}{\sqrt{1000\cdot 0.0475}} \leq \frac{X-950}{\sqrt{1000\cdot 0.0475}} < \frac{950.5-950}{\sqrt{1000\cdot 0.0475}}\right)$

$\approx \mathbb{P}\left(\frac{949.5-950}{\sqrt{1000\cdot 0.0475}} \leq Y < \frac{950.5-950}{\sqrt{1000\cdot 0.0475}}\right)$ By CLT

$= \Phi\left(\frac{950.5-950}{\sqrt{1000\cdot 0.0475}}\right) - \Phi\left(\frac{949.5-950}{\sqrt{1000\cdot 0.0475}}\right)$

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

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

Still an Approximation

$\binom{1000}{950} \cdot (0.95)^{950} \cdot (0.05)^{50} \approx 0.05779$ 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 $\mathbb{P}(X \leq 940)$

$$\mathbb{P}(X \leq 940)$$
$$= \mathbb{P}\left(\frac{X-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}} \leq \frac{940-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}}\right)$$
$$\approx \mathbb{P}\left(Y \leq \frac{940-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}}\right) \textit{ By CLT}$$
$$= \Phi\left(\frac{940-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}}\right)$$
$$\approx \Phi(-1.45) = 1 - \Phi(1.45)$$
$$\approx 1 - 0.92647 = 0.07353.$$

$$\mathbb{P}(X \leq 940.5)$$
$$= \mathbb{P}\left(\frac{X-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}} \leq \frac{940.5-1000\cdot .95}{\sqrt{1000\cdot .0475}}\right)$$
$$\approx \mathbb{P}\left(Y \leq \frac{940.5-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}}\right) \textit{ By CLT}$$
$$= \Phi\left(\frac{940.5-1000\cdot 0.95}{\sqrt{1000\cdot 0.0475}}\right)$$
$$\approx \Phi(-1.38) = 1 - \Phi(1.38)$$
$$\approx 1 - 0.91621 = 0.08379.$$

True answer: 0.08673

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

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

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

≈ 0.62552

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. 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.