Statistics,
Law of Large Numbers

Fredo Durand & Ana Bell
MIT EECS 6.00
CAN MY BOYFRIEND COME ALONG?

I'M NOT YOUR BOYFRIEND!

I'M CASUALLY DATING A NUMBER OF PEOPLE.

BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.

YOUR MATH IS IRREFUTABLE.

FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.
Plotting
Pylab

- plotting functions for python
- best reference: textbook
- Central functions
  - `pylab.plot(x_list, y_list)`
  - `pylab.hist(data, numbin)`
- Formatting functions (title(), xlabel(), ylabel(), ...)
- Need to call `pylab.show()`
import pylab

def drunkTestP(numTrials = 20):
    meanDistances = []
    stepsTaken = [10, 100, 1000, 10000]
    for numSteps in stepsTaken:
        distances = simWalks(numSteps, numTrials)
        meanDistances.append(sum(distances)/len(distances))
    pylab.plot(stepsTaken, meanDistances)
    pylab.title('Distance from Origin in Steps')
    pylab.xlabel('Steps Taken')
    pylab.ylabel('Mean Steps from Origin')
    pylab.show()
Plot

Drunkards Random Walk

Distance from origin

Number of steps
Label your axes

I think we should give it another shot.

We should break up, and I can prove it.

Our relationship

Huh.

Maybe you're right.

I knew data would convince you.

No, I just think I can do better than someone who doesn't label her axes.
Cool Monte-Carlo simulation
Election prediction

Monte-Carlo Ray Tracing

- e.g. http://www.mitsuba-renderer.org/
  http://madebyevan.com/webgl-path-tracing/
  http://renderer.ivank.net/
Ray tracing
Random walk in path space
Statistics
Statistics

Science of data

gathering

analysis

presentation

Truth despite/with uncertainty
Statistics

• Science of data
  – gathering
  – analysis
  – presentation

• Extracting truth with controlled uncertainty
Data

- Scientific experiments
- Medical trials
- Stochastic simulation samples
- DNA
- Stock market
- Weather
- Traffic
- Student grades
- Sales, inventory
- Movie rating by viewers
- Music characteristics
- MOOC usage
- Votes, election poll
- User study
- Energy consumption
- Sports information
- Economics
- Crops
- e.g. https://www.data.gov/
Descriptive vs. inferential stats

**Descriptive**
- average, standard deviation of quiz
- summarize big data

**Inferential**
- small → big
- big → small
Descriptive vs. inferential stats

• Descriptive
  – from big data to small summaries
  – e.g. test score average and standard deviation, histogram

• Inference
  – from small data to conclusions about big populations
  – e.g. deduce drug efficiency from small trial
Big questions

Can we measure precision with small samples? How many samples?

Can we estimate if a drug is working based on small samples?
Big questions

• Can we estimate quantities over the whole population by sampling?
• Can we get a margin of error?
• How many samples do we need?
• Can we test the effectiveness of a treatment from a small trial?
• Can we compare conditions that are mutually exclusive (with/without drug)
Inferential statistics

Full population
- finite set of individuals
- infinite set (e.g. light rays)
- hypotheticals

Sample
small subset of population
hopefully representative
ideally purely random
Inferential statistics

• Full population: distribution of values/outcomes
  – finite but big population
  – hypothetical population/scenarios (with without medical treatment)
  – Or (infinite) stochastic process

• Measurement on small sample
  – polls, trial, Monte-Carlo sampling

• Infer something about full population
  – mean value

• Test hypothesis
  – is a drug working?
Inferential statistics tasks

From small sample estimate value + confidence margin of error

test by pothesis

Fit models regression

effect:
\[ y = ax + b \]
x: drug dose
From small Sample
- estimate value + confidence margin of error
- test hypothesis

Fit models: regression

\[ y_i \rightarrow y = a x + b \]

Predict: Machine learning
Inferential statistics tasks

• From (small) sample:
  - Estimate values with confidence interval
    - e.g. population mean
  - Test hypotheses
    - is drug working
  - Fit models
    - e.g. data \((x_i, y_i)\Rightarrow model \ y = ax + b\)
  - Predict
    - future outcome, new test data
This class

• Stochastic simulation to understand statistics
• Statistics to understand stochastic simulation
• probabilities to help with everything
Estimating the mean
Big question

- Can we estimate quantities over the whole population by sampling?
- Can we get a margin of error?
- How many samples do we need?
Fair die
Fair die

