The Most Important Lecture of The Year

and the physically most dangerous for Frédo

Fredo Durand & Ana Bell
MIT EECS 6.00
The annual death rate among people who know that statistic is one in six.
Quiz

• email Emily about conflicts
Do read

THE CARTOON GUIDE TO

STATISTICS

LARRY GONICK
Author of The Cartoon History of the Universe

& WOOLLCOTT SMITH
For more

Central limit theorem
Central limit theorem (Laplace)

• Standard deviation falls off as $\frac{1}{\sqrt{N}}$

• Converge to Normal distribution
Variance for N samples

\[ \text{variance for 1 sample of a random experiment} \]

\[ \text{coins} : \frac{6^2}{0.25} = 6 = 0.5 \]

\[ \text{variance for N samples divide by} \ N \]
Variance for N samples

- Variance for N samples = Variance for 1 sample divided by N

- assuming it’s all random and independent
Big rule of thumb

Increase precision by 10

# samples × 100
Big rule of thumb

• To multiply precision by $k$
  – need $k^2$ times as many samples

  – e.g. 100 times more samples to increase precision by 10
Fuzzy central limit theorem
Fuzzy central limit theorem

• Effects that are influenced by many small and unrelated random effects are approximately normally distributed

• For example:
  – people’s height, weight
  – SAT scores
  – temperatures
Group size and variance

Small data and variance

- Smaller datasets have higher variance
  - hence can exhibit more extreme average values
  - be wary of any correlation between population and extreme outcomes (e.g. small schools, small counties, small states)

Today
Plan today

Tuesday: God’s view

Today back to Earth

N samples
we don’t know much about the full population
Plan today

• Tuesday:
  we know everything,
  let’s understand what happens to random samples

• Today:
  we only have a limited set of samples,
  what can we tell?
Confidence interval
Political poll

ask 1000 people

540 will vote for A

54%
Political poll

• Ask 1,000 people if they would vote for a candidate
  - 540 say yes

• Can we bound the true mean of the full population?
• i.e. say the true percentage is 54 +/- x%
Visually
Probabilistic answer

5% prob of error
Probabilistic answer

- Give ourselves a confidence target
  - for example, 5% probability of being wrong is acceptable
Central limit to the rescue

1000 samples: pretty big

If hypothesized distribution is Normal

approximate by 6 of sampled poll

mean confidence interval

54%

unknown average true mean of population

percentage given by a poll of 1000 people
Central limit to the rescue

\frac{s}{\sqrt{n}}

The probability that the true mean is by symmetry

26 gives us 5% proba to get roll value if were true mean

36 1%
Central limit to the rescue

• 1000 is big enough, the distribution of hypothetical polls should be normal
  – our one poll is one instance in this distribution
Two unknowns
Two unknowns

• Mean of the Gaussian
  – the true percentage of likely voters
  – This is the main value we want to derive

• Variance/stddev of the Gaussian
  – How precise our estimate is
Standard deviation estimate

\[ \mu = 0.34 \]

\[ \sigma^2 = 0.54 \left(1 - 0.54\right) + 0.46 \times \left(0 - 0.54\right)^2 \]

\[ \text{Weight for weighted average} \]

\[ \text{Square difference} \]

\[ \approx 0.25 \times \frac{1}{1000} \]
\[ \sigma = \sqrt{\frac{(1 - 0.54)^2 + 0.46 \times (0 - 0.54)^2}{\text{mean square difference}}} \]

\[ \sigma \approx \frac{0.25}{\sqrt{1000}} \]

\[ \sigma \approx 0.015 \]

for average of N samples

\[ \sigma = 0.03 \]

3\%
\[ \sigma^2 \]

\[ \sigma \approx \text{standard deviation of the sample} \]

\[ \text{mean: } 0.54 \]
\[ \text{variance for 1 sample} = 0.54 \times (1 - 0.54)^2 + 0.46 \times (0 - 0.54)^2 \]
\[ = 0.248 \]
\[ \text{stddev for 1 sample} = 0.498 \]
\[ \text{sddev for 1000 samples:} \]
\[ 0.498 / \sqrt{1000} = 0.015, \text{ i.e. } 1.5\% \]

\[ \bar{X} \sim \text{for average of } N \text{ samples} \]

\[ 2\sigma \approx 0.015 \]

\[ 2\sigma = 0.03 \]

\[ 3\% \]

