ST5201: Basic Statistical Theory
Chapter 5: Limit Theorems

Choi, Yunjin
stachoiy@nus.edu.sg

Department of Statistics and Applied Probability
National University of Singapore (NUS)

19th September, 2017

Announcement

- Midterm on 3rd October (in class):
  - From lecture 1 to lecture 5.
  - One sheet of two-sided A4 allowed
  - A non-programmable calculator is allowed and might be necessary
- If you need a make-up exam:
  - official document needed
  - need to inform me by 27th of September; for those who do not notify by the date, a make-up exam will not be available

Outline

- Review
- Introduction
- Three Types of Convergence
- The Law of Large Numbers
- The Central Limit Theorem

- Covariance
  - Definition:
    $Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$
  - Interpretation: positively correlated/negatively correlated/uncorrelated
  - Remark: Independence $\Rightarrow$ Uncorrelated; Uncorrelated $\nRightarrow$ Independence
  - Properties: $\text{Var}(X) = Cov(X, X)$,
    $Cov(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j)$.

- Correlation coefficient
  - Definition: $Corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$
  - Remark: $-1 \leq Corr(X, Y) \leq 1$.
  - Properties: $Corr(a + bX, c + dY) = Corr(X, Y)$. It does not change under linear transformation for 1 r.v..

- Conditional Expectation
  - Definition: $E[X|Y=y] = \sum_x x p_{X|Y}(x|y)$ and $E[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx$.
  - Interpretation: Note that $X|Y=y$ is a new r.v., $E(X|Y=y)$ is the expectation on this r.v.
- Law of Total Expectation
  - $E[X|Y=y]$ assigns a number to each y. Therefore, $E(X|Y)$ can be viewed as a function of Y, which is a new r.v..
  - $E[E(X|Y)] = E(X)$
  - Useful for problems with several cases.
- Moment Generating Function
  - Definition: $M_X(t) = E(e^{tX})$ is a function of t
  - Characterization of a r.v. (similar as CDF/PDF/PMF)
  - Calculate moments: $E[X^k] = \frac{d^k}{dt^k} M_X(t)|_{t=0}$
  - Property 1: $M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t)$ for indept. r.v.'s $X_1, X_2, ..., X_t$
  - Property 2: $M_{aX+b}(t) = e^{at} M_X(bt)$.
  - MGF for common distributions

Introduction

Learning Outcomes

- Questions to Address: What will happen when we repeat the experiment many times? ⋆ What the relationship between parameter and stat. ⋆ How to approximate the probability for i.i.d samples

Concept & Terminology

- i.i.d. ⋆ Convergence in probability ⋆ Convergence in distribution ⋆ almost sure convergence
- Weak Law of Large Numbers ⋆ Strong Law of Large Numbers ⋆ Monte Carlo Method
- Central Limit Theorem ⋆ Normal Approximation

Mandatory Reading

Textbook: Section 5.1 – Section 5.3

Motivation

Recall

- Mean: long-run average
- Mode: If we repeated the experiment many times independently, the most frequently outcome

Questions:

1. Why do we care about long-run results?

2. What results do we have?

In many fields, people care about many experiments instead of 1

- In a casino, there are hundreds of players playing with the slot machine
- For a company, there might be tens of thousands of customers in one day
- In a hospital, there are hundreds of patients in one day
- In a survey, there might be thousands of participants
- In our class, we have 68 students working on assessments/tests
- ...

Results are based on all these experiments:

- In a casino with hundreds of players playing, the manager cares about the sum of all these experiments
- For a company with tens of thousands of customers, the manager cares about the cost of all these customers
- In a hospital with hundreds of patients, the manager cares about the rate of recovery
- In a survey with thousands of participants, the researcher is interested in statistics from all these participants
- In my class, I care about the average performance of students

- It is impossible to find the distribution for each subject; a simplified and reasonable assumption that the outcomes of these experiments have the same distribution, and independent

Definition

For r.v.'s $X_1, X_2, \cdots, X_n$, if they follow the same distribution and independently distributed, then we say they are *independent and identically distributed (i.i.d)*

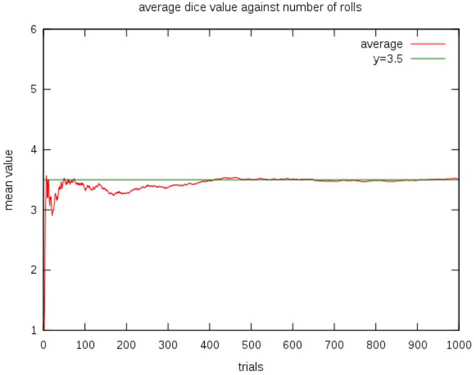
- If the r.v.'s are $i.i.d.$, intuitively,

$$\text{Statistics} \Leftrightarrow \text{Parameter/Distribution}$$

  - the distribution of one r.v. can be estimated by the empirical distribution of all these r.v.'s
  - some statistics (e.g., average) can be found according to the distribution of one r.v.

A Naive but Non-trivial Example

- Roll a fair six-sided die ($\Omega = \{1, 2, 3, 4, 5, 6\}$)
- Expectation: $E(X) = (1 + 2 + 3 + 4 + 5 + 6)/6$
- Average when we have 1, 2, 3, 4, and even more samples: $\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\} = X_1, \frac{X_1+X_2}{2}, \frac{X_1+X_2+X_3}{3}, \frac{X_1+X_2+X_3+X_4}{4}, \ldots$



- Intuitively, Gerolamo Cardano (1501–1576) stated that the accuracy of empirical statistics ($\{\frac{1}{n}\sum_{i=1}^{n} X_i\}$) tends to improve with more trials ($\{\frac{1}{n}\sum_{i=1}^{n} X_i\} \to E(X) = 3.5$), without leaving a proof
- Bernoulli proved the a form of Law of Large Numbers (LLN) and published on *Ars Conjectandi* in 1713; but the proof took him 20 years!
- Random trials are later named as Bernoulli trials
- We will prove it using Chebyshev's inequality (proved in 1867)
- Formally introduce *LLN* and *convergence in probability*

Different Types of Convergence

How to measure accuracy of empirical statistics?

