# Section 10: Law of large numbers and central limit theorem

## STOR 155.02, Spring '21

updated 2021-03-30

# What you will learn

- parameters and point estimates
- model intro
- law of large numbers
- central limit theorem

# Resources

- Textbook: 5.1, 5.4

# Parameters

## numbers that define a random variable's distribution

## examples:

**Binomial(N, p): N and p are parameters!**

**Normal(m, s): m and s are parameters!**

# If you know the parameters, you can

## calculate the probabilities of distribution

## sample from it

**using a computer**

## e.g. homework: input m and s to sample from Normal

# Modeling

## Abstractions of reality

In this class:

- a model is a probability distribution
- along with its parameters
- that describes a random process.

## Example

**Numbers of field goals made**

- for a known N number of shots
- assumed independent
- with success of shot p

**Q: What kind of distribution to use?**

## 'True' parameters are unknown!

## Use data to plug in estimates for them

# Point estimates and the law of large numbers

# A point estimate is a statistic we use to guess at an unknown parameter

$$\text{sample mean is a point estimate for the true mean}$$

$$\text{sample s.d. is a point estimate for the true s.d.}$$

# This works because of the law of large numbers

which says

> > if $X_1 \ldots X_n$ are independent (random) samples of a random variable X, then

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx E(X) \quad \text{for large n}$$

and the same for sample standard deviation vs. true trandard deviation

# Example: Bernoulli

**Say I have $n$ samples $X_1 \ldots X_n$ from a Bernoulli(p)**

**but p is unknown**

$$\hat{p} = \frac{\#\ \text{successes}}{n} \approx p$$

# Example: Binomial

**$n$ samples $X_1 \ldots X_n$ from a Binomial(N, p)**

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx Np$$

# Example: Modeling Paris Kea's 2017-18 field goals

Old data on 2017-18 season for UNC basketball player Paris Kea.

## Model

$$X = \text{ field goals made in a game with N attempts}$$

$$X \text{ modeled as Binomial(N, p)}$$

# Why does this model make sense for the data, or not?

# Point estimates for the model

**To make life easy: We will say N is known**

$$\hat{p} = \frac{\text{\# field goals made across all games}}{\text{\# field goals attempted across all games}}$$

## Q: Why does this make sense as a point estimate?

**Once we have a point estimate p**

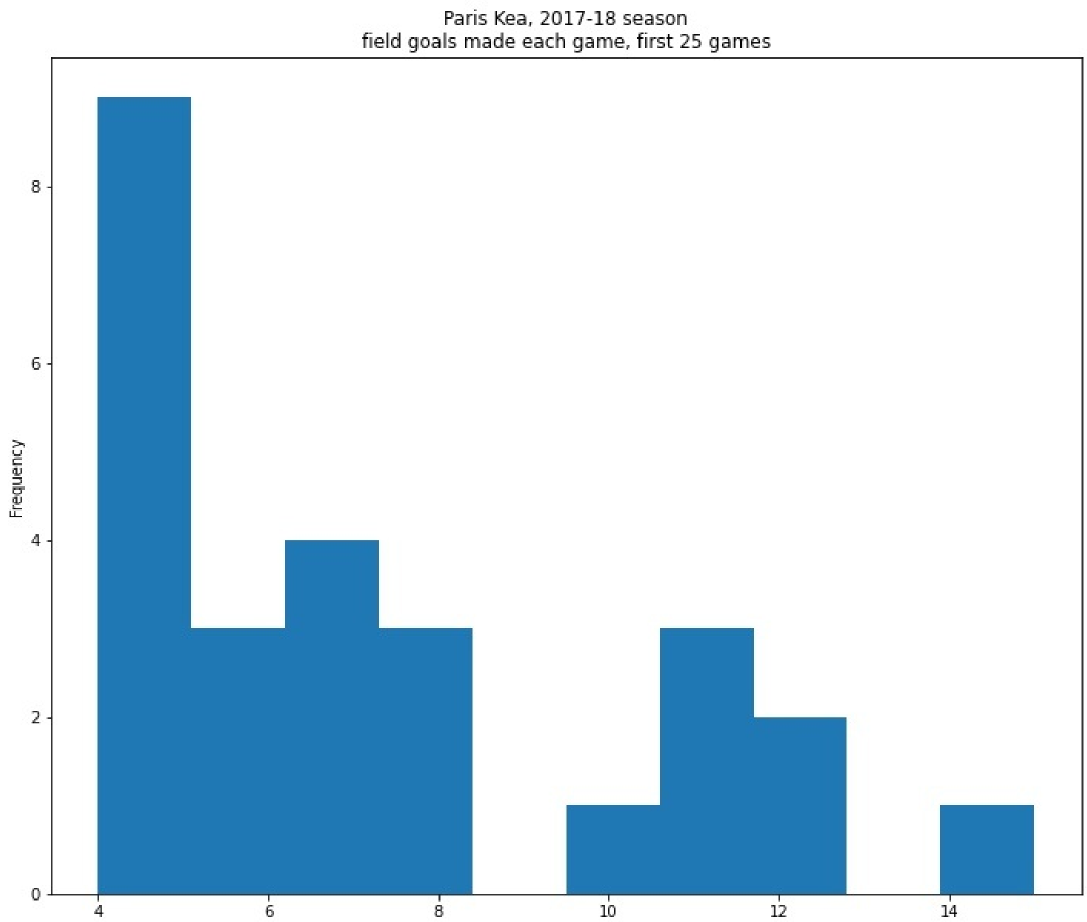
**we can predict the number of field goals Kea will make for any N attempts**