Poll: estimate 54\% ± 3\%

With 85\% proba candidate A will be between 51\% and 57\%
Standard deviation estimate

• Approximate it by the standard deviation of the sample

• variance of 540 1’s and 460 0’s:
  - mean: 0.54
  - variance for 1 sample = 0.54*(1−0.54)^2+0.46*(0−0.54)^2
    = 0.248
  - stddev for 1 sample = 0.498
  - stddev for 1000 samples:
    - 0.498/sqrt(1000) = 0.015, i.e. 1.5%
Confidence interval
Confidence interval

• If N is big enough
  – Distribution is Normal

• Compute variance/stddev from the data

• Confidence interval at 5% margin of error is at 2 sigmas
  – recall sigma = 1.5% for this sample size

• with a 5% chance of error,
  the true percentage is between 51% and 57%
  – You’ll note that polls usually claim a +/-3% margin of error
What if we want less probability of error?

• For example to get 1%, use 3 sigmas
NB

polls.
samples are not random
stratification
NB

• Political polls violate many of our assumptions

• In particular they are not truly random samples
  – sometimes in good ways (better than random)
    – stratification
  – often in bad ways (lots of selection and convenience bias)
    – only people will landlines
    – only people willing to respond

• Also people lie
Boostrapping
Goal of bootstrapping

back to God's perspective
assume sample is representative
coin tosses
virtual exp

again
Goal of bootstrapping

• Get back to the God perspective where we can simulate lots of experiments
  – But for this we need to know the probability for the whole population

• Simpler, more general method
  – If you can compute a statistics, you can get its confidence interval

• Invented by Efron in 1979
  – Building on a bunch of previous ideas, e.g. jackknife
Physically bootstrapping
Bootstrapping
Bootstrapping

• Getting something from nothing
• Idea:
  – create virtual datasets
  – By resampling the data
Bootstrapping

• Say we believe the data is representative from the population
• Let’s (re)sample from the data to simulate different experiments
  – same number of samples
  – can possibly reuse the same original sample multiple times

– in python: use random.choice(data)
def bootstrap_mean(data, num_trials=10000):
    """return a list of bootstrapped mean estimate from random re-samples""
    N = len(data)
    L = []  # store all the virtual averages
    for i in xrange(num_trials):
        total = 0.0  # for average
        for i in xrange(N):
            sample = random.choice(data)
            total += sample
        L.append(total/N)
    return L

data_mean=sum(data)/float(len(data))
print data_mean

simulated_trials = bootstrap_mean(data)

pylab.hist(simulated_trials, bins = 100)
pylab.xlabel('average value')
pylab.ylabel('number of trials')
pylab.title('Distribution of average estimates for N coin tosses')
pylab.show()
Booststrapping confidence interval
Confidence interval
Confidence interval

• We have a list of virtual mean estimates

• What is the value mini so that 2.5% of the virtual means lie below and maxi so that 2.5% lie above?

• Just sort the data and pick the value at 2.5% of the length of the list, and that at 97.5%

$\{0.4, 0.5, 0.5, 0.1, 0.5, 0.2\}$
```python
import random
import pylab

data = [1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1]

def bootstrap_mean(data, num_trials=10000):
    """return a list of bootstrapped mean estimate from random re-samples""
    N = len(data)
    L = []
    for i in xrange(num_trials):
        total = 0.0
        for i in xrange(N):
            sample = random.choice(data)
            total += sample
        L.append(total/N)
    return L

def confidence_interval(data, p, num_trials=10000):
    simulated_data = bootstrap_mean(data, num_trials)
    sorted_data = sorted(simulated_data)
    mini = sorted_data[int(p/2.0*len(sorted_data))]
    maxi = sorted_data[int((1-p/2.0)*len(sorted_data))]
    return mini, maxi

data_mean=sum(data)/float(len(data))
print "The mean of the data is ", data_mean

p=0.05
mini, maxi = confidence_interval(data, p)
print "confidence interval for p =", p*100, "% is [", mini, ",", maxi, "]"
```