- Define $Y_1 = X_1$, $Y_2 = \frac{X_1+X_2}{2}$, $Y_3 = \frac{X_1+X_2+X_3}{3}$, $\cdots$, $X_n = \frac{\sum_{i=1}^{n} X_i}{n}$, $\cdots$
- $\{Y_n\}$: a sequence of r.v.'s, with the same sample space Ω
- Interest in the limiting behavior of $\{Y_n\}$; especially, hope that $\{Y_n\} \to E(X)$.

What does the convergence of r.v.'s ($Y_n \to E(X)$, $n \to \infty$) mean?

- Converge in distribution
- Converge in probability
- Almost sure convergence

Definition: Convergence in Distribution

Let $\{X_n\} \equiv X_1, \cdots, X_i, \cdots$, be a sequence of r.v.'s with CDF $F_1, \cdots, F_i, \cdots$, and let X be a r.v. with CDF F. We say that *X_n converges in distribution to a r.v. X (i.e., $X_n \xrightarrow{d} X$)* if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

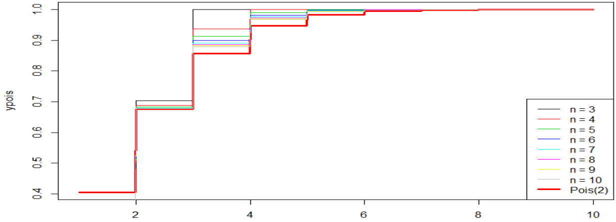
at every point at which F is cont.

- $\{X_n\}$ and X can be dependent or independent
- Convergence:
  - If X is discrete, the convergence stands at points F does not jump
  - If X is cont., the convergence stands at every point
- Interpretation: for any constant a, b,

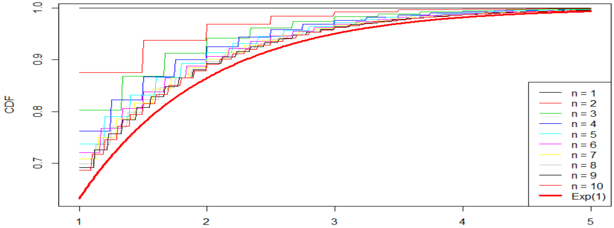
$$P(X_n \le b) \to P(X \le b), P(X_n > a) \to P(X > a),$$
$$P(a < X_n \le b) \to P(a < X \le b)$$

Example: Convergence in Distribution

- $X_n \sim Bin(n, 2/n)$, for $n = 3, 4, \cdots$, $X \sim Pois(2)$ (Poisson approximation for binomial r.v.'s)



- $Y_n \sim Geo(1/n)$, for $n = 1, 2, \cdots$, $X_n = Y_n/n$, $X \sim Exp(1)$



Definition: Convergence in Probability

For a sequence of r.v.'s $\{X_n\} = X_1, X_2, \cdots, X_i, \cdots$, we say they *converge in probability towards the r.v. X (i.e., $X_n \xrightarrow{P} X$)* if for any $\epsilon > 0$

$$P(|X_n - X| \geq \epsilon) \to 0 \text{ as } n \to \infty$$

or, equivalently,

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$$

- The target X has the same sample space with all the X_i's
- $\{X_n\}$ are usually dependent
- Practically, find the sequence of events $A_n = \{\omega \in \Omega, |X_n - X| \geq \epsilon\}$ by obtaining $|X_n - X|$ as a new r.v., and check if $P(A_n) \to 0$ when $n \to \infty$
- Interpretation: for any ϵ, the event that $|X - X_n| \geq \epsilon$ has probability smaller than δ when n is large enough. It concerns more about the probability measure and r.v., instead of the CDF only.

- Let X be a r.v. with prob 1 at 1, and $X_n \sim N(1, \frac{1}{n^2})$.

  $$P(|X - X_n| \geq \epsilon) = P(|N(0, \frac{1}{n^2})| \geq \epsilon) \leq \frac{1/n^2}{\epsilon^2} = \frac{1}{n^2\epsilon^2} \leq \delta, \quad n \geq \frac{1}{\epsilon\sqrt{\delta}}.$$

  So, $X_n \xrightarrow{P} X$.
  Remark: For any constant a, we can define a r.v. X with prob 1 at a. So the example here can also be written as

  $$X_n \xrightarrow{P} 1,$$

  where 1 denotes the r.v. with probability mass 1 at the point 1.

- Let $X_n \sim Ber(0.5)$, and $X \sim Ber(0.5)$, X and X_n are independent. Obviously, $X_n \xrightarrow{d} X$ as the CDF's for X_n and X are the same. However, X_n does NOT converge to X in probability. Note for any n,

  $$\begin{aligned} P(|X_n - X| \geq 1) &= P(\{X_n = 1, X = 0\} \cup \{X_n = 0, X = 1\}) \\ &= P(\{X_n = 1, X = 0\}) + P(\{X_n = 0, X = 1\}) \\ &= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = 1/2 \nrightarrow 0. \end{aligned}$$

Definition: Almost Sure Convergence

For a sequence of r.v.'s $\{X_n\} = X_1, X_2, \cdots, X_i, \cdots$ and X with the same sample space Ω, we say *X_n almost surely converge to X (i.e., $X_n \overset{a.s.}{\to} X$)* if

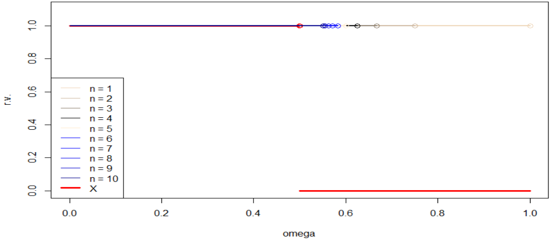
$$P(\lim_{n\to\infty} X_n(\omega) = X(\omega)) = 1$$

