Central Limit Theorem & Preview of Markov Chains

Math 20

August 3, 2012

Announcements

- Homework 6 has been posted and is due on Wednesday, August 8.
- Midterm Exam is Thursday, August 9 from 4-6pm in Moore Hall Filene Auditorium. (If you have a conflict you must let me know by Monday!!)
- Monday and Wednesday of next week, we have a guest speaker. The lectures will be recorded.
- Tuesday will be an (optional) Matlab tutorial session during x-hour.

The Four Versions of the Central Limit Theorem:

Theorem (1)

(Central Limit Theorem for the Binomial Distribution) For the binomial distribution $b(n, p, j)$*, we have*

$$\lim_{n\to\infty} \sqrt{npq} \cdot b(n, p, \langle np + x\sqrt{npq}\rangle) = \varphi(x).$$

The Four Versions of the Central Limit Theorem:

Theorem (1)

(Central Limit Theorem for the Binomial Distribution) For the binomial distribution $b(n, p, j)$, we have

$$\lim_{n\to\infty} \sqrt{npq} \cdot b(n, p, \langle np + x\sqrt{npq}\rangle) = \varphi(x).$$

Notice that stated another way (that may be easier to remember), this says:

$$\lim_{n\to\infty} \sqrt{npq} \cdot b(n, p, j) = \varphi(j^*)$$

where

$$j^* = \frac{j - np}{\sqrt{npq}}.$$

The Four Versions of the Central Limit Theorem:

What kind of questions is this version of the theorem meant to answer?

The Four Versions of the Central Limit Theorem:

What kind of questions is this version of the theorem meant to answer?

- Estimate the probability of rolling exactly 15 sixes when you roll a die 100 times.

The Four Versions of the Central Limit Theorem:

What kind of questions is this version of the theorem meant to answer?

- Estimate the probability of rolling exactly 15 sixes when you roll a die 100 times.
- Estimate the probability of getting exactly 480 heads when you flip a coin 1000 times.

The Four Versions of the Central Limit Theorem:

What kind of questions is this version of the theorem meant to answer?

- Estimate the probability of rolling exactly 15 sixes when you roll a die 100 times.
- Estimate the probability of getting exactly 480 heads when you flip a coin 1000 times.
- Estimate the probability that a box with 120 times has exactly 6 defective items if the probability that a single item is defective is $p = 0.05$.

The Four Versions of the Central Limit Theorem:

What kind of questions is this version of the theorem meant to answer?

- Estimate the probability of rolling exactly 15 sixes when you roll a die 100 times.
- Estimate the probability of getting exactly 480 heads when you flip a coin 1000 times.
- Estimate the probability that a box with 120 times has exactly 6 defective items if the probability that a single item is defective is p = 0.05.
- In general, used to estimate the probability of j successes in n Bernoulli trials where p is the probability of success, where n is large.

The Four Versions of the Central Limit Theorem:

Theorem (2)

(Central Limit Theorem for Bernoulli Trials) Let S_n be the number of successes in n Bernoulli Trials with probability p for success and let a and b be two fixed real numbers. Then

$$\lim_{n\to\infty} P(a \leq S_n \leq b) = \lim_{n\to\infty} P(a^* \leq S_n^* \leq b^*) = \int_{a^*}^{b^*} \varphi(x)\, dx$$

where

$$a^* = \frac{a - np}{\sqrt{npq}}$$

and

$$b^* = \frac{b - np}{\sqrt{npq}}.$$

The Four Versions of the Central Limit Theorem:

This version is used to answer questions like:

- Estimate the probability of rolling more than 15 sixes when you roll a die 100 times.

The Four Versions of the Central Limit Theorem:

This version is used to answer questions like:

- Estimate the probability of rolling more than 15 sixes when you roll a die 100 times.
- Estimate the probability of getting between 480 and 520 heads when you flip a coin 1000 times.

