CSE 312 – Section 7

Spring 2026

Review of Main Concepts

- **Normal (Gaussian, "bell curve")**: $X \sim \mathcal{N}(\mu,\ \sigma^2)$ iff X has the following probability density function:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad x \in \mathbb{R}$$

  $\mathbb{E}[X] = \mu$ and $Var(X) = \sigma^2$. The "standard normal" random variable is typically denoted Z and has mean 0 and variance 1: if $X \sim \mathcal{N}(\mu,\ \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$. The CDF has no closed form, but we denote the CDF of the standard normal as $\Phi(z) = F_Z(z) = \mathbb{P}(Z \leq z)$. Note from symmetry of the probability density function about $z = 0$ that: $\Phi(-z) = 1 - \Phi(z)$.

- **Standardizing**: Let X be any random variable (discrete or continuous, not necessarily normal), with $\mathbb{E}[X] = \mu$ and $Var(X) = \sigma^2$. If we let $Y = \frac{X-\mu}{\sigma}$, then $\mathbb{E}[Y] = 0$ and $Var(Y) = 1$.

- **Closure of the Normal Distribution**: Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, $aX+b \sim \mathcal{N}(a\mu+b, a^2\sigma^2)$. That is, linear transformations of normal random variables are still normal.

- **"Reproductive" Property of Normals**: Let $X_1, \ldots, X_n$ be independent normal random variables with $\mathbb{E}[X_i] = \mu_i$ and $Var(X_i) = \sigma_i^2$. Let $a_1, \ldots, a_n \in \mathbb{R}$ and $b \in \mathbb{R}$. Then,

$$X = \sum_{i=1}^{n}(a_i X_i + b) \sim \mathcal{N}\left(\sum_{i=1}^{n}(a_i\mu_i + b), \sum_{i=1}^{n} a_i^2\sigma_i^2\right)$$

  There's nothing special about the parameters – the important result here is that the resulting random variable is still normally distributed.

- **Z-score:** I have not used this term in class, but a Z-score measures how many standard deviations a specific data point is above or below the mean, indicating its position relative to the average. A Z-score of 0 equals the mean, while a Z-score of 1 is one standard deviation above the mean.

- **Central Limit Theorem (CLT)**: Let $X_1, \ldots, X_n$ be iid random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. Let $X = \sum_{i=1}^{n} X_i$, which has $\mathbb{E}[X] = n\mu$ and $Var(X) = n\sigma^2$. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, which has $\mathbb{E}\left[\overline{X}\right] = \mu$ and $Var(\overline{X}) = \frac{\sigma^2}{n}$. $\overline{X}$ is called the *sample mean*. Then, as $n \to \infty$, $\overline{X}$ approaches the normal distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Standardizing, this is equivalent to $Y = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ approaching $\mathcal{N}(0,1)$. Similarly, as $n \to \infty$, X approaches $\mathcal{N}(n\mu, n\sigma^2)$ and $Y' = \frac{X-n\mu}{\sigma\sqrt{n}}$ approaches $\mathcal{N}(0,1)$.

  It is no surprise that $\overline{X}$ has mean μ and variance σ^2/n – this can be done with simple calculations. The importance of the CLT is that, for large n, regardless of what distribution

X_i comes from, $\overline{X}$ is *approximately normally distributed with mean μ and variance σ^2/n.* Don't forget the continuity correction, only when $X_1, \ldots, X_n$ are discrete random variables.

- **Continuity Correction:** When we use the Central Limit Theorem (CLT) to approximate a discrete random variable X (such as a Binomial or Poisson random variable) with a continuous Normal random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$, we encounter a structural mismatch.

  A discrete random variable takes on exact integer values, meaning probabilities like $\mathbb{P}(X = k)$ are strictly positive. However, for a continuous random variable, the probability of any single exact point is zero: $\mathbb{P}(Y = k) = 0$.

  To account for this, we use a "continuity correction". We associate the discrete probability at the integer k with the continuous probability *interval* from $k - 0.5$ to $k + 0.5$. Geometrically, think of this as matching the area of a discrete histogram bar (centered at k with width 1) to the corresponding area under the smooth continuous Normal curve.

  How to apply the Correction When converting discrete bounds to continuous bounds, we expand the interval by 0.5 in the relevant directions to ensure the entire "histogram bar" is captured. The following assumes that the discrete random variable X takes consecutive integer values.

