Probability, Distributions, Variance

6.100 LECTURE 6

SPRING 2026

Announcements

- **Exam 1** next Wednesday, Feb 25
  - during class time
  - in **Walker 50-340**
  - will send logistics announcement tomorrow
- Last term's Midterm 1 released after class for practice
  - website top bar > **Course Info > Exam Practice**
  - some questions have been modified to better match course content this semester so far
  - TAs will review in recitation this Friday
  - solutions will be posted after recitations conclude

Exam studying tips

- **Python mechanics**
  - be completely clear how environment diagrams work
  - try to avoid mental shortcuts
- **Problem-solving, organizing code**
  - why should the code be designed that way?
  - is there another valid way?
    - see finger exercise solutions
    - reflect on Pset 1 checkoff
- **Be fluent enough with object operations**
  - need to engage more than just reading code
  - run your own experiments on the REPL

Exam grading expectations

- Before the exam, we will decide rough ranges for **middle-A/B/C**
- After grading, we may adjust these ranges, but never up, only down
- We will announce the **final ranges** when releasing scores

- If your performance is not what you had hoped, please reach out to **6.100-instructors@mit.edu** to meet with me

Python modules

matplotlib library

- A third-party module
  - installation instructions in Pset 2
- Python modules are "namespaces"
  - look like frames: map names to
  - but they are objects on the heap
  - module name is a variable in global frame, pointing to modul object on heap
- dot-notation
  - ***module.name*** → any object on the heap
- `import matplotlib.pyplot`
- **`import matplotlib.pyplot as plt`**

Probability and Distributions

What is a probability distribution

- When a run a simulation, **any outcome has a certain probability of occurring**
  - if outcome space is continuous, any outcome has a probability ***density***
- The **probabilities over all possible outcomes** is the distribution
  - all probabilities must sum (or integrate) to 1
- A **random variable** represents sampling a single outcome from a distribution
  - uppercase is random variable $\boldsymbol{X}$
  - lowercase is outcome $\boldsymbol{x}$**,** like a normal math variable

Properties of sampling

- **Fundamental property**
  - after sampling enough outcomes from a distribution, the samples themselves begin to reflect the shape of the distribution
- → **Law of Large Numbers**
  - the mean of the samples converges to the mean of the distribution
- **Monte Carlo** simulation
  - simply repeat an experiment many times
  - compute an aggregate of the results (such as a mean)
  - (name is a reference to casino in Monaco)

Measuring spread in a distribution

- For LLN, how fast does the sample mean converge?
  - i.e., how many samples are needed to get a good estimate?
- Need to measure how "spread out" a distribution is
- Given samples $\boldsymbol{x_1}$, … $\boldsymbol{x_n}$, consider all differences relative to mean $\boldsymbol{\mu_x}$
- Simply summing them results in 0, not useful

Variance and standard deviation

- Idea: sum their magnitudes
  - **absolute deviation:** $\left(\sum_{i=1}^{n}|x_i - \mu_x|\right) \ / \ n$
  - intuitive, but awkward to analyze mathematically
- Idea 2: sum their differences squared
  - **variance σ^2:** $\left(\sum_{i=1}^{n}(x_i - \mu_x)^2\right) \ / \ n$
  - penalizes more for outliers
- Variance's units are squared, take square root to get comparable units
  - **standard deviation σ**

Some common ideal distributions

- Uniform
  - https://en.wikipedia.org/wiki/Continuous_uniform_distribution
  - https://en.wikipedia.org/wiki/Discrete_uniform_distribution
- Normal / Gaussian
  - https://en.wikipedia.org/wiki/Normal_distribution
- Exponential
  - https://en.wikipedia.org/wiki/Exponential_distribution

