Final Exam: 8 days
→ HW7 is optional (100% credit)
→ Vitamin 7 is not optional

Normal RVs

Lecture 6D: Continuous Probability II

UC Berkeley CS70
Summer 2023
Nikki Suzani

Gaussian (Normal) Random Variables

A normal (or Gaussian) distribution is a distribution with two parameters, μ – its expectation – and σ^2 – its variance.

$X \sim \text{Normal}(\mu, \sigma^2)$

mean is μ

variance is σ^2

PDF of X

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

68 - 95 - 99.7

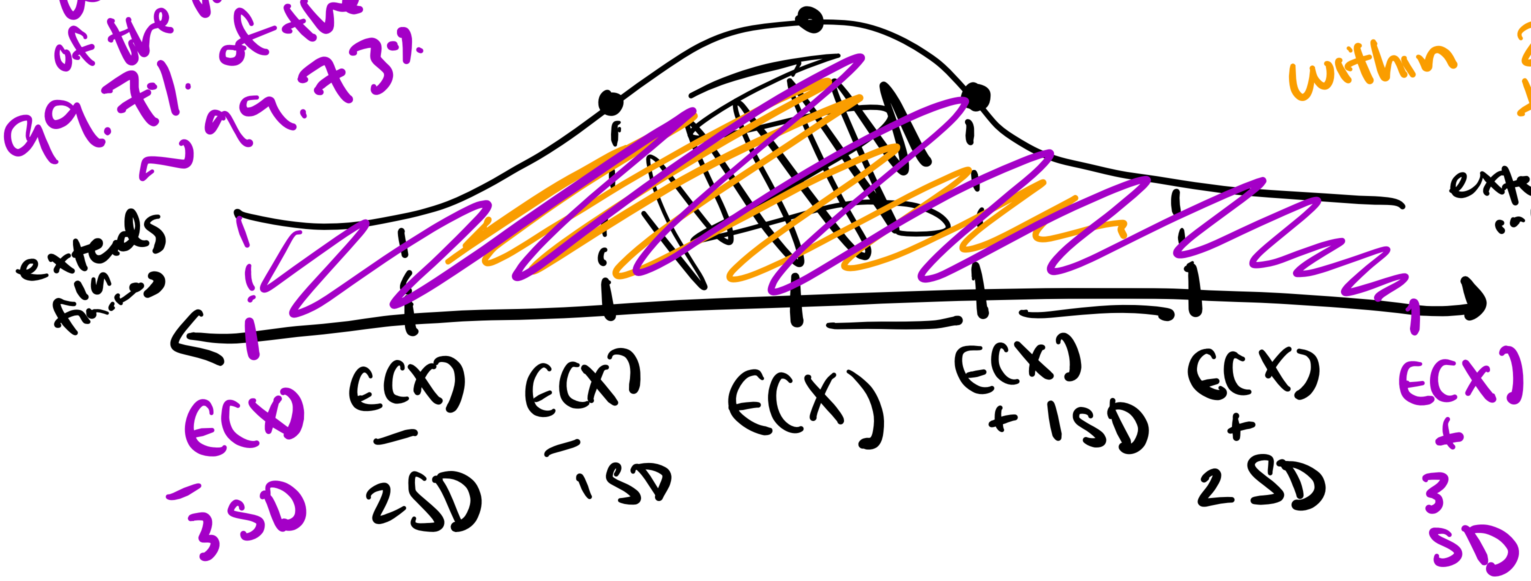
Some Cool Properties of the Normal Distribution

within 1 SD of the mean, 68% of the data

Symmetric about the mean

within 3 SD of the mean 99.7% of the data
~99.73%

within 2 SD of the mean, 95% of the data
↓
95.45
↓
(1.96 instead of 2)



Gaussian (Normal) Random Variables (cont.)