**and you can use sampling**

**to estimate how risky your predictions are**

**future homework!**

$$\text{predicted field goals} = N\hat{p}$$


Paris Kea, 2017-18 season
field goals made each game, first 25 games
Frequency
0
2
4
6
8
4
6
8
10
12
14

# Modeling procedure

1. Estimate $\hat{p}$ based on data from the first 25 games (the 'training data')

2. Predict outcomes for the next 6 games, for which you have data (the 'test data')

3. Evaluate how good your model is

**this is a common basic approach to predictive modeling**

| phat |
|---:|
| 0.4818653 |

$$\text{predicted field goals} = \hat{p} \times \text{fg\_attempts}$$

| | game | field_goals_made | fg_made_pred |
|---|---|---:|---:|
| 26 | gtech | 6 | 7.227979 |
| 27 | louisville | 5 | 5.782383 |
| 28 | syracuse | 5 | 7.227979 |
| 29 | duke | 3 | 5.300518 |
| 30 | boston college | 12 | 10.601036 |
| 31 | nc state | 9 | 11.082902 |

remember we are pretending we knew what $N$, number of field goal attempts, was for each game.

**How good is our prediction? Is it better or worse than what we should have expected?**

**A start: How close should $\hat{p}$ be to the true p?**

# Central limit theorem

# Central limit theorem

If $X_1 \ldots X_n$ are **independent** (random) samples of a random variable $X$ where

$$E(X) = m, \qquad Var(X) = s^2$$

then writing $\hat{m}_n = \frac{1}{n} \sum_1^n X_i$ for the **sample average**

$$\frac{\hat{m}_n - m}{s / \sqrt{n}} \approx \text{Normal}(0, 1) \quad \text{for large enough sample size n}$$

which is the same as saying

$$\hat{m}_n \approx \text{Normal}\left(m, \frac{s}{\sqrt{n}}\right) \quad \text{for large enough sample size n}$$

**in words**

> the distribution of the sample mean for a sample of size n (written $\hat{m}_n$) is approximately normal, with the **same expected value** as the original $X$ you sampled from and standard deviation of the original $X$ divided by $\sqrt{n}$

# This is how I feel about it

# Why is the CLT so amazing?

## 1. At the heart of much of science for the past 300 years

It's why we take averages of measurements.

## 2. Basis for most statistical tests of 'significance' (future lectures)

and therefore of statistical estimation in economics, business, public policy, physical sciences, medical studies ...