The Four Versions of the Central Limit Theorem:

This version is used to answer questions like:

- Estimate the probability of rolling more than 15 sixes when you roll a die 100 times.
- Estimate the probability of getting between 480 and 520 heads when you flip a coin 1000 times.
- What is the probability that a 100-seat plane is overbooked if 105 tickets were sold and the probability a person shows up for the flight is 0.9? That is, what is the probability that more than 100 people show up?

The Four Versions of the Central Limit Theorem:

This version is used to answer questions like:

- Estimate the probability of rolling more than 15 sixes when you roll a die 100 times.
- Estimate the probability of getting between 480 and 520 heads when you flip a coin 1000 times.
- What is the probability that a 100-seat plane is overbooked if 105 tickets were sold and the probability a person shows up for the flight is 0.9? That is, what is the probability that more than 100 people show up?
- In general you want to know the probability that the number of successes in n Bernoulli trial lies in some interval.

The Four Versions of the Central Limit Theorem:

Also, used to answer questions about polling and hypothesis testing:

- If you interview 1000 people and 600 of them express support a ballot initiative, what is the 95% confidence interval for the actual proportion of people in the population who support this initiative?

The Four Versions of the Central Limit Theorem:

Also, used to answer questions about polling and hypothesis testing:

- If you interview 1000 people and 600 of them express support a ballot initiative, what is the 95% confidence interval for the actual proportion of people in the population who support this initiative?
- If a newspaper claims that 30% of people are afraid of heights and you interview a random sample of size 2000 and 800 of them tell you that they are afraid of heights, do you reject the hypothesis put forth by the newspaper?

The Four Versions of the Central Limit Theorem:

Theorem (3)

(Central Limit Theorem for Discrete Independent Trials) Let $S_n = X_1 + X_2 + \cdots + X_n$ be the sum of n discrete independent identically distributed random variables with parameters $\mu = E(X_i)$ and $\sigma^2 = V(X_i)$. Then

$$\lim_{n\to\infty} P(a \leq S_n \leq b) = \lim_{n\to\infty} P(a^* \leq S_n^* \leq b^*) = \int_{a^*}^{b^*} \varphi(x)\,dx$$

where

$$a^* = \frac{a - n\mu}{\sqrt{n\sigma^2}}$$

and

$$b^* = \frac{b - n\mu}{\sqrt{n\sigma^2}}.$$

The Four Versions of the Central Limit Theorem:

Also notice that Version 2 of the Central Limit Theorem is just a special case of Version 3!

To show this, just notice that in Version 2, each X_i is a Bernoulli trial where $X_i = 1$ with probability p and $X_i = 0$ with probability $q = 1 - p$. Then

$$\mu = p, \, \sigma^2 = pq$$

and after plugging in these values, we get exactly the statement of Version 2 of the Central Limit Theorem.

The Four Versions of the Central Limit Theorem:

Version 3 is used to answer questions like:

- Estimate the probability that you get a sum between 3500 and 3600 when you roll a die 1000 times.

The Four Versions of the Central Limit Theorem:

Version 3 is used to answer questions like:

- Estimate the probability that you get a sum between 3500 and 3600 when you roll a die 1000 times.
- Estimate the probability that don't lose any money when playing roulette (or any other gambling game) if you play 100 times.

The Four Versions of the Central Limit Theorem:

Version 3 is used to answer questions like:

- Estimate the probability that you get a sum between 3500 and 3600 when you roll a die 1000 times.
- Estimate the probability that don't lose any money when playing roulette (or any other gambling game) if you play 100 times.
- In general, estimate the probability that S_n lies within some interval, where the X_i's are not necessarily Bernoulli trials.

The Four Versions of the Central Limit Theorem:

Theorem (4)

(General Version of the Central Limit Theorem) Let $X_1, X_2, \cdots, X_n$, be a sequence of independent discrete random variables (that are not necessarily identically distributed!) and $S_n = X_1 + X_2 + \cdots + X_n$. Define $E(X_i) = \mu_i$, $V(X_i) = \sigma_i^2$ (these could all be different for different i). Denote the mean of S_n to be $m_n = E(S_n) = \mu_1 + \mu_2 + \cdots + \mu_n$. Denote the variance of S_n to be $s_n^2 = V(S_n) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$ and assume that $s_n \to \infty$ as $n \to \infty$. If there exists a constant A such that $|X_n| \leq A$ for all n, then

$$\lim_{n\to\infty} P(a \leq \frac{S_n - m_n}{s_n} \leq b) = \int_a^b \varphi(x)\, dx.$$