- $\{X_n\}$ and X are usually dependent
- Practically, to show the a.s. convergence,
  - For each outcome ω, find the sequence $X_1(\omega), X_2(\omega), X_3(\omega), \cdots$ (sequence of real numbers) and the real number $X(\omega)$. Figure out whether $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ is true or not
  - Let the event $A = \{\omega \in \Omega, \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$
  - Check if $P(A) = 1$.
- Interpretation: for almost all the outcomes ω, when n is large enough, $|X_n(\omega) - X(\omega)| \leq \epsilon$ for any $\epsilon > 0$

- Let the sample space $\Omega = [0,1]$, with a probability measure that is uniform on this space, i.e. $P([a,b]) = b - a$ for any $0 \le a \le b \le 1$. Let

$$X_n(\omega) = \begin{cases} 1, & 0 \le \omega < \frac{n+1}{2n} \\ 0, & \text{otherwise} \end{cases}, \text{ and } X(\omega) = \begin{cases} 1, & 0 \le \omega < \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

$X_n \sim Ber(\frac{n+1}{2n}) : [0,1] \to \{0,1\}$; $X \sim Ber(1/2) : [0,1] \to \{0,1\}$.



For each $\omega \in [0,1]$, if $\omega \in [0,1/2)$, then $X_n(\omega) = 1 = X(\omega)$. If $\omega \in (1/2,1]$, then $X_n(\omega) = 0 = X(\omega)$ when $\frac{n+1}{2n} < \omega$, which is equivalent with $n > 1/(2\omega - 1)$. When $\omega = 1/2$, then $X_n(\omega) = 1 \nrightarrow X(\omega) = 0$. So $A = [0,1/2) \cup (1/2,1]$

$$X_n \overset{a.s.}{\sim} X \Rightarrow X_n \overset{P}{\sim} X \Rightarrow X_n \overset{d}{\sim} X$$

- The requirement is stronger and stronger
- Can be proved by the definition

$$X_n \overset{d}{\sim} X \not\Rightarrow X_n \overset{P}{\sim} X \not\Rightarrow X_n \overset{a.s.}{\sim} X$$

- $X_n \overset{d}{\sim} X \not\Rightarrow X_n \overset{P}{\sim} X$: Example 2 on Page 15
- $X_n \overset{P}{\sim} X \not\Rightarrow X_n \overset{a.s.}{\sim} X$: External reading: `http://math.stackexchange.com/questions/149775/convergence-of-random-variables-in-probability-but-not-almos`

Let $\{X_n\}$, X, $\{Y_n\}$ and Y be r.v.'s,

- If $X_n \xrightarrow{P} \mu$ and $g(\cdot)$ is a continuous function, then $g(X_n) \xrightarrow{P} g(\mu)$
- If $X_n \xrightarrow{d} \mu$ and $g(\cdot)$ is a continuous function, then $g(X_n) \xrightarrow{d} g(\mu)$

- If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$
- If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n Y_n \xrightarrow{P} XY$

- No such properties for a.s. convergence

WLLN (a simple one)

Let $\{X_n\} \equiv X_1, \cdots, X_i, \cdots$, be a sequence of independent r.v.'s with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. let $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$. Then,

$$\bar{X}_n \xrightarrow{P} \mu,$$

where μ denotes the r.v. with probability 1 at the point μ.

- $\{\bar{X}_n\}$ forms a sequence of dependent r.v.'s
- Obtained by applying Chebyshev's inequality (see next slide)
- X_i's are not required to be i.i.d in WLLN but should be independent and sharing the same 1st moments and 2nd moments
- According to properties for convergence in probability, for any cont. function $g(\cdot)$,

$$g(\bar{X}_n) = g\left(\frac{\sum_{i=1}^n X_i}{n}\right) \xrightarrow{P} g(\mu)$$

To prove the convergence in probability, we need that, for any $\epsilon > 0$,

$$P(|\bar{X}_n - \mu| \geq \epsilon) \to 0.$$

To give an upper bound for the probability, recall Chebyshev's inequality:

$$P(|\bar{X}_n - E(\bar{X}_n)| \geq \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2}.$$

So we need $E(\bar{X}_n)$ and $\mathrm{Var}(\bar{X}_n)$.

Because X_i's are independent, $E(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \mu$ and $\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{\sigma^2}{n}$.

Introduce $E(\bar{X}_n) = \mu$ and $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$ in, then

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0, \quad \text{as } n \to \infty$$

SLLN

Let $\{X_n\} \equiv X_1, \cdots, X_i, \cdots$, be a sequence of independent r.v.'s with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. let $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$. Then, for any $\epsilon > 0$,

$$P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1,$$

or, equivalently,

$$\bar{X}_n \overset{a.s.}{\to} \mu$$

- $\{\bar{X}_n\}$ forms a sequence of dependent r.v.'s
- X_i's are not required to be i.i.d in SLLN but should be independent and sharing the same 1st moments and 2nd moments
- Different from WLLN, for a cont. function $g(\cdot)$, we cannot claim that

$$g(\bar{X}_n) \overset{a.s.}{\to} g(\mu)$$

	WLLN	SLLN
Conditions	Independent Share μ and σ^2	Independent Share μ and σ^2
Results	$\bar{X}_n \xrightarrow{P} \mu$ $\Updownarrow$ $P(\lvert\bar{X}_n - \mu\rvert > \epsilon) \to 0$	$\bar{X}_n \xrightarrow{a.s.} \mu$ $\Updownarrow$ $P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$
Property	cont. $g(\cdot)$, $g(\bar{X}_n) \xrightarrow{P} g(\mu)$	cont. $g(\cdot)$, $g(\bar{X}_n) \overset{a.s.}{\nrightarrow} g(\mu)$
Interpretation	For any $\epsilon > 0$, $\delta \geq 0$ $P(\lvert\bar{X}_n - \mu\rvert \leq \epsilon) \geq 1 - \delta$ for n large enough	For any $\epsilon > 0$, $\delta \geq 0$ $P(\cup_{n\geq N}^{\infty} \lvert\bar{X}_n - \mu\rvert \leq \epsilon) \geq 1 - \delta$ for N large enough