  - Exact value:
  $$\mathbb{P}(X = k) \approx \mathbb{P}(k - 0.5 \leq Y \leq k + 0.5)$$
  - Less than or equal to:
  $$\mathbb{P}(X \leq k) \approx \mathbb{P}(Y \leq k + 0.5)$$
  - Greater than or equal to:
  $$\mathbb{P}(X \geq k) \approx \mathbb{P}(Y \geq k - 0.5)$$
  - Strict inequalities: First, convert strict inequalities to non-strict inequalities (since X only takes integer values), and then apply the correction.
  $$\mathbb{P}(X < k) = \mathbb{P}(X \leq k - 1) \approx \mathbb{P}(Y \leq (k - 1) + 0.5) = \mathbb{P}(Y \leq k - 0.5)$$
  $$\mathbb{P}(X > k) = \mathbb{P}(X \geq k + 1) \approx \mathbb{P}(Y \geq (k + 1) - 0.5) = \mathbb{P}(Y \geq k + 0.5)$$

  A helpful rule of thumb: Always sketch the histogram bars and *include* the entire bar for any integers that satisfy the discrete inequality. If you want $X \leq k$, you must include the entire bar for k, which extends up to $k + 0.5$. If you want $X < k$, you do not include the bar for k; you only include the bar for $k - 1$, which extends its right edge up to $k - 0.5$.

- **General template for solving CLT problems:** Sometimes we'll be trying to solve for the probability of something (e.g., $P(X \leq 10)$, and sometimes, we'll be trying to find a value of some parameter that will allow for the probability to be in a certain range (e.g., $P(X \leq 10) \leq 0.2$). Regardless, we still will want to apply CLT on X, and follow the same process (the only difference is that we may be solving for different things).

  a) Setup the problem - write event you are interested in, in terms of sum of random variables. (what do we want to solve for/what is the probability we want to be true?)
     - Write the random variable we're interested in as a sum of i.i.d., random variables

- Apply CLT to $X = X_1 + X_2 + ... + X_n$ (we can approximate X as a normal random variable $Y \sim N(\mu, \sigma^2)$)
- Write the probability we're interested in

b) *If the RVs are discrete*, apply continuity correction.

c) Normalize RV to have mean 0 and standard deviation 1: $Z = \frac{Y-\mu}{\sigma}$

d) Replace RV in probability expression with $Z \sim N(0, 1)$

e) Write in terms of $\Phi(z) = P(Z \leq z)$

f) Look up in the Phi table (or do a reverse Phi table lookup if we're looking for a value of z that gives us a certain probability)

Announcements & Plan for Section

Announcements

- Midterm today, 4/14 @ 6pm. Please bring a photo ID.
- Pset 6 is due next week, 5/20 @ 11:59pm.

Plan for Section

- Answer any questions about the midterm: content, practice tests, etc.
- Content Review (Problem 1)
- Problem 4: Bad Computer (if time remaining).
- Problem 5: Tweets or Problem 6: Ping Pong (if time remaining).

Midterm Prep Resources

- Link to information about exam.
- Link to draft cheat sheet.
- Link to practice midterm and solutions to practice midterm

1 Content Review - understanding the Central Limit Theorem

The Central Limit Theorem (CLT) is one of the most powerful results in probability because it allows us to approximate the distribution of sums and averages of independent random variables, regardless of their original distribution.

a) (4 points) You roll a fair, 6-sided die 100 times. The outcome of a single roll has a discrete uniform distribution with a mean of $\mu = 3.5$ and a variance of $\sigma^2 \approx 2.92$. Let $\bar{X}$ be the **average** value of your 100 rolls. According to the Central Limit Theorem, which of the following best describes the approximate distribution of $\bar{X}$?

(a) A discrete uniform distribution from 1 to 6.

(b) A normal distribution with mean 3.5 and variance 2.92.

(c) A normal distribution with mean 3.5 and variance 0.0292.

(d) A normal distribution with mean 350 and variance 292.

b) (4 points) A shipping company loads 100 identical packages onto a truck. The weight of a single package is a random variable with a mean of 20 lbs and a standard deviation of 5 lbs. Let S be the **total combined weight** of all 100 packages. Using the CLT, what is the approximate distribution of S?

(a) $S \approx \mathcal{N}(2000, 500)$

(b) $S \approx \mathcal{N}(2000, 2500)$

(c) $S \approx \mathcal{N}(20, 0.25)$

(d) $S \approx \mathcal{N}(2000, 5)$

c) (4 points) The waiting time for a bus is modeled by an Exponential distribution, which is heavily right-skewed. If you record the waiting times for 64 independent bus rides and calculate your average waiting time $\bar{X}$, what will the shape of the distribution of $\bar{X}$ look like?

(a) It will be exactly Exponential, because the sum of exponentials is exponential.

(b) It will be heavily right-skewed, matching the underlying population.

(c) It will be approximately a bell-shaped Normal curve.

(d) It will be perfectly uniform, since all averages balance out.

d) (4 points) A researcher takes a sample of size $n = 36$ from a population with mean $\mu = 50$ and variance $\sigma^2 = 144$. They want to find the probability that the sample mean $\bar{X}$ is greater than 54. To use the Standard Normal table, they must convert $\bar{X}$ to a standard Z-score. Which of the following is the correct calculation for Z?

