MATH 321: Mathematical Statistics

Lecture 5: The Central Limit Theorem

Chapter 5: Distributions of Functions of Random Variables (5.6 and 5.7)

The Central Limit Theorem (CLT)

Introduction

- The sample mean is one statistic whose large-sample behavior is quite important. In particular, we want to investigate its limiting distribution. This is summarized in one of the most important in statistics, the central limit theorem (CLT).
- In MATH 320, we introduced the CLT and thought about it as a sum of random variables.
- **Central Limit Theorem**: Let $X_1, \ldots, X_n$ be independent random variables, all of which have the same probability distribution and thus the same mean μ and variance σ^2. If n is large, the sum

$$S = X_1 + X_2 + \cdots + X_n$$

will be approximately normal with mean $n\mu$ and variance $n\sigma^2$.

- Written succinctly: If $X_i \overset{iid}{\sim} f(x)$ with mean μ and variance σ^2, then

$$S = \sum_{i=1}^{n} X_i \overset{approx}{\sim} \text{Normal}\left(n\mu, n\sigma^2\right) \quad \text{if } n \text{ is large}$$

- We used it to solve problems like this for example:

  Suppose the number of claims filed on for a particular policy follow a Poisson distribution with a mean of 2 claims per year and the company has a portfolio of 500 active policies this year, which are assumed to be independent.

  Find the distribution of the total number of filed claims for the entire portfolio.

→ $X_i \sim \text{Poisson}(\lambda = 2)$

↳ $= E(X) = V(X)$

$n = 500$

→ $S = X_1 + \cdots + X_{500}$

$\overset{approx}{\sim} N(n\mu, n\sigma^2)$

$= 500(2), 500(2)$

$= 1000, 1000$

⇒ normalcdf( ) for probabilities

CLT - Different perspective

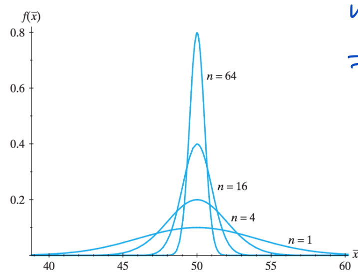
- Now we will think about the CLT from a convergence (in distribution) point of view.
- Build up to CLT / convergence in distribution.

  Let $X_i \overset{iid}{\sim} \text{Normal}\,(\mu, \sigma^2)$ and let $\bar{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.

  1. For a fixed n:

     $\bar{X}_n \sim \text{Normal}\,(\mu, \frac{\sigma^2}{n})$

  2. As $n \to \infty$:

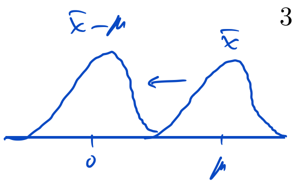
     pdfs of means of samples from $N(50, 16)$

     $V(\bar{X}_n) \to 0$

     $\Rightarrow \bar{X}_n \xrightarrow{p} \mu$

     The variance decreases until all probability is a single point.

  3. Suppose we "center the distribution" with $\bar{X}_n - \mu$:

     $\bar{X}_n - \mu \sim N(0, \sigma^2/n)$

     as $n \to \infty$, $(\bar{X}_n - \mu) \xrightarrow{p} 0$

     Even with the location adjustment, the variance of $\bar{X}_n$ disappears when the sample size n increases (without bound).

  4. We want to stop (or slow down) the "decay" of the variance, or we can think about this as spreading out the probability, so that when $n \to \infty$, $\bar{X}_n$ (and $\bar{X}_n - \mu$) does not converge to a constant, but rather distribution that still has some variation.

     To do this, we multiply the quantity of interest by a factor of n: $\sqrt{n}$

     $\underbrace{\sqrt{n}(\bar{X}_n - \mu)}_{\sim N(0,\ \sigma^2/n)} \sim \text{Normal}(0, \sigma^2)$

     $E[\sqrt{n}(\bar{X}_n - \mu)] = \sqrt{n}\, E(\bar{X}_n - \mu) = 0$

     $V[\sqrt{n}(\bar{X}_n - \mu)] = (\sqrt{n})^2\, V(\bar{X}_n - \mu) = n \frac{\sigma^2}{n} = \sigma^2$

     Now this result doesn't converge to a constant because the variance doesn't depend on n and remains "in tact" when $n \to \infty$.