## 3. Shows up in magical ways throughout mathematics

## Key assumptions

**1. Independent samples**

**2. $E(X)$ and $Var(X)$ are finite**

second assumption is always true in this class

**it works for any distribution under those assumptions**

## But how big does n have to be?

**exact answer needs math beyond this course**

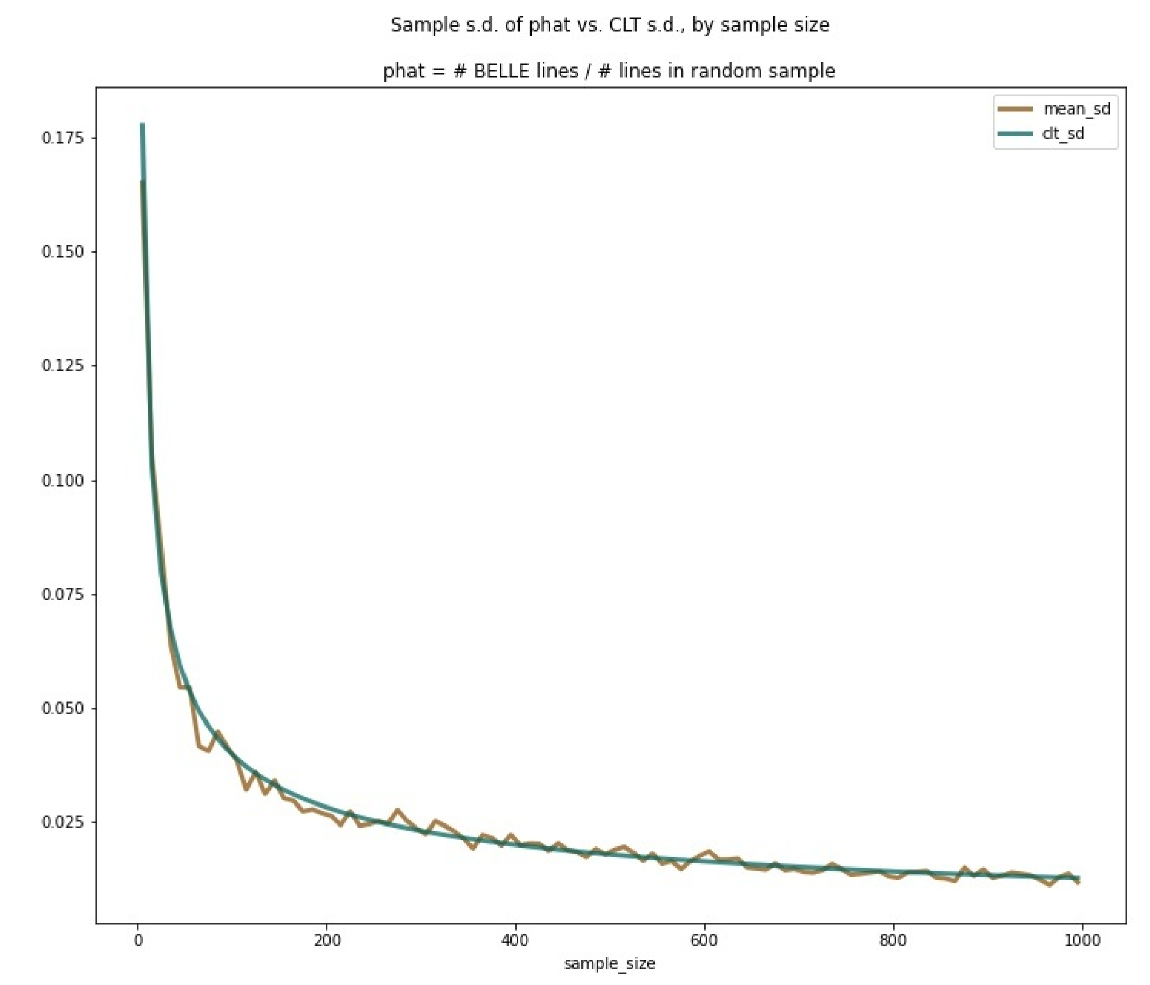
**rough answer**

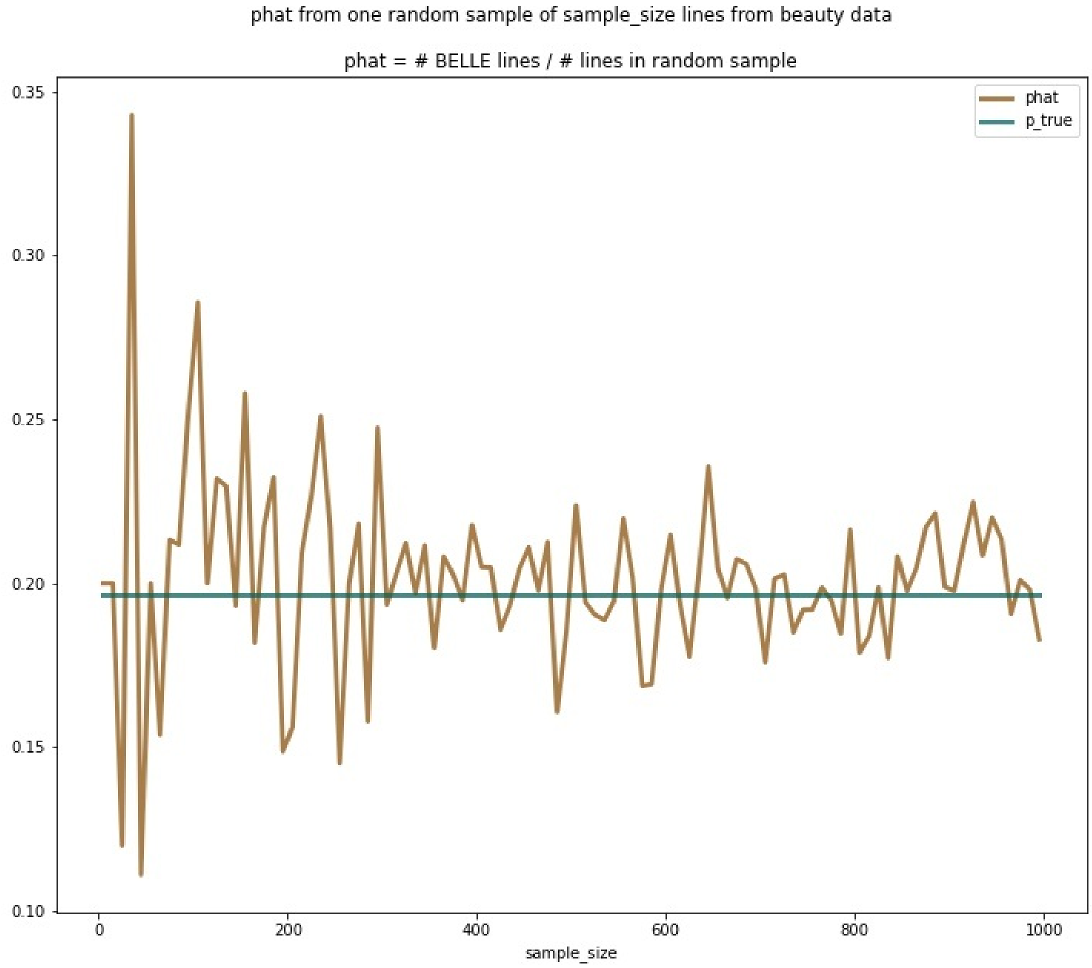
> > you need **larger n the more skewed** the sample distribtion is

skewed meaning the histogram is lopsided on the left or right

**play with this demo**

CLT demo


Sample s.d. of phat vs. CLT s.d., by sample size
phat = # BELLE lines / # lines in random sample
mean_sd
clt_sd
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0
200
400
600
800
1000
sample_size


phat from one random sample of sample_size lines from beauty data
phat = # BELLE lines / # lines in random sample
phat
p_true
0.35
0.30
0.25
0.20
0.15
0.10
0
200
400
600
800
1000
sample_size

# Example: Using the CLT

A simplified version of the Paris Kea modeling example, to avoid dealing with different numbers of field goal attempts.

## Setup

Think of our estimated field goal percentage now as estimated *number* of field goals made in $N = 20$ attempts.