(a) $Z = \frac{54-50}{144/\sqrt{36}}$

(b) $Z = \frac{54-50}{12/36}$

(c) $Z = \frac{54-50}{12/\sqrt{36}}$

(d) $Z = \frac{54-50}{144/36}$

2 More review

(a) (4 points) What does the Central Limit Theorem fundamentally guarantee about a sequence of i.i.d. random variables as the sample size n becomes very large?

i. The distribution of the individual random variables gradually becomes normal.

ii. The sample mean gets closer and closer to the true expected value of the distribution.

iii. The distribution of the sample mean (or sample sum) approaches a normal distribution, regardless of the original distribution's shape.

iv. The sum of any two normally distributed random variables will also be normally distributed.

(b) (4 points) Suppose the weight of an apple in an orchard has an unknown, highly skewed distribution with an expected value of 150 grams and a standard deviation of 20 grams. You pick a random sample of 100 apples. Let S be the total weight of these 100 apples. Using the CLT, what is the approximate probability that the total weight exceeds 15,200 grams?

Hint: First, find the expected value $\mathbb{E}[S]$ *and the variance VarS for the sum of the 100 apples. Then standardize to find a Z-score (. Assume* $\Phi(1) \approx 0.84$.

i. 0.84
ii. 0.16
iii. 0.50
iv. 0.05

(c) (4 points) A computer server processes requests with an expected response time of 40 milliseconds and a standard deviation of 12 milliseconds. The distribution of response times is heavily right-skewed (an exponential-like tail). If you take a random sample of 36 requests, what is the approximate distribution of the **sample mean**, $\bar{X}$?

Hint: Remember that the variance of the sample mean is $Var(\bar{X}) = \frac{\sigma^2}{n}$. *What is its standard deviation (the standard error)?*

i. Heavily right-skewed with a mean of 40 and a standard deviation of 12.
ii. Approximately Normal with a mean of 40 and a standard deviation of 12.
iii. Approximately Normal with a mean of 40 and a standard deviation of 2.
iv. Approximately Normal with a mean of 1440 and a standard deviation of 72.

3 Round off error

Let X be the sum of 100 real numbers, and let Y be the same sum, but with each number rounded to the nearest integer before summing. If the roundoff errors are independent and uniformly distributed between -0.5 and 0.5, what is the approximate probability that $|X - Y| > 3$?

4 Bad Computer

Each day, the probability your computer crashes is 10%, independent of every other day. Suppose we want to evaluate the computer's performance over the next 100 days.

a) Let X be the number of crash-free days in the next 100 days. What distribution does X have? Identify $\mathbb{E}[X]$ and $Var(X)$ as well. Write an exact (possibly unsimplified) expression for $\mathbb{P}(X \geq 87)$.

b) Approximate the probability of at least 87 crash-free days out of the next 100 days using the Central Limit Theorem. Use continuity correction.

Important: continuity correction says that if we are using the normal distribution to approximate

$$\mathbb{P}(a \leq \sum_{i=1}^{n} X_i \leq b)$$

where $a \leq b$ are integers and the X_i's are i.i.d. **discrete** random variables taking integer values, then, as our approximation, we should use

$$\mathbb{P}(a - 0.5 \leq Y \leq b + 0.5)$$

where Y is the appropriate normal distribution that $\sum_{i=1}^{n} X_i$ converges to by the Central Limit Theorem. (The intuition here is that, to avoid a mismatch between discrete distributions (whose range is a set of integers) and continuous distributions, we get a better approximation by imagining that a discrete random variable, say W, is a continuous distribution with density function

$$f_W(x) := p_W(i) \quad \text{when } i - 0.5 \leq x < i + 0.5 \text{ and } i \text{ integer}$$

)

For more details see pages 209-210 in the Tsun book.

5 Tweets

A prolific twitter user tweets approximately 350 tweets per week. Let's assume for simplicity that the tweets are independent, and each consists of a uniformly random number of characters between 10 and 140. (Note that this is a discrete uniform distribution.) Thus, the central limit theorem (CLT) implies that the number of characters tweeted by this user is approximately normal with an appropriate mean and variance. Assuming this normal approximation is correct, estimate the probability that this user tweets between 26,000 and 27,000 characters in a particular week. (This is a case where continuity correction will make virtually no difference in the answer, but you should still use it to get into the practice!).

6 Ping Pong

You're playing ping pong with your friend, and want to keep playing until you've scored 15 points. Unfortunately, your friend is a much more skilled ping pong player than you, so you only win points 25% of the time (with each point being independent of the other points). Approximate the probability that you'll need to play at least 50 points before stopping.

7 More normal stuff (10 points)

Let X be a normal random variable with mean 12 and variance 4. Find the value of c such that $\mathbb{P}[X < c] = 0.1$

8 CLT for stocks (10 points)

Suppose that the daily price change of a certain stock on the stock market is a random variable with mean 0 and variance σ^2. Thus, if Y_n is the price of the stock on the n-th day, then

$$Y_n = Y_{n-1} + X_n, \quad n \geq 1$$

where $X_1, X_2, ...$ are independent, identically distributed random variables with mean 0 and variance σ^2. Suppose also that today's stock price is 100 and $\sigma^2 = 16$. Use the Central Limit Theorem to estimate the probability that the stock price will exceed 110 after 10 days.