- WLLN shows that each single r.v. X_n is very likely to be near to μ when with $n \geq N$; SLLN further shows that the probability is small even when we consider the sequence $\{X_n\}_n^N$ when N is large enough
- SLLN has the same conditions with WLLN, but stronger results. You could always use SLLN only
- WLLN is important for historical reasons and future studies in measure theory, not in our class

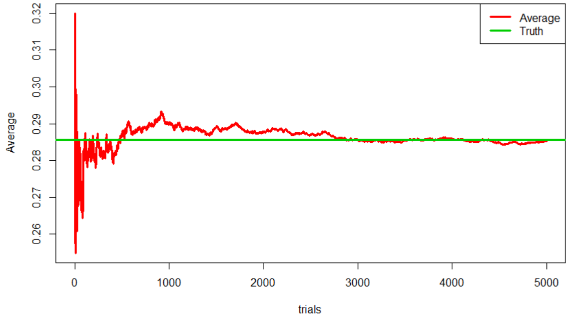
Example of LLN: Calculate Expectation

Recall: $E(X) = \int_{-\infty}^{\infty} x f(x) dx$, where $f(x)$ is pdf of X.

- Generate n trials with pdf $f(x)$, and calculate the $\bar{X}_n$. When n is very large, $E(X) \approx \bar{X}_n$
- Example: Beta distribution with parameters $a = 2$, $b = 5$.

$$E(X) = \int_0^1 x \times \frac{\Gamma(7)}{\Gamma(2)\Gamma(5)} x^{2-1}(1-x)^{5-1} dx.$$

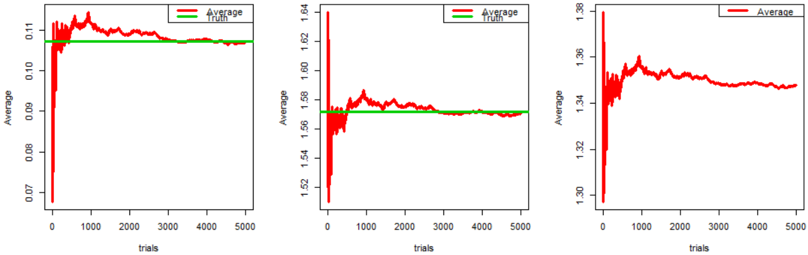
Hard to Calculate!



- What's more, $g(E(X)) \approx g(\bar{X}_n)$, e.g., $[E(X)]^2 \approx \bar{X}_n^2$

- LLN can also be used to find $E(g(X))$, where $g(\cdot)$ is a function
- Generate n i.i.d. trials $\{X_i\}_{i=1}^n$ with pdf $f(x)$, and let $Y_i = g(X_i)$. When n is very large, $E(g(X)) \approx \bar{Y}_n$
- Example: Beta distribution with parameters $a = 2$, $b = 5$.

Let $Y = X^2$ $\quad\quad$ $Z = 2X + 1$ $\quad\quad$ $W = e^X$



- $\text{Var}(X) \approx \bar{X^2}_n - (\bar{X}_n)^2$

Examples of Using LLN: Integration

- Suppose we wish to calculate

$$\int_0^1 g(x)dx$$

  where the integration is not easy to compute.

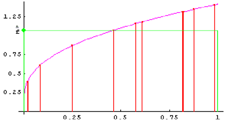
- Let $X \sim Unif(0,1)$, then the pdf of X is 1 on $[0,1]$. For function $g(\cdot)$,

$$E[g(X)] = \int_0^1 g(x) \cdot 1\, dx = \int_0^1 g(x)dx$$

Procedure (apply the method in previous slide for mean):

- Generate n i.i.d trials $X_i \sim Unif(0,1)$, and calculate $g(X_i)$ correspondingly
- Compute $E[g(X)] \approx \overline{g(X_i)} = \sum_{i=1}^n g(X_i)/n$, and so $\int_0^1 g(x)dx = E[g(X)]$
- This method is called Monte Carlo method

Integral of $\int_0^1 \sqrt{x+\sqrt{x}}dx$



Points Generated n = 10

The average of $\{f[x_i]\}_{i=1}^{10}$ is

$$\hat{f} = \frac{1}{10}\sum_{i=1}^{10} f[x_i] = 1.06496$$

Approximation for the integral

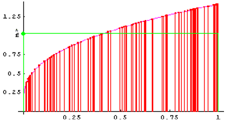
$$\int_0^1 \left(\sqrt{\sqrt{x}+x}\right)dx \approx \frac{1}{n}\sum_{i=1}^{n} f[x_i]$$

$$\int_0^1 \left(\sqrt{\sqrt{x}+x}\right)dx \approx \frac{1}{10}\sum_{i=1}^{10} f[x_i]$$

$$\int_0^1 \left(\sqrt{\sqrt{x}+x}\right)dx \approx \hat{f}$$

$$\int_0^1 \left(\sqrt{\sqrt{x}+x}\right)dx \approx 1.06496$$

Actual |Area-approx| $\approx$ 0.0196588



Points Generated n = 100

The average of $\{f[x_i]\}_{i=1}^{100}$ is

$$\hat{f} = \frac{1}{100}\sum_{i=1}^{100} f[x_i] = 1.02576$$

Approximation for the integral

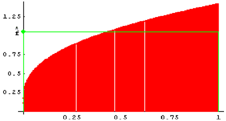
$$\int_0^1 \left(\sqrt{\sqrt{x}+x}\right)dx \approx \frac{1}{n}\sum_{i=1}^{n} f[x_i]$$

$$\int_0^1 \left(\sqrt{\sqrt{x}+x}\right)dx \approx \frac{1}{100}\sum_{i=1}^{100} f[x_i]$$