The Four Versions of the Central Limit Theorem:

Theorem (4)

(General Version of the Central Limit Theorem) Let $X_1, X_2, \cdots, X_n$, *be a sequence of independent discrete random variables (that are not necessarily identically distributed!) and* $S_n = X_1 + X_2 + \cdots + X_n$. *Define* $E(X_i) = \mu_i$, $V(X_i) = \sigma_i^2$ *(these could all be different for different* i*). Denote the mean of* S_n *to be* $m_n = E(S_n) = \mu_1 + \mu_2 + \cdots + \mu_n$. *Denote the variance of* S_n *to be* $s_n^2 = V(S_n) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$ *and assume that* $s_n \to \infty$ *as* $n \to \infty$. *If there exists a constant* A *such that* $|X_n| \leq A$ *for all* n, *then*

$$\lim_{n \to \infty} P(a \leq \frac{S_n - m_n}{s_n} \leq b) = \int_a^b \varphi(x)\, dx.$$

The Four Versions of the Central Limit Theorem:

Theorem (4)

(General Version of the Central Limit Theorem) Let $X_1, X_2, \cdots, X_n$, be a sequence of independent discrete random variables (that are not necessarily identically distributed!) and $S_n = X_1 + X_2 + \cdots + X_n$. Define $E(X_i) = \mu_i$, $V(X_i) = \sigma_i^2$ (these could all be different for different i). Denote the mean of S_n to be $m_n = E(S_n) = \mu_1 + \mu_2 + \cdots + \mu_n$. Denote the variance of S_n to be $s_n^2 = V(S_n) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$ and assume that $s_n \to \infty$ as $n \to \infty$. If there exists a constant A such that $|X_n| \leq A$ for all n, then

$$\lim_{n\to\infty} P(a \leq \frac{S_n - m_n}{s_n} \leq b) = \int_a^b \varphi(x)\, dx.$$

The Four Versions of the Central Limit Theorem:

Theorem (4)

(General Version of the Central Limit Theorem) Let $X_1, X_2, \cdots, X_n$*, be a sequence of independent discrete random variables (that are not necessarily identically distributed!) and* $S_n = X_1 + X_2 + \cdots + X_n$*. Define* $E(X_i) = \mu_i$*,* $V(X_i) = \sigma_i^2$ *(these could all be different for different* i*). Denote the mean of* S_n *to be* $m_n = E(S_n) = \mu_1 + \mu_2 + \cdots + \mu_n$*. Denote the variance of* S_n *to be* $s_n^2 = V(S_n) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$ *and assume that* $s_n \to \infty$ *as* $n \to \infty$*. If there exists a constant* A *such that* $|X_n| \leq A$ *for all* n*, then*

$$\lim_{n\to\infty} P(a \leq \frac{S_n - m_n}{s_n} \leq b) = \int_a^b \varphi(x)\, dx.$$

The Four Versions of the Central Limit Theorem:

Theorem (4)

(General Version of the Central Limit Theorem) Let $X_1, X_2, \cdots, X_n$, be a sequence of independent discrete random variables (that are not necessarily identically distributed!) and $S_n = X_1 + X_2 + \cdots + X_n$. Define $E(X_i) = \mu_i$, $V(X_i) = \sigma_i^2$ (these could all be different for different i). Denote the mean of S_n to be $m_n = E(S_n) = \mu_1 + \mu_2 + \cdots + \mu_n$. Denote the variance of S_n to be $s_n^2 = V(S_n) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$ and assume that $s_n \to \infty$ as $n \to \infty$. If there exists a constant A such that $|X_n| \leq A$ for all n, then

$$\lim_{n\to\infty} P(a \leq \frac{S_n - m_n}{s_n} \leq b) = \int_a^b \varphi(x)\, dx.$$