Normal w/ a mean of 0 & a variance of 1

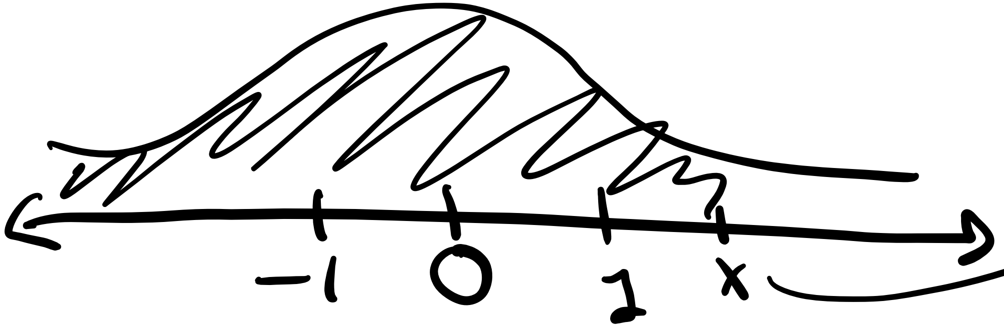
CDF of the Normal(0, 1) distribution is referred to as $\Phi(x)$.

Standard Normal

standard normal

value x

$$\Phi(x) = P(Z \le x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{ -\frac{u^2}{2} \right\} du.$$



$\Phi(x)$

$\Phi(x)$

Values x	$\Phi(x)$
1.96	0.975
$\vdots$	$\vdots$

“Scaling” Normal Random Variables

Y is a standard normal RV

Normal(μ, σ^2)

Lemma 21.1. *If* $X \sim N(\mu, \sigma^2)$*, then* $Y = \frac{X-\mu}{\sigma} \sim N(0,1)$*. Equivalently, if* $Y \sim N(0,1)$*, then* $X = \sigma Y + \mu \sim N(\mu, \sigma^2)$.

$$E[Y] = E\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma}E[X-\mu] = \frac{1}{\sigma}\left(E[X] - E[\mu]\right)$$

$$= \frac{1}{\sigma}\left(E[X] - \mu\right) = \frac{1}{\sigma}(\mu - \mu) = 0$$

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\mathrm{Var}(X-\mu) = \frac{1}{\sigma^2}\mathrm{Var}(X)$$

$$= \frac{1}{\sigma^2}\cdot\sigma^2 = 1$$

“Scaling” Normal Random Variables (cont.)

$X \sim \text{Normal}(\mu, \sigma^2)$

$$\mathbb{P}[X \le a] = \mathbb{P}\left[Y \le \frac{a-\mu}{\sigma}\right] = \Phi\left(\frac{a-\mu}{\sigma}\right)$$

where Y is standard normal.

This allows us to use the CDF!

$$P[X \le a] = P[X-\mu \le a-\mu] = P\left[\frac{X-\mu}{\sigma} \le \frac{a-\mu}{\sigma}\right]$$

$$= P\left[Y \le \frac{a-\mu}{\sigma}\right]$$

Sum of Independent Normal RVs

$\downarrow Var(X)$ $\quad\quad$ $\swarrow Var(Y)$

Corollary 21.1. *Let $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ be independent normal random variables. Then for any constants $a, b \in \mathbb{R}$, the random variable $Z = aX + bY$ is also normally distributed with mean $\mu = a\mu_X + b\mu_Y$ and variance $\sigma^2 = a^2\sigma_X^2 + b^2\sigma_Y^2$.*

$$Z \sim \text{Normal}(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$$

$$E[Z] = E[aX + bY] = E[aX] + E[bY] = aE[X] + bE[Y]$$

$$= a\mu_x + b\mu_y$$

$$Var(Z) = Var(aX + bY) = Var(aX) + Var(bY)$$

by independence

$$= a^2 Var(X) + b^2 Var(Y) = a^2 \sigma_x^2 + b^2 \sigma_y^2$$

Central Limit Theorem

independent & identically distributed

Let X_1, X_2, … X_n be a sequence of n i.i.d. random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. When **n** is large, both the sample sum and the sample mean can be approximated as normal random variables.

$X_i \to E(X_i) = \mu$

$\to Var(X_i) = \sigma^2$

Sample Sum: $X = X_1 + X_2 + \dots X_n$

$E(X) = E(X_1 + X_2 + \dots X_n) = nE(X_1)$

$= n\mu$

Really important: X_i can follow any distribution

$Var(X) = Var(X_1 + X_2 + \dots X_n) = Var(X_1) + \dots Var(X_n)$

$= nVar(X_1) = n\sigma^2$

Sample Sum Expectation & Variance

Let X_1, X_2, … X_n be a sequence of n i.i.d. random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Find the distribution of $S_n = X_1 + \ldots + X_n$ for large n.