• Good model for lots of stuff

• e.g. imagine computing the ratio of males vs. females in a population by sampling
  – assuming we have a good sample
Monte-Carlo simulation

def simulate_one_coin():
    random.random() < 0.5
Monte-Carlo simulation

```python
import random
import pylab

def simulate_one_coin():
    return random.random() < 0.5

def simulate_n_coins(num_samples):
    count = 0
    for i in xrange(num_samples):
        if simulate_one_coin():
            count += 1
    return float(count) / num_samples

def print_n_coin_simu(num_samples):
    percent = simulate_n_coins(num_samples)
    print "average of ", num_samples, " coins is ", percent*100, "%"

print_n_coin_simu(10)
```
import random
import pylab

def simulate_one_coin():
    return random.random() < 0.5

def simulate_n_coins(num_samples):
    count = 0
    for i in xrange(num_samples):
        if simulate_one_coin():
            count += 1
    return float(count) / num_samples

def print_n_coin_simu(num_samples):
    percent = simulate_n_coins(num_samples)
    print "average of ", num_samples, "coins is", percent*100, "%"

print_n_coin_simu(10)
Message to the wise

import random
import pylab

def simulate_one_coin():
    return random.random() < 0.5

def simulate_n_coins(num_samples):
    count = 0
    for i in xrange(num_samples):
        if simulate_one_coin():
            count += 1
    return float(count) / num_samples

def print_n_coin_simu(num_samples):
    percent = simulate_n_coins(num_samples)
    print "average of ", num_samples, "coins is", percent*100, 

def print_n_coin_simu(10)

• Don’t forget to cast to float
  – use 0.0
  – use float()
Basic behavior

• 1 sample
  \[ \rightarrow \quad 0\%, \quad 100\% \]

• 2 samples
  \[ \rightarrow \quad 0\%, \quad \frac{50\%}{2x \text{ as much}}, \quad 100\% \]

• ...

• N samples

4 samples

0\%, 25\%, 50\%, 75\%, 100\%

small probability
Basic behavior

- 1 sample
- 2 samples
- ...
- N samples

\[\begin{align*}
\text{1 coin} & : \frac{1}{2} \\
\text{2 coins} & : \frac{1}{4} \\
\text{N coins} & : \left(\frac{1}{2}\right)^N
\end{align*}\]

When N increases, the probability to get 0% after N coins goes to 0.

\[\%\]

\[\begin{align*}
\text{50\%} & , \text{75\%} , \text{100\%}
\end{align*}\]
Behavior when N increases

```python
import random
import pylab

def simulate_one_coin():
    return random.random() < 0.5

def progression_of_average_of_N_coins(max_N):
    average_list = []
    count = 0
    for i in xrange(max_N):
        if simulate_one_coin():
            count += 1
            current_average = float(count) / (i+1)
        average_list.append(current_average)
    return average_list

def plot_coin_improvement(N):
    y_data = progression_of_average_of_N_coins(N)
    pylab.plot(range(1, N+1), y_data)
    pylab.title('coins flip average')
    pylab.xlabel('number of coins')
    pylab.ylabel('average')
    pylab.show()

plot_coin_improvement(3)
```
Law of large numbers
Informally

\# samples $\rightarrow \infty$

probe (wrong) $\rightarrow 0$
Informally

• When the number of samples increases the chance that the measured mean deviates from the true mean goes to zero
Law of large number
Law of large number

- aka Bernoulli’s law

- in repeated independent tests
  (flips in this case)
  with the same actual probability $p$
  of a particular outcome in each test
  (e.g., an actual probability of 0.5 of getting a head for each flip),
  the chance that the fraction of times that outcome occurs differs from $p$
  converges to zero
  as the number of trials goes to infinity
Business consequences
Gambler’s fallacy
The fallacy
The fallacy

• If the roulette has landed many times in a row on black, it’s more likely to land on white
  – kind of to get us back on track for the promises of the law of large numbers

• But remember, the samples are independent
  – The future couldn’t care less about the past
August 18, 1913

“at the casino in Monte Carlo, black came up a record twenty-six times in succession [in roulette]. … [There] was a near-panicky rush to bet on red, beginning about the time black had come up a phenomenal fifteen times. In application of the maturity [of the chances] doctrine, players doubled and tripled their stakes, this doctrine leading them to believe after black came up the twentieth time that there was not a chance in a million of another repeat. In the end the unusual run enriched the Casino by some millions of francs.”

Huff and Geis, How to Take a Chance, pp. 28-29.
Think about it

lands on black 25 times

$\frac{1}{2}$  \[ \frac{1}{30} \text{M} \]

lands on red after 25 blocks

50%