The Four Versions of the Central Limit Theorem:

Theorem (4)

(General Version of the Central Limit Theorem) Let $X_1, X_2, \cdots, X_n$, *be a sequence of independent discrete random variables (that are not necessarily identically distributed!) and* $S_n = X_1 + X_2 + \cdots + X_n$. *Define* $E(X_i) = \mu_i$, $V(X_i) = \sigma_i^2$ *(these could all be different for different* i*). Denote the mean of* S_n *to be* $m_n = E(S_n) = \mu_1 + \mu_2 + \cdots + \mu_n$. *Denote the variance of* S_n *to be* $s_n^2 = V(S_n) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$ *and assume that* $s_n \to \infty$ *as* $n \to \infty$. *If there exists a constant* A *such that* $|X_n| \leq A$ *for all* n, *then*

$$\lim_{n\to\infty} P(a \leq \frac{S_n - m_n}{s_n} \leq b) = \int_a^b \varphi(x)\,dx.$$

The Four Versions of the Central Limit Theorem:

Theorem (4)

(General Version of the Central Limit Theorem) Let $X_1, X_2, \cdots, X_n$, be a sequence of independent discrete random variables (that are not necessarily identically distributed!) and $S_n = X_1 + X_2 + \cdots + X_n$. Define $E(X_i) = \mu_i$, $V(X_i) = \sigma_i^2$ (these could all be different for different i). Denote the mean of S_n to be $m_n = E(S_n) = \mu_1 + \mu_2 + \cdots + \mu_n$. Denote the variance of S_n to be $s_n^2 = V(S_n) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$ and assume that $s_n \to \infty$ as $n \to \infty$. If there exists a constant A such that $|X_n| \leq A$ for all n, then

$$\lim_{n\to\infty} P(a \leq \frac{S_n - m_n}{s_n} \leq b) = \int_a^b \varphi(x)\, dx.$$

The Four Versions of the Central Limit Theorem:

Why is this significant?

The Four Versions of the Central Limit Theorem:

Why is this significant?

Essentially anything that can be thought of as being made up of as the sum of many small independent (or weakly dependent) pieces is approximately normal.

For example: Height, Weight, IQ, Blood Pressure, time it takes to run a mile, and other traits that are governed by multiple genetic and environmental factors. Other examples are errors in measurements, measurements of natural phenomenon, and SAT scores.

The Four Versions of the Central Limit Theorem:

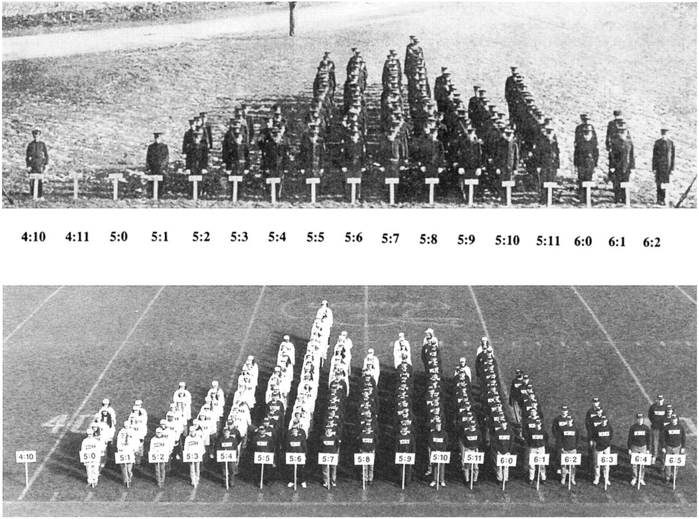


Figure: Living Histogram

The Four Versions of the Central Limit Theorem:

Height –

- Certainly has some genetic factors – If your parents are tall, there's a better chance you will be tall as well.
- However, it isn't the case that there is a "tall" gene. In reality, there are many factors (genetic and environmental) which contribute to your height.

The Four Versions of the Central Limit Theorem:

Height –

- Certainly has some genetic factors – If your parents are tall, there's a better chance you will be tall as well.
- However, it isn't the case that there is a "tall" gene. In reality, there are many factors (genetic and environmental) which contribute to your height. No single one of these is overwhelming.