$X \sim Normal(n\mu, n\sigma^2)$

Sample Sum: $X = X_1 + X_2 + \ldots X_n$

$E(X) = E(X_1 + X_2 + \ldots X_n) = nE(X_1)$
$= n\mu$

$Var(X) = Var(X_1 + X_2 + \ldots X_n) = Var(X_1) + \ldots Var(X_n)$
$= nVar(X_1) = n\sigma^2$

$X_i \rightarrow E(X_i) = \mu$

$\rightarrow Var(X_i) = \sigma^2$

Really important:
X_i can follow
<u>any</u> distribution

Scaling to Standard Normal

$X \sim \text{Normal}(n\mu, n\sigma^2)$

mean

$$Y = \frac{X - n\mu}{\sqrt{n\sigma^2}}$$

SD or $\sqrt{\text{var}}$

$Y \sim \text{Normal}(0, 1)$

Sample Mean Expectation and Variance

Let X_1, X_2, … X_n be a sequence of n i.i.d. random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Find the distribution of X-bar = $S_n/n = X_1 + \ldots + X_n$ for large n.

$$\bar{X} = \frac{S_n}{n} = \frac{X_1 + X_2 + \ldots X_n}{n}$$

Normally dist. by CLT

$$\boxed{E(\bar{X})} = E\left(\frac{X_1 + X_2 + \ldots X_n}{n}\right) = \frac{1}{n} E(X_1 + X_2 + \ldots X_n)$$

$$= \cancel{\frac{1}{n}} \cdot \cancel{n} \cdot E(X_1) = \underline{\mu}$$

$$\boxed{Var(\bar{X})} = Var\left(\frac{X_1 + X_2 + \ldots X_n}{n}\right) = \frac{1}{n^2} Var(X_1 + X_2 + \ldots X_n)$$

$$= \cancel{\frac{1}{n^2}} \cdot \cancel{n} \cdot Var(X_1)$$

$$= \frac{Var(X_1)}{n} = \frac{\sigma^2}{n}$$

Scaling to Standard Normal

$$\bar{X} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Y = \frac{\bar{X} - \mu}{\left(\sqrt{\frac{\sigma^2}{n}}\right)} = \frac{(\bar{X} - \mu)\sqrt{n}}{\sigma}$$

$$Y \sim Normal(0,1)$$

How to Solve CLT Questions

1. Identify that you can use CLT – sample mean/sum of i.i.d. random variables for **large n**.

2. Calculate the Mean and Variance of the sample mean/sum

≤ 3 $\qquad \dfrac{3 - E(X)}{SD(X)}$

3. Convert your **value** into the standard normal value (subtract mean, divide by standard deviation)

4. Use Φ(standard normal value) $\rightarrow \Phi\left(\dfrac{3 - E(X)}{SD(X)}\right)$

Dice Example

2. My friend and I gamble on rolls of a die. Each time the die is rolled,

- my friend gives me a dollar if the number of spots is five or six,
- I give my friend a dollar if the number of spots is one or two,
- and no money changes hands if the number of spots is three or four.

this dist

If we play this game 400 times, approximately what is the chance that my net gain is more than 20 dollars?

total value

$$X = X_1 + X_2 + \dots X_{400}$$

$$X_i \begin{cases} 1 & \text{w/ probability } 2/6 \\ -1 & \text{w/ probability } 2/6 \\ 0 & \text{w/ probability } 2/6 \end{cases}$$

CLT : n is large, sum, independent & identically dist

$$E(X_i) = 1 \cdot 2/6 - 1 \cdot 2/6 = 0$$

$$Var(X_i) = E(X_i^2) - E(X_i)^2$$

Dice Example (cont.)