$$\int_0^1 \left(\sqrt{\sqrt{x}+x}\right)dx \approx \hat{f}$$

$$\int_0^1 \left(\sqrt{\sqrt{x}+x}\right)dx \approx 1.02576$$

Actual |Area-approx| $\approx$ 0.019539



Points Generated n = 1000

The average of $\{f[x_i]\}_{i=1}^{1000}$ is

$$\hat{f} = \frac{1}{1000}\sum_{i=1}^{1000} f[x_i] = 1.05106$$

Approximation for the integral

$$\int_0^1 \left(\sqrt{\sqrt{x}+x}\right)dx \approx \frac{1}{n}\sum_{i=1}^{n} f[x_i]$$

$$\int_0^1 \left(\sqrt{\sqrt{x}+x}\right)dx \approx \frac{1}{1000}\sum_{i=1}^{1000} f[x_i]$$

$$\int_0^1 \left(\sqrt{\sqrt{x}+x}\right)dx \approx \hat{f}$$

$$\int_0^1 \left(\sqrt{\sqrt{x}+x}\right)dx \approx 1.05106$$

Actual |Area-approx| $\approx$ 0.00575848

- An extension to integration over interval (a, b)

$$\int_a^b g(x)dx$$

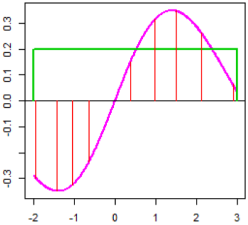
- Let $X \sim Unif(a, b)$, then the pdf of X is $\frac{1}{b-a}$ on $[a, b]$. For function $g(\cdot)$,

$$E[g(X)] = \int_a^b g(x) \cdot \frac{1}{b-a}\, dx = \frac{1}{b-a} \int_a^b g(x)\, dx$$

- Hence, $\int_a^b g(x)\, dx = (b-a)E[g(X)]$

Procedure:

- Generate n i.i.d trials $X_i \overset{i.i.d.}{\sim} Unif(a, b)$ $(1 \leq i \leq n)$ and calculate $g(X_i)$ correspondingly
- Compute the average $\frac{1}{n}\sum_{i=1}^n g(X_i)$, and $\int_a^b g(x)\, dx \approx \frac{(b-a)}{n} \sum_{i=1}^n g(X_i)$



Points Generated n = 10

The average of $\{g(x_i)\}_{i=1}^{10}$ is

$\hat{g} = \frac{1}{10}\sum_{i=1}^{10} g(x_i) = 0.09500501$

Approximation for the Integral:

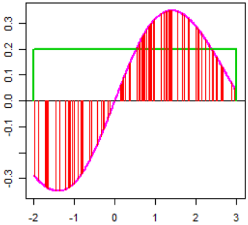
$$\int_{-2}^{3} \frac{\sin(x)}{\sqrt{x^2+6}}dx \approx (3-(-2)) * \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

$$\int_{-2}^{3} \frac{\sin(x)}{\sqrt{x^2+6}}dx \approx (3-(-2)) * \frac{1}{10}\sum_{i=1}^{10} g(x_i)$$

$$\int_{-2}^{3} \frac{\sin(x)}{\sqrt{x^2+6}}dx \approx 5 * \hat{g}$$

$$\int_{-2}^{3} \frac{\sin(x)}{\sqrt{x^2+6}}dx \approx 5 * 0.09501 = 0.47503$$

Actual $|\text{Area} - \text{approx}| \approx 0.3083$



Points Generated n = 100

The average of $\{g(x_i)\}_{i=1}^{100}$ is

$\hat{g} = \frac{1}{100}\sum_{i=1}^{100} g(x_i) = 0.04916243$

Approximation for the Integral:

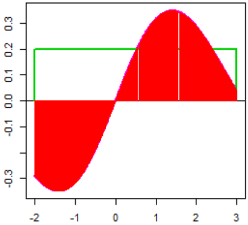
$$\int_{-2}^{3} \frac{\sin(x)}{\sqrt{x^2+6}}dx \approx (3-(-2)) * \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

$$\int_{-2}^{3} \frac{\sin(x)}{\sqrt{x^2+6}}dx \approx (3-(-2)) * \frac{1}{100}\sum_{i=1}^{100} g(x_i)$$

$$\int_{-2}^{3} \frac{\sin(x)}{\sqrt{x^2+6}}dx \approx 5 * \hat{g}$$

$$\int_{-2}^{3} \frac{\sin(x)}{\sqrt{x^2+6}}dx \approx 5 * 0.04916 = 0.24581$$

Actual $|\text{Area} - \text{approx}| \approx 0.0790$



Points Generated n = 1000

The average of $\{g(x_i)\}_{i=1}^{1000}$ is

$\hat{g} = \frac{1}{1000}\sum_{i=1}^{1000} g(x_i) = 0.03862086$

Approximation for the Integral:

$$\int_{-2}^{3} \frac{\sin(x)}{\sqrt{x^2+6}}dx \approx (3-(-2)) * \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

$$\int_{-2}^{3} \frac{\sin(x)}{\sqrt{x^2+6}}dx \approx (3-(-2)) * \frac{1}{1000}\sum_{i=1}^{1000} g(x_i)$$

$$\int_{-2}^{3} \frac{\sin(x)}{\sqrt{x^2+6}}dx \approx 5 * \hat{g}$$

$$\int_{-2}^{3} \frac{\sin(x)}{\sqrt{x^2+6}}dx \approx 5 * 0.03862 = 0.19310$$

Actual $|\text{Area} - \text{approx}| \approx 0.0263$

- Further extension to double integration over interval $(a,b)\times(c,d)$

$$I(g) = \int_a^b \left(\int_c^d g(x,y)dy \right) dx$$

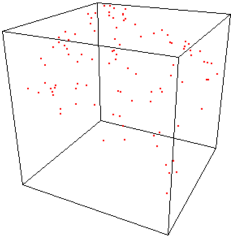
- The products of $(d-c)$, $(b-a)$, and the expectation of $g(X,Y)$ where $X \sim Unif(a,b)$, and $Y \sim Unif(c,d)$.