The Four Versions of the Central Limit Theorem:

Height –

- Certainly has some genetic factors – If your parents are tall, there's a better chance you will be tall as well.
- However, it isn't the case that there is a "tall" gene. In reality, there are many factors (genetic and environmental) which contribute to your height. No single one of these is overwhelming.
- Clearly, the probability distributions associated with each factor will not be the same.

The Four Versions of the Central Limit Theorem:

We can let $X_1, X_2, \cdots, X_n$ represent each factor's contribution to your height. Some contribute more than others. Some may be negative (make you shorter). Then our theorem suggests that $S_n = X_1 + \cdots + X_n$, which represents height is normally distributed.

The Four Versions of the Central Limit Theorem:

We can let $X_1, X_2, \cdots, X_n$ represent each factor's contribution to your height. Some contribute more than others. Some may be negative (make you shorter). Then our theorem suggests that $S_n = X_1 + \cdots + X_n$, which represents height is normally distributed.

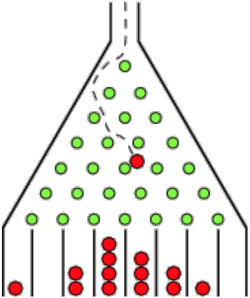


Figure: Galton Board

The Four Versions of the Central Limit Theorem:

So most people will land in the middle!

Galton: “The beautiful regularity in the structures of a population, whenever they are statistically marshaled in order of their heights, is due to the number of variable and quasi-independent elements of which Structure is the sum.”

The Four Versions of the Central Limit Theorem:

Any Questions??

Markov Chains

On Monday, we'll start Chapter 11 on Markov Chains!

Markov chains will give a way of studying a DEPENDENT trial process. (So the likelyhood of an event depends on what happened last.)

Markov Chains

On Monday, we'll start Chapter 11 on Markov Chains!

Markov chains will give a way of studying a DEPENDENT trial process. (So the likelyhood of an event depends on what happened last.)

Applications:

- weather
- genetics, neuroscience
- physics (thermodynamics), chemistry (enzyme activity), economics (dynamic macroeconomics, stock prices)
- Google (PageRank)
- Games (Monopoly, Chutes and Ladders, baseball)
- music
- cryptography

Markov Chains

Here is an example from Stanford's Statistics Department. A psychologist from the state prison statement showed up with a collection of coded messages...

Markov Chains

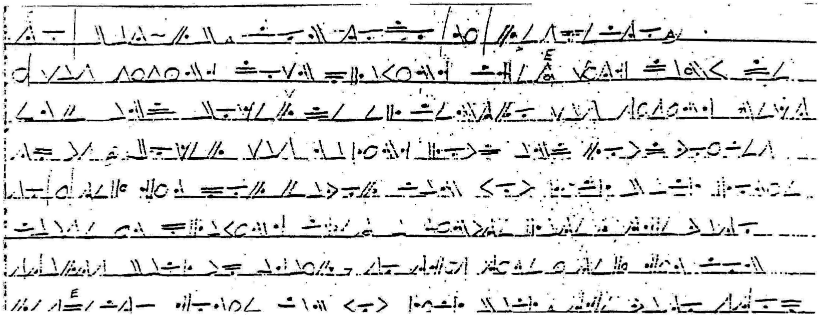


Figure: Encoded Message

Markov Chains

How to decode?

Markov Chains

How to decode?

In this case, it was safe to assume that the code was a simple substitution cipher (so each symbol stands for a letter, number, punctuation mark or space). What would you do?

Markov Chains

How to decode?

In this case, it was safe to assume that the code was a simple substitution cipher (so each symbol stands for a letter, number, punctuation mark or space). What would you do?

Could try to make a substitution and see if it works.

Markov Chains

How to decode?

In this case, it was safe to assume that the code was a simple substitution cipher (so each symbol stands for a letter, number, punctuation mark or space). What would you do?

Could try to make a substitution and see if it works. But even if you are only working with letters, how many possible substitutions are there?