$$= \underline{1 \cdot 2/6 + 1 \cdot 2/6} - \underline{0}^2 = 2/3$$

$$\Phi(x) = P(\underline{Z \le x})$$

$$X \sim \text{Normal}(\_\_\_, \_\_\_)$$

$$E(X) = E(X_1 + X_2 + \ldots X_{400}) = 400\, E(X_1)$$
$$= 400 \cdot 0 = 0$$

$$Var(X) = Var(X_1 + X_2 + \ldots X_{400}) = 400 \cdot Var(X_1)$$
$$= Var(X_1) + Var(X_2) + \ldots Var(X_{400}) \qquad = 400 \cdot 2/3$$

$$P(\underline{X} > \underline{20}) = \boxed{1 - P(X \leq 20)}$$

$$Y = \frac{X - E(X)}{\sqrt{Var(X)}} = \frac{X - E(X)}{SD(X)} = \frac{X - 0}{\sqrt{400 \cdot 2/3}}$$

$$= 1 - P\left(Y \leq \frac{20 - 0}{\sqrt{400 \cdot 2/3}}\right)\Bigg] \quad X \sim \text{Normal}(0, 400 \cdot \tfrac{2}{3})$$

$$= 1 - \Phi\left(\frac{20}{\sqrt{400 \cdot 2/3}}\right)$$

$Y \sim \text{Normal}(0, 1)$

Exam Example

A large number $N = 400$

N students take an exam where the ~~average is 50~~, and the variance is 5. Let S_n be the sum of their n scores, and assume all of their scores are independent. What is the probability that their average score is greater than 72.5?

CLT: large n, average, identically dist. & independent

$$\bar{X} = \frac{S_n}{n} = \frac{X_1 + X_2 + \dots X_n}{n} \qquad \bar{X} \sim \text{Normal}$$

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \dots X_n}{n}\right) = \frac{1}{n} E(X_1 + X_2 + \dots X_n)$$

$$= \frac{1}{n} \cdot n \cdot E(X_1) = E(X_1) =$$

50

Exam Example

$$Var(\bar{X}) = Var\left(\frac{X_1+X_2+\dots X_n}{n}\right) = \frac{1}{n^2} Var(X_1+X_2+\dots X_n)$$

$$= \frac{1}{n^2} \cdot n \cdot Var(X_1) = \frac{Var(X_1)}{n}$$

$$= \frac{5}{400}$$

$$\bar{X} \sim Normal\left(50, \frac{5}{400}\right)$$

$$P(\bar{X} > 72.5) = \underbrace{1 - P(\bar{X} \le 72.5)}$$

$$Y = \frac{X - E(X)}{SD(X)} = \frac{X - 50}{\sqrt{5/400}} = \frac{(X-50)20}{\sqrt{5}}$$

$$= 1 - P\left(Y < \frac{72.5-50}{\sqrt{5/400}}\right) \qquad Y \sim \text{Normal}(0,1)$$

$$= 1 - \Phi\left(\frac{72.5-50}{\sqrt{5/400}}\right)$$

Polling Example

We want to estimate the true proportion **p** of students who like Olivia Rodrigo more than Joshua Bassett. We poll **n** people, for large **n**. How many people do we have to poll to get an estimate with accuracy **0.1** with **95%** ~~probability~~ confidence?

poll all n people → 1 if they like Olivia better
→ 0 otherwise

$$X_i \sim \text{Bernoulli}(p)$$

$$X = \frac{X_1 + X_2 + \dots X_n}{n} \rightarrow \text{proportion for our estimate} = \hat{p}$$

$-\infty \leftarrow -0.1 \qquad 0.1 \rightarrow \infty$

Polling Example

$$P(|\hat{p}-p| \geq 0.1) \leq 0.05$$

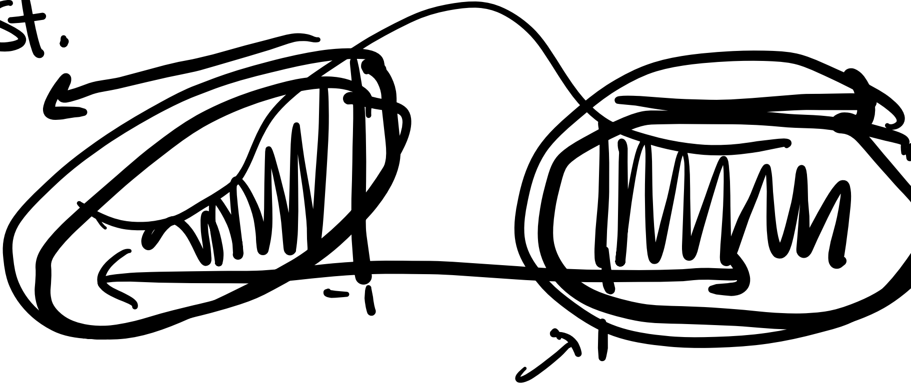
CLT

$\hat{p}$ is normally distributed

$\hat{p}-p$ is normally dist.