$$\begin{aligned} (d-c)(b-a)E(g(X,Y)) &= (d-c)(b-a)\int_a^b \int_c^d g(x,y)\cdot\frac{1}{b-a}\cdot\frac{1}{d-c}\,dy\,dx \\ &= \int_a^b \int_c^d g(x,y)dydx \end{aligned}$$

Procedure:

- Generate i.i.d. trials (X_i,Y_i) $(1\le i\le n)$ in the rectangle $(a,b)\times(c,d)$ and calculate $g(X_i,Y_i)$ correspondingly
- Compute the mean $\frac{1}{n}\sum_{i=1}^n g(X_i,Y_i)$,and $\hat{I}(g) \approx \frac{(b-a)*(d-c)}{n}\sum_{i=1}^n g(X_i,Y_i)$ (area $(b-a)$ is replaced by volume $(b-a)*(d-c)$)

Integral of $\int_0^{5/4}\left(\int_0^{5/4}(4-x^2-y^2)dy\right)dx$



Points Generated $n = 100$

The average of $\{f[x_i]\}_{i=1}^{100}$ is

$$\hat{f} = \frac{1}{100}\sum_{i=1}^{100} f[x_i] = 2.93043$$

Approximation for the integral

$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx (b-a)*(d-c)*\left(\frac{1}{n}\sum_{i=1}^{n} f[x_i,y_i]\right)$$

$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx (b-a)*(d-c)*\frac{1}{100}\sum_{i=1}^{100} f[x_i]$$

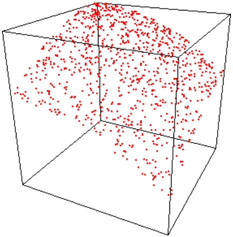
$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx (b-a)*(d-c)*\hat{f}$$

$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx \left(\frac{25}{16}\right)*(2.93043)$$

$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx 4.57879$$

The 'error estimate' $\approx$ 0.105135

Actual |Volume-approx| $\approx$ 0.0436045



Points Generated $n = 1000$

The average of $\{f[x_i]\}_{i=1}^{1000}$ is

$$\hat{f} = \frac{1}{1000}\sum_{i=1}^{1000} f[x_i] = 2.93203$$

Approximation for the integral

$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx (b-a)*(d-c)*\left(\frac{1}{n}\sum_{i=1}^{n} f[x_i,y_i]\right)$$

$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx (b-a)*(d-c)*\frac{1}{1000}\sum_{i=1}^{1000} f[x_i]$$

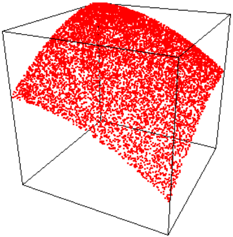
$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx (b-a)*(d-c)*\hat{f}$$

$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx \left(\frac{25}{16}\right)*(2.93203)$$

$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx 4.58129$$

The 'error estimate' $\approx$ 0.0330176

Actual |Volume-approx| $\approx$ 0.0411014



Points Generated $n = 10000$

The average of $\{f[x_i]\}_{i=1}^{10000}$ is

$$\hat{f} = \frac{1}{10000}\sum_{i=1}^{10000} f[x_i] = 2.95871$$

Approximation for the integral

$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx (b-a)*(d-c)*\left(\frac{1}{n}\sum_{i=1}^{n} f[x_i,y_i]\right)$$

$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx (b-a)*(d-c)*\frac{1}{10000}\sum_{i=1}^{10000} f[x_i]$$

$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx (b-a)*(d-c)*\hat{f}$$

$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx \left(\frac{25}{16}\right)*(2.95871)$$

$$\int_0^{\frac{5}{4}}\left(\int_0^{\frac{5}{4}}(4-x^2-y^2)dy\right)dx \approx 4.62298$$

The 'error estimate' $\approx$ 0.0103204

Actual |Volume-approx| $\approx$ 0.000585008

Suppose that a fair coin is tossed 100 times. What is the probability that the total number of heads is no smaller than 60?

Let X be the total number of heads, then $X \sim Bin(100, 1/2)$.
We are interested in $P(X \geq 60)$

- Calculate directly means calculating 40 probs ($P(X = i)$, $i = 60, 61, \cdots$) and take the summation. COMPLICATED.
- Poisson approximation cannot be applied as $np = 100(1/2) = 50$ is too large.
- X can be seen as the summation of 100 Bernoulli trials with $p = 1/2$, and limit theorems can be applied.
    - With LLN, we only know $X/100 \xrightarrow{P} 1/2$, CANNOT get $P(X \geq 60)$
    - New Limit Theorem is required to describe the behaviour of X more accurately

Central Limit Theorem (CLT)

CLT

Let $\{X_n\} = X_1, X_2, \cdots$ be a sequence of i.i.d r.v.'s with mean $E(X)$, variance $\mathrm{Var}(X)$ and moment-generating function M defined in a neighbourhood of zero. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then

$$\frac{\bar{X}_n - E(X)}{\mathrm{SD}(X)/\sqrt{n}} \xrightarrow{d} Z,$$

where $Z \sim N(0,1)$

- It means that $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ converge to a standard normal r.v. as long as the mean, variance, and moment-generating function exist, no matter what the distribution for X_i is.
- The convergence is converge in distribution, so that we can apply it for $P(\bar{X}_n \geq c)$ (let $\mu = E(X)$, $\sigma = \mathrm{SD}(X)$).
  - $P(\bar{X}_n \geq c) = P(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \geq \frac{c-\mu}{\sigma/\sqrt{n}})$
  - As $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z$, $P(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \geq \frac{c-\mu}{\sigma/\sqrt{n}}) \to P(Z \geq \frac{c-\mu}{\sigma/\sqrt{n}}) = 1 - \Phi(\frac{c-\mu}{\sigma/\sqrt{n}})$
  - Check z-table for $\Phi(\frac{c-\mu}{\sigma/\sqrt{n}})$