Markov Chains

How to decode?

In this case, it was safe to assume that the code was a simple substitution cipher (so each symbol stands for a letter, number, punctuation mark or space). What would you do?

Could try to make a substitution and see if it works. But even if you are only working with letters, how many possible substitutions are there? $26!$

Markov Chains

How to decode?

In this case, it was safe to assume that the code was a simple substitution cipher (so each symbol stands for a letter, number, punctuation mark or space). What would you do?

Could try to make a substitution and see if it works. But even if you are only working with letters, how many possible substitutions are there? 26! That's huge! More than the number of stars in the universe! It would take a very long time.

Markov Chains

As an alternative, one could see which symbol appears the most often and replace that with “e” (the most common letter). Then replace the second most common symbol with the second most common letter and so on....

Markov Chains

As an alternative, one could see which symbol appears the most often and replace that with “e” (the most common letter). Then replace the second most common symbol with the second most common letter and so on....

However, this has been shown to not work, especially if what you’re trying to decipher is short.

Markov Chains

So what you should look at instead is the relationship between letters. How likely is it that one specific letter is followed by another?

For example, “q” is almost always followed by “u”.

To determine these probabilities, the people at Stanford used *War and Peace* to estimate the probability that one letter is followed by another.

Markov Chains

They started with some random substitution and started making random transpositions of letters and then used this information to determine what the best substitution should be.

Markov Chains

They started with some random substitution and started making random transpositions of letters and then used this information to determine what the best substitution should be.

For example:

GCN

Markov Chains

They started with some random substitution and started making random transpositions of letters and then used this information to determine what the best substitution should be.

For example:

GCN

GCH

Markov Chains

They started with some random substitution and started making random transpositions of letters and then used this information to determine what the best substitution should be.

For example:

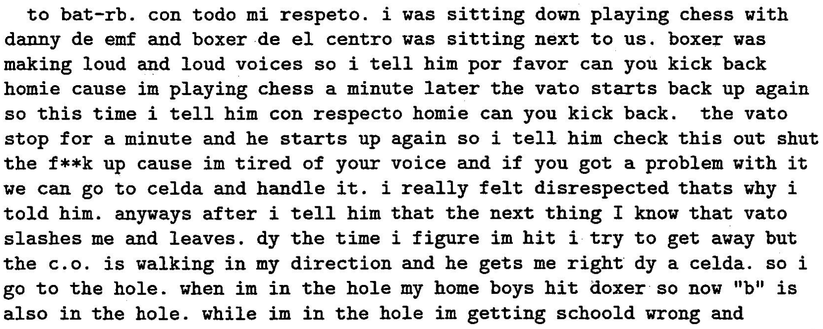
GCN

GCH

Is CN more common or is CH? If CH is more common, we accept this transposition. If it is not more likely, flip a coin. If heads, accept it. If tails, reject it and stay with CN.

Markov Chains

We continue this process as many times as necessary (maybe a few thousand times or more) and we get...



to bat-rb. con todo mi respeto. i was sitting down playing chess with danny de emf and boxer de el centro was sitting next to us. boxer was making loud and loud voices so i tell him por favor can you kick back homie cause im playing chess a minute later the vato starts back up again so this time i tell him con respecto homie can you kick back. the vato stop for a minute and he starts up again so i tell him check this out shut the f**k up cause im tired of your voice and if you got a problem with it we can go to celda and handle it. i really felt disrespected thats why i told him. anyways after i tell him that the next thing I know that vato slashes me and leaves. dy the time i figure im hit i try to get away but the c.o. is walking in my direction and he gets me right dy a celda. so i go to the hole. when im in the hole my home boys hit doxer so now "b" is also in the hole. while im in the hole im getting schoold wrong and

Figure: Deciphered Message

Markov Chains

Worked even though there is a mix of English, Spanish, and prison jargon.

Markov Chains

Worked even though there is a mix of English, Spanish, and prison jargon.

Takeaway Message: Markov chains are really useful! Plus, there are really interesting applications to the real world (we'll get to see a few more as we study them).