$P(1 \rightarrow \infty)$

$P(-1 \rightarrow -\infty)$



$$2P(\hat{p}-p \geq 0.1) \leq 0.05$$

$$2P(\hat{p} - p \geq 0.1) = 2(1 - P(\hat{p} - p \leq 0.1))$$

α

$$P(\hat{p} - p \leq 0.1)$$

$\hat{p} - p$ is normally dist. by CLT

$$\hat{p} - p \sim \text{Normal}\left(0, \frac{p(1-p)}{n}\right) \qquad E(\hat{p} - p) = E(\hat{p}) - p = p - p = 0$$

$$\leq \frac{1}{4n} \qquad \hat{p} = \frac{X_1 + X_2 + \dots X_n}{n}$$

$$E(\hat{p}) = \frac{1}{n} \cdot n \cdot E(X_1) = E(X_1) = p$$

$$Y \sim \frac{X - E(X)}{SD(X)}$$

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

$$\text{Var}\left(\frac{X_1 + X_2 + \dots X_n}{n}\right) = \frac{1}{n^2} \text{Var}(X_1 + X_2 + \dots X_n)$$

$$= \frac{1}{n^2} \cdot n \cdot \text{Var}(X_1)$$

$$= \frac{p(1-p)}{n} \leq \frac{1}{4n}$$

$$Y \sim \frac{X - 0}{\left(\sqrt{\frac{1}{4n}}\right)} =$$

$$Y = X \cdot 2\sqrt{n}$$

$$P(\hat{p} - p \leq 0.1) = P(Y \leq (0.1 - 0) 2\sqrt{n})$$

$$p(1-p) \leq \frac{1}{4} \qquad Y \leq 0.1 \cdot 2 \cdot \sqrt{n}$$

$$2(1 - \Phi(0.1 \cdot 2\sqrt{n}))$$

$$2 - 2\Phi(0.1 \cdot 2\sqrt{n})$$

$$2 - 2\Phi(0.1 \cdot 2\sqrt{n}) \le 0.05$$

$$\Phi(0.1 \cdot 2\sqrt{n}) \ge 0.975$$

$$\Phi(1.96) = 0.975$$

$$\rightarrow 0.1 \cdot 2\sqrt{n} \ge 1.96$$

$$\rightarrow (0.1)^2 \cdot 4n \ge 1.96^2$$

$$n \ge \frac{1.96^2}{(0.1)^2 \cdot 4}$$

n as below below

Poisson Example

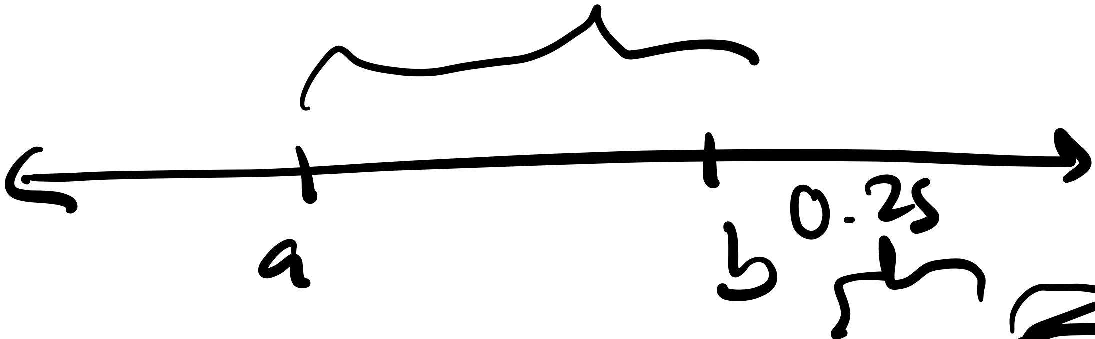
The number of times I listen to Olivia Rodrigo's *Vampire* is modeled by a Poisson(λ) distribution per day, but I don't know λ. All I know about λ is that it is at most 20. If I track the number of times I listen to *Vampire* each day, for a large number of days, how many days do I need to track to get a 95% confidence interval for λ of **width 3.**

$$X = \frac{X_1 + X_2 + \dots X_n}{n}$$

$$Var(X) = Var\left(\frac{X_1 + X_2 + \dots X_n}{n}\right) = \frac{1}{n^2} Var(X_1 + X_2 + \dots X_n) = \frac{1}{n^2} \cdot n \cdot Var(X_1) = \frac{Var(X_1)}{n} = \frac{\lambda}{n}$$