Assume her field goal stats across 25 games are **independent.**

$$\hat{m}_{25} = 20 \times \hat{p}$$

$$\approx 9.64$$

$$\hat{s} = \sqrt{20 \times \hat{p}(1 - \hat{p})}$$

## Q: Why are these my point estimates for the true parameters $m$ and $s$?

# using the CLT

**We don't know what the true $p$ is.**

**can use the CLT and LLN to see how spread out from the true value our estimate should be**

$$\sqrt{25} \times \frac{\hat{m}_{25} - 20p}{\sqrt{20p(1-p)}} \approx 5 \times \frac{\hat{m}_{25} - 20p}{\sqrt{20\hat{p}(1-\hat{p})}}$$

$$\approx \text{Normal}(0, 1)$$

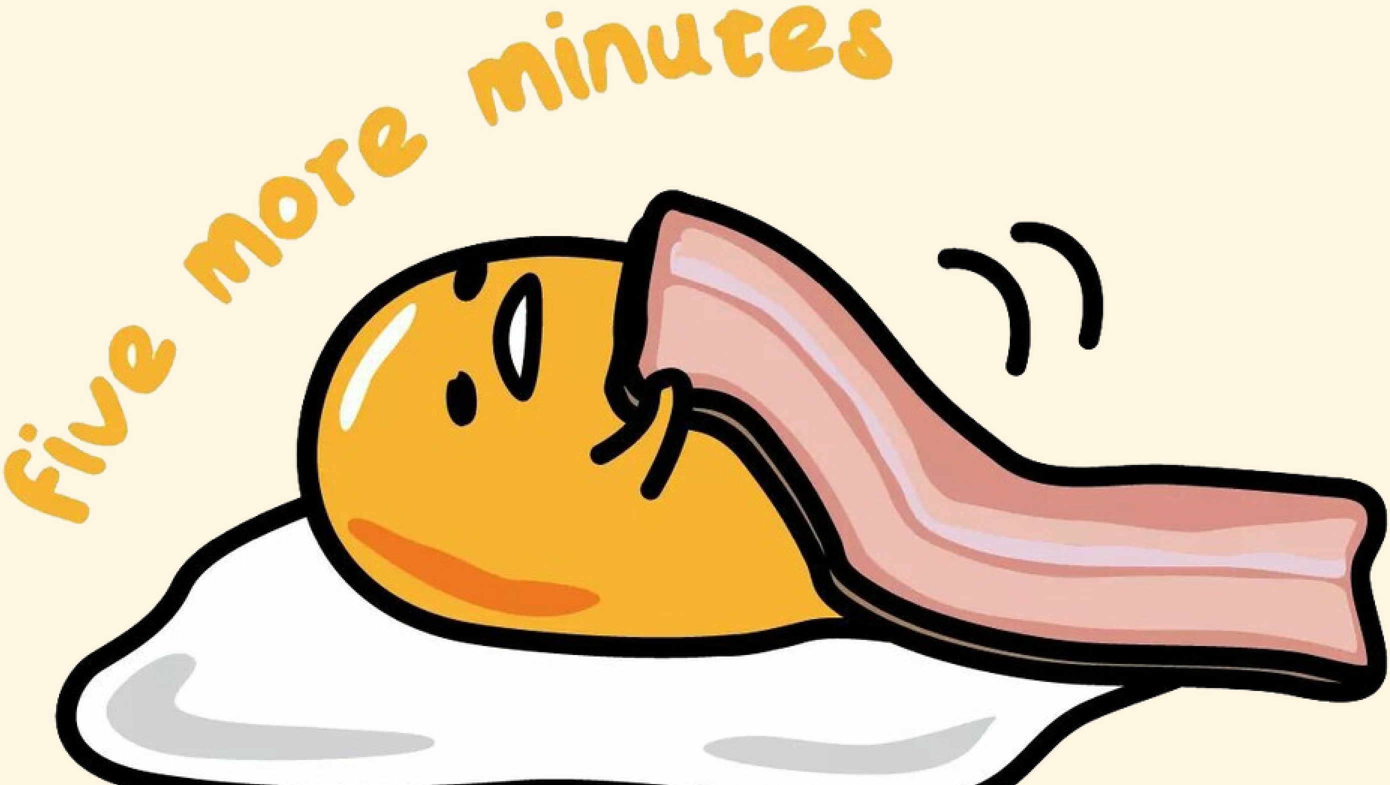
**will make this precise next class**

# Q: Assuming the CLT normal approximation is correct, how would I calculate

$$P\left(\hat{m}_{25} \text{ is within two standard deviations of the true m}\right)$$

PollEv.com/brendanbrown849

poll closes at __________


five more minutes