Comments on CLT

CLT is the most important theorem in statistics

- CLT means that, the sample mean will be approximately normally distributed for large sample sizes, *regardless* of the distribution of the samples
- Many statistics (say, $\bar{X}_n$, $\bar{X^2}_n$) have distributions that are approximately normal, *even the population distribution is not normal* $\Longleftarrow$ The dist. of statistics can be approximated
- Statistical inference can be derived based on normality, provided the sample size is large
- In practice, it gives a very rough guideline to approximate $\bar{X}_n$ when n is large (a few hundreds or even more)
- However, the convergence is the weakest convergence, converge in distribution. With the result, for statistics (e.g., $\bar{X}_n$), we can only calculate

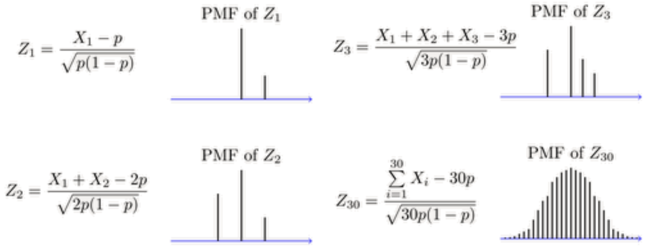
$$P(\bar{X}_n \geq a), \quad P(\bar{X}_n \leq a), \quad P(a \leq \bar{X}_n \leq b), \ldots$$

Comparison Between LLN and CLT

	LLN	CLT
Conditions	Independent Share μ and σ^2	Independent & identically dist. μ, σ^2, and mgf exist
Results	Focus on $\bar{X}_n$ $\bar{X}_n \xrightarrow{P/a.s.} \mu$	Focus on $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$ $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z$
Convergence	In probability/a.s.	In distribution
Interpretation	$\bar{X}_n$ converges to μ	The rate that $\bar{X}_n$ converges to μ
Usage	Monte Carlo Method	Statistical Inference

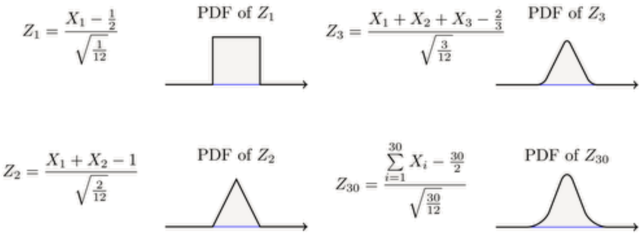
- LLN and CLT are not contradictory. LLN says $\bar{X}_n \to E(X) = \mu$. CLT evaluates the enlarged difference $(\bar{X}_n - E(X)) \times \sqrt{n}/\mathrm{SD}(X)$, which is a more detailed description
- Compare to real numbers, LLN means that $\frac{2\sqrt{n}+1}{\sqrt{n}} \to 2$, and CLT means that $\sqrt{n}(\frac{2\sqrt{n}+1}{\sqrt{n}} - 2) \to 1$.

- Let $X_i \sim Ber(p)$, $i = 1, 2, \cdots$ with mean $E(X) = p$ and $\sigma = \sqrt{p(1-p)}$, then $n\bar{X}_n \sim Bin(n, p)$, where $p = 1/3$.
- The PMF for $Z_i = \frac{\bar{X}_n - p}{\sqrt{p(1-p)}/\sqrt{n}} = \frac{n\bar{X}_n - np}{\sqrt{np(1-p)}}$. Figures are for Z_n when $n = 1, 2, 3, 30$.
- When $n = 30$, Z_{30} is very close to normal.

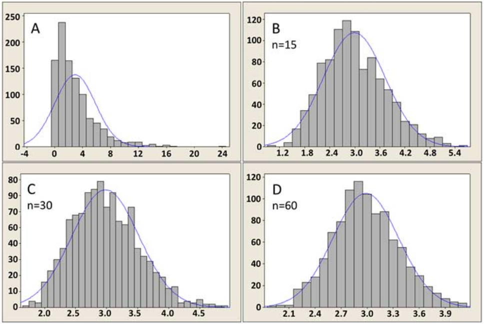


CLT - Illustration, II

- Let $X_i \sim Unif(0,1)$, $i = 1, 2, \cdots$ with mean $E(X) = 1/2$ and $\sigma = \sqrt{1/12}$.
- The PMF for $Z_i = \frac{\bar{X}_n - 1/2}{\sqrt{1/12}/\sqrt{n}} = \frac{n\bar{X}_n - n/2}{\sqrt{n/12}}$. Figures are for Z_n when $n = 1, 2, 3, 30$.
- When $n = 30$, Z_{30} is very close to normal.



- draw samples from an "unknown" distribution, but clearly skewed to the right (A)
- histogram of sample mean of different sizes are illustrated in B, C and D



Suppose that a fair coin is tossed 100 times. What is the probability that the total number of heads is no smaller than 60?

Let X_i be 1 if a head is obtained at the ith roll and 0 otherwise $(1 \leq i \leq 100)$. The coin is fair, so $X_i \sim Ber(1/2)$, $E(X_i) = 1/2$ and $\text{Var}(X) = 1/4$. We need to find

$$P(S_{100} \geq 60) = P(\frac{S_{100}}{100} \geq \frac{60}{100}) = P(\bar{X}_{100} \geq \frac{60}{100})$$

where $S_{100} = \sum_{i=1}^{100} X_i$ and $\bar{X}_{100} = \frac{S_{100}}{100}$. By CLT, $\frac{\bar{X}_{100} - 1/2}{\sqrt{1/4}/\sqrt{100}} \overset{approx}{\sim} N(0, 1)$

$$P(\frac{\bar{X}_{100} - 1/2}{\sqrt{\frac{1}{4*100}}} \geq \frac{60/100 - 1/2}{\sqrt{\frac{1}{4*100}}})$$
$$= P(Z \geq 2) = 1 - P(Z \leq 2) = 0.0228.$$

If a fair die is rolled 30 times, find the probability that the sum obtained is between 100 and 110?