$b - a = 3$

$2 \cdot$ ___

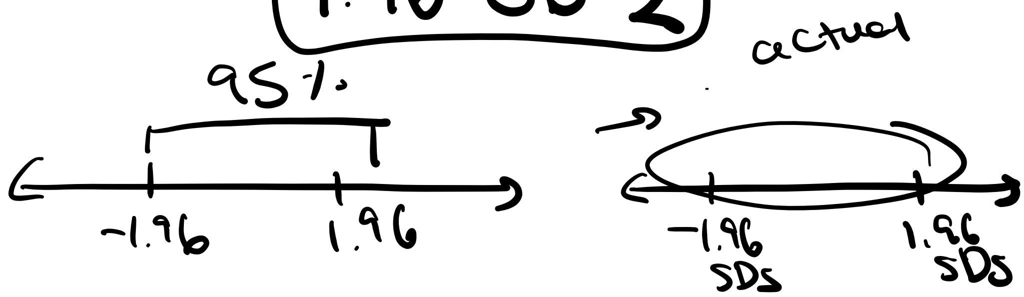


Standard lookup: 0.975 below it $\overset{+}{1.96}$ SDs $\overset{\downarrow}{1.96}$ SDs = 2·1.96 SDs

between −1.96 SD and 1.96 SDs

SD(X) calculate this

$$\boxed{1.96 \cdot SD \cdot 2}$$



$$2 \cdot 1.96 \cdot SD(X) = 3$$

$$Var(X) = \frac{\lambda}{n}$$

$$SD(X) = \sqrt{\frac{\lambda}{n}} \qquad 2 \cdot 1.96 \cdot \sqrt{\frac{\lambda}{n}} = 3$$

$$2 \cdot 1.96 \cdot \sqrt{\frac{20}{n}} = 3$$

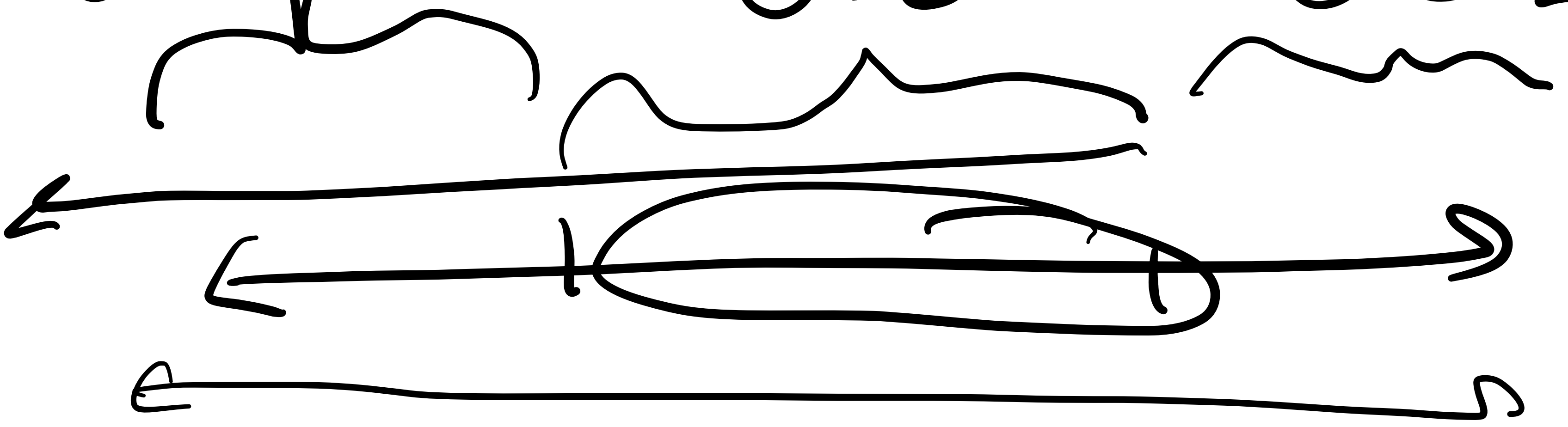
$$= 4 \cdot 1.96^2 \cdot \frac{20}{n} = 9$$

$$\boxed{n = \frac{4 \cdot 1.96^2 \cdot 20}{9}}$$

1.96 0 1.96

Poisson Example (cont.)