Distributions of Means

Evaluating accuracy of sample mean

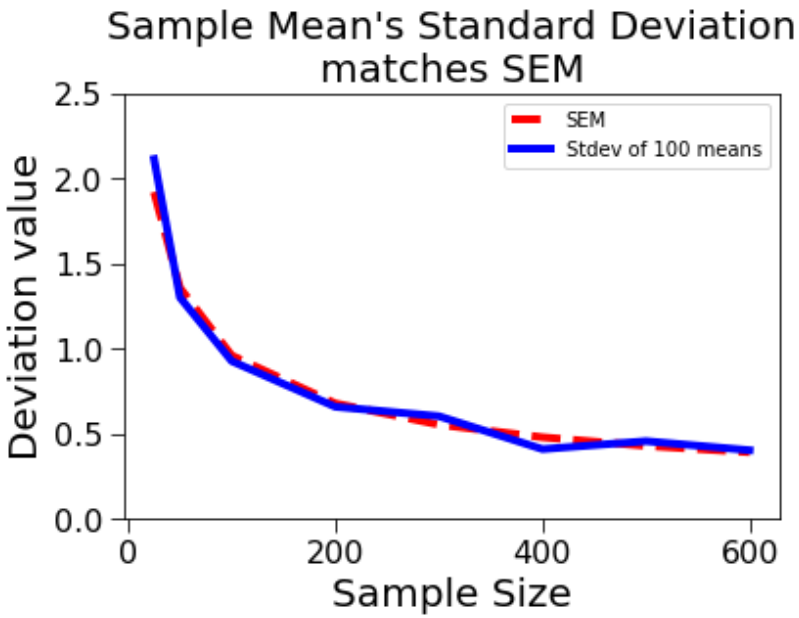
- As we collect more sample, get a better sense of the underlying distribution
  - know its mean and stdev better
- But can't just look at the variance of the sample
- Consider mean as a random variable
  - $\boldsymbol{M = (X_1 + X_2 + \cdots + X_n) \ / \ n}$
- $\boldsymbol{M}$ has its own distribution
  - we're interested in the shape/variance of that distribution
  - would like it to be narrow and tall

Estimating distribution of sample mean

- Sample $\boldsymbol{M}$ many times
  - which means running $\boldsymbol{n}$ trials of sampling $\boldsymbol{X}$
  - for a total of $\boldsymbol{k}$ times
- Want to see how accuracy improves with more samples of X
  - keep $\boldsymbol{k}$ fixed, increase $\boldsymbol{n}$
- Terminology
  - sometimes, $\boldsymbol{n}$ trials of $\boldsymbol{X}$ is called a single sample with a sample size of $\boldsymbol{n}$
    - as if taking a group of size $\boldsymbol{n}$ out of an infinite population for $\boldsymbol{X}$
  - and $\boldsymbol{k}$ samples of $\boldsymbol{M}$ is called $\boldsymbol{k}$ trials of our experiment

Central Limit Theorem

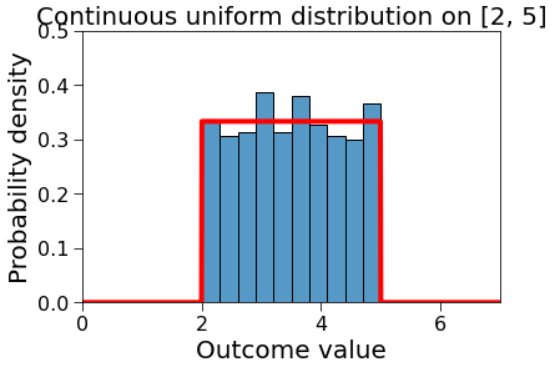
- The distribution for $\boldsymbol{M}$ approaches a normal distribution as the sample size $\boldsymbol{n}$ increases
- The mean of that distribution is the same as the mean of $\boldsymbol{X}$
  - $\boldsymbol{\mu_M = \mu_X}$
- The variance of that distribution is the variance of $\boldsymbol{X}$ divided by the sample size
  - $\boldsymbol{\sigma_M^2 = \sigma_X^2 \,/\, n}$
- Hence, $\boldsymbol{M}$'s standard deviation scales inversely with the square root of the sample size
  - $\boldsymbol{\sigma_M = \sigma_X \,/\, \sqrt{n}}$
- $\boldsymbol{M}$'s standard deviation is called the **standard error of the mean (SEM or SE)**