Let X_i be the number obtained at the ith roll ($1 \le i \le 30$). The die is fair, so we have $E(X_i) = 7/2$ and $E(X_i^2) = 91/6$, thus $\mathrm{Var}(X_i) = 35/12$. We need to find

$$P(100 \le S_{30} \le 110) = P(\frac{100}{30} \le \frac{S_{30}}{30} \le \frac{110}{30}) = P(\frac{100}{30} \le \bar{X}_{30} \le \frac{110}{30})$$

where $S_{30} = \sum_{i=1}^{30} X_i$ and $\bar{X}_{30} = \frac{S_{30}}{30}$. By CLT, $\frac{\bar{X}_{30} - 7/2}{\sqrt{35/12}/\sqrt{30}} \overset{approx}{\sim} N(0,1)$

$$P(\frac{100/30 - 7/2}{\sqrt{\frac{35}{12*30}}} \le \frac{\bar{X}_{30} - (7/2)}{\sqrt{\frac{35}{12*30}}} \le \frac{110/30 - 7/2}{\sqrt{\frac{35}{12*30}}})$$
$$= P(-0.53 \le Z \le 0.53) = 1 - 2P(Z \ge 0.53) = 0.4038$$

Normal Approximation to Poisson (when λ is large)

Suppose $X_n \sim \text{Poisson}(n)$ where $n \to \infty$. What does the cdf for X_n look like?

Recall that if indept. r.v.'s $X \sim Pois(a)$ and $Y \sim Pois(b)$, then $X + Y \sim Pois(a + b)$ (in Lecture 3). Hence, if we take $Y_i \overset{i.i.d}{\sim} Pois(1)$, $1 \leq i \leq n$, then $S_n = \sum_{i=1}^n Y_i \sim Pois(1 + 1 + \cdots 1) = Pois(n)$, the same with X_n.

Therefore,

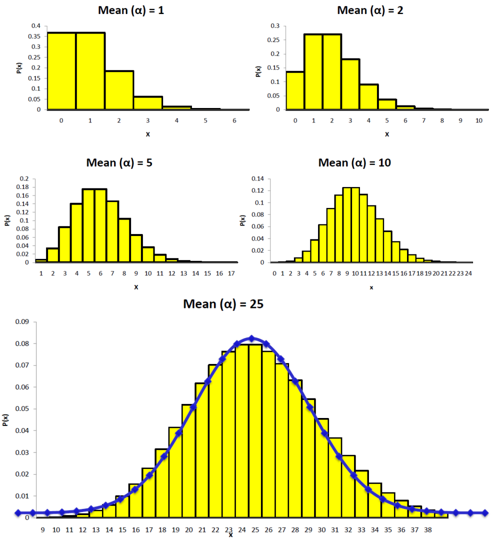
$$F_{X_n}(c) = P(X_n \leq c) = P(S_n \leq c).$$

As Y_i has mean 1 and variance 1, with CLT there is $\frac{S_n/n-1}{1/\sqrt{n}} \overset{approx}{\sim} N(0, 1)$. For fixed n, by linear transformation of normal r.v.'s, there is $S_n \overset{approx}{\sim} N(n, n)$.

Hence, we can approximate the distribution of X_n by $N(n, n)$.

Remark: When λ_n is large but not an integer, it can also be proved that $N(\lambda_n, \lambda_n)$ is a good approximation for $Pois(\lambda_n)$

Normal Approximation to Poisson - cont'd



Example: Normal Approximation

Let X be a Poisson random variable with mean 900. We find $P(X > 950)$ by standardizing:

$$P(X > 950) = P(\frac{X - 900}{\sqrt{900}} > \frac{950 - 900}{\sqrt{900}})$$

$$\approx 1 - \Phi(\frac{5}{3}) = .04779$$

The exact probability is .04712

The number of students who enroll in a psychology class is a Poisson r.v. with mean 100. The professor in charge of the course decided that if the number of enrollment is 120 or more, he will teach the course in 2 separate sessions, whereas if the enrollment is under 120 he will teach all the students in a single session. What is the probability that the professor will have to teach 2 sessions?

Let X be the enrollment in the psychology class. Given that $X \sim \text{Pois}(100)$ with $E(X) = 100 = \text{Var}(X)$. The required probability

$$P(X \geq 120) = P(\frac{X - 100}{\sqrt{100}} \geq \frac{120 - 100}{\sqrt{100}})$$

$$\approx 1 - \Phi(2) = 0.0228$$

Appendix: Properties of Normal Distribution

Suppose $X \sim N(\mu, \sigma^2)$, then

- $E(X) = \mu$, $\text{Var}(X) = \sigma^2$, $\text{SD}(X) = \sigma$
- $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$
- If $Y = a + bX$, then
  - $Y \sim N(a + b\mu, b^2\sigma^2)$
  - $M_Y(t) = e^{(a+b\mu)t + b^2\sigma^2 t^2/2}$.

If X and Y are independent, $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$, then

- $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

Suppose (X, Y) is a bivariate normal vector with parameters μ_X, μ_Y, σ_X, σ_Y, ρ, then

- $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$
- $Corr(X, Y) = \rho$, $Cov(X, Y) = \rho\sigma_X\sigma_Y$
- $X|Y = y \sim N(\mu_X + \rho\sigma_X(y - \mu_Y)/\sigma_Y, (1 - \rho^2)\sigma_X^2)$
  $Y|X = x \sim N(\mu_Y + \rho\sigma_Y(x - \mu_X)/\sigma_X, (1 - \rho^2)\sigma_Y^2)$
- $E(X|Y) \sim N(\mu_X, \rho^2\sigma_X^2)$, $E(Y|X) \sim N(\mu_Y, \rho^2\sigma_Y^2)$