$0.2/2$ $\quad$ 0.8 $\quad$ $0.2/2$



$$1 - 0.8 = 0.2$$

80% confidence

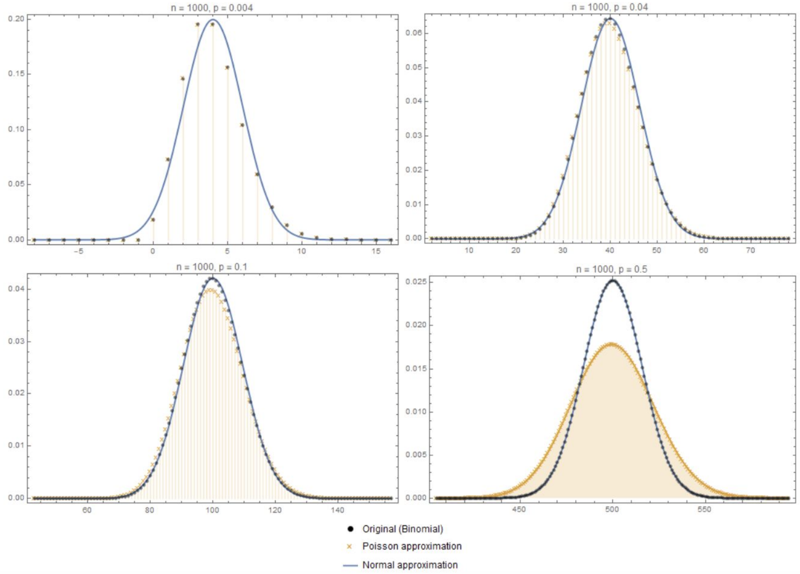
Poisson Example (cont.)

80% confidence

$$\Phi(x) = 0.8 + \frac{0.2}{2}$$
$$= 0.9$$

Normal as a Limiting Distribution of Binomial

When **n** is large, and **p** is not small, the normal distribution is a good approximation to the binomial. Source: https://math.stackexchange.com/questions/3278070/approximation-of-binomial-distribution-poisson-vs-normal-distribution.



(If time)
The following are slides that didn’t make it into recording for 6C so I will redo them if there is time, so that they can be in a recording.

Exponential Distribution

Def: For some λ > 0, a continuous random variable X is an exponential random variable with parameter λ if it has the following PDF,

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

We can write X ~ Exp(λ) if is an exponential random variable.

Let's check that f(x) satisfies the two PDF properties

Expectation & Variance of Exponential

For a random variable X ~ Exp(λ),

$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad \textit{and} \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

Example Exponential

Going back to our terrible alarm clock, we know the behavior is:

Once plugged in, the alarm will randomly ring once after some amount of time, however we know it goes off at a rate of 1 time every 10 minutes.

Let X be the amount of time it takes the alarm to sound. $X \sim \text{Exp}(1/10)$, because our rate of rings is 1/10 per minute.

How many minutes should we expect to wait before the alarm rings?

An important property to note

$$\mathbb{P}[X > t] = e^{-\lambda t}$$

Proof:

Exponential Relation to Geometric

Let's try to draw a more rigorous connection between Exponential and Geometric distributions. Take our Geometric r.v. and consider running trials after every d seconds. The probability of a success is p = λ**δ**.

Then, let Y denote the amount of time/seconds before a successful trial:

$$\mathbb{P}[Y > k\delta] = (1-p)^k = (1-\lambda\delta)^k,$$

To translate our trials to a continuum, consider taking the limit of d → 0. Then, for any time t,

$$\mathbb{P}[Y > t] = \mathbb{P}[Y > (\tfrac{t}{\delta})\delta] = (1-\lambda\delta)^{t/\delta} \approx e^{-\lambda t}$$

Recap

Discussed Gaussian (Normal) Random Variables

- Properties of Normal Distributions
- Scaling Normal Distributions
- Summing Independent Random Normal Variables
- The CDF of a Normal Distribution $\Phi(x)$
- Solving Problems using the **Central Limit Theorem (CLT)**