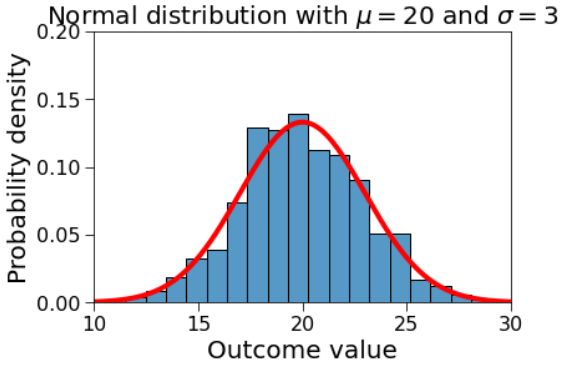


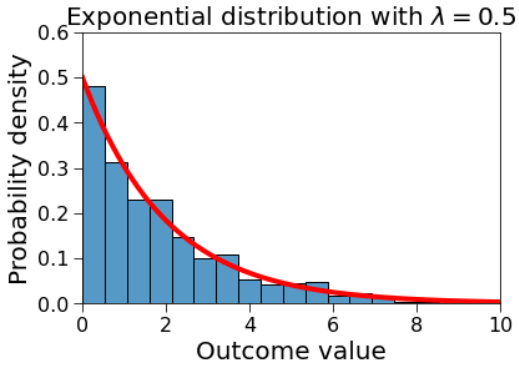
Estimating standard error

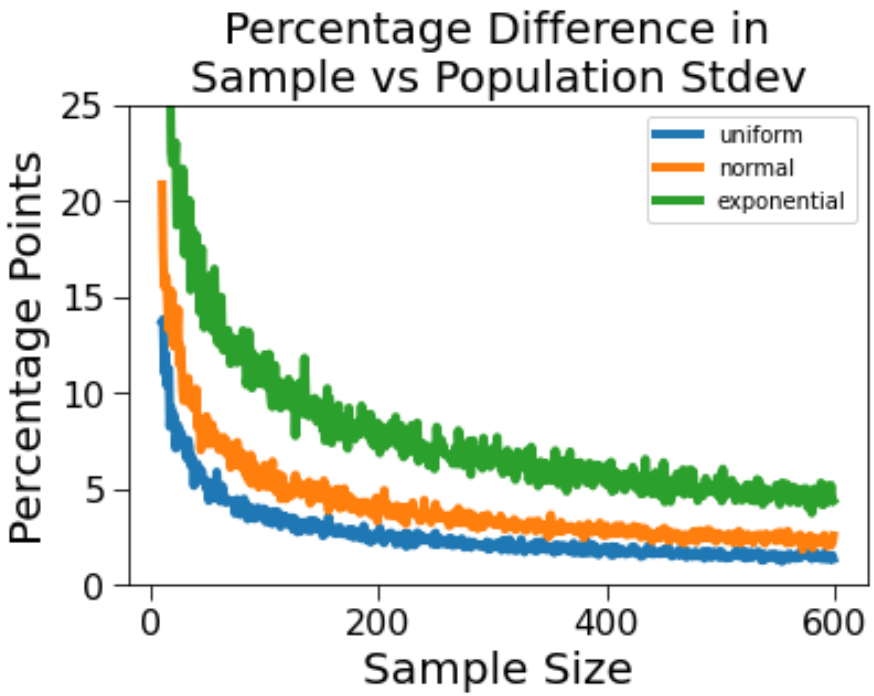
- Expensive to sample M multiple times
  - k trials of n trials
  - would prefer to know σ_M from a single sample of M
- Just need to know σ_X and sample size n
  - but we don't know X's underlying distribution
- But we have n samples of X
  - if n is large enough, we have some approximation of X's distribution
- Is the variance of the samples σ_x a good enough replacement for the true distribution variance σ_X?

Sample stdev vs distribution stdev









- Convergence depends on **skew**
  - a measure of the asymmetry in a distribution
- **More skewed distributions require more samples for sample stdev to converge**

Next time

- **Next few lectures have been rearranged**
  - start with optimization and curve-fitting
  - after Exam 1, resume curve-fitting discussion
  - then discuss statistical significance
  - better matches flow of Pset 3
  - Pset 3 release pushed back Feb 25, after Exam 1