     $\sqrt{n}(\bar{X}_n - \mu) \sim$ Normal regardless of n

$V[\frac{\sqrt{n}}{\sigma}(\bar{X}_n-\mu)] = \frac{n}{\sigma^2}(\frac{\sigma^2}{n}) = 1$

5. Then, we standardize the variance (adjusting the scale) by dividing by σ:

$X_i \overset{iid}{\sim}$ Normal $\qquad \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim$ Normal$(0,1) = Z$

6. Lastly, it turns out that regardless of the distribution of X_i (so we are dropping the normal assumption), this result is always true!

$X_i \overset{iid}{\sim} f(x) \qquad \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} Z$ as $n \to \infty$

This is the CLT.

- Convergence in distribution
  - Definition: A sequence of random variables, $Y_1, Y_2, \ldots$, **converges in distribution** to a random variable Y if

$$\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$$

at all points y where $F_Y(y)$ is continuous (notation: $Y_n \xrightarrow{d} Y$).

  - Although we talk of a sequence of random variables converging in distribution, it is really **the cdfs that converge, not the random variable (or statistic)**.

So for the CLT, we technically have:

$$\lim_{n\to\infty} \quad F_{Y_n}(y) \quad = \quad F_Y(y)$$

$$\lim_{n\to\infty} \quad F\left[\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)\right] = F_Z(z)$$

Restating CLT

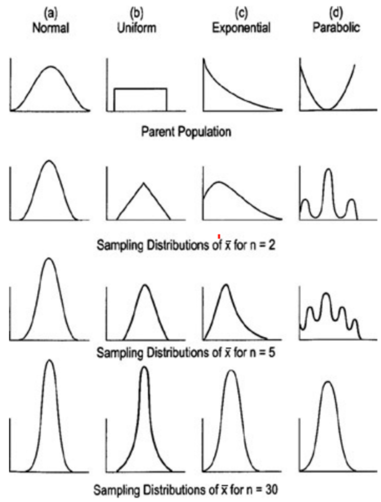
- **Theoretical result**

Central Limit Theorem: Let $X_i \overset{iid}{\sim} f(x)$ with $E(X) = \mu$ and $V(X) = \sigma^2 > 0$. Then the distribution of

$$W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}\,(0,1) \quad \text{as } n \to \infty$$

- **In practice**

This means for **any** random variable X with $E(X) = \mu$ and $V(X) = \sigma^2 > 0$, as n gets larger the distribution of $W = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ can be more closely approximated by the standard normal distribution.



Distribution of sample mean, $\overline{X}$

- Using this practical result

  $Z = \frac{X-\mu}{\sigma}$

  - Needed theorem: If $Z \sim \mathrm{N}(0,1)$, and μ and $\sigma > 0$ are constants, then

$$\boxed{X = \sigma Z + \mu \sim \mathrm{N}(\mu, \sigma^2)}$$

Proof: $M_X(t) = M_{\sigma Z+\mu}(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{\frac{1}{2}(\sigma t)^2}$

$= e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

$\sim \mathrm{Normal}(\mu, \sigma^2)$

  - Results

$W = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \Rightarrow \bar{X}_n = \frac{\sigma}{\sqrt{n}} W + \mu \Rightarrow \bar{X}_n \approx \frac{\sigma}{\sqrt{n}} Z + \mu$

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

$\overset{\text{approx}}{\sim} N(\mu, \frac{\sigma^2}{n})$

(a) $\frac{\sigma}{\sqrt{n}} W + \mu = \bar{X}$ can be approximated by $\frac{\sigma}{\sqrt{n}} Z + \mu \sim \mathrm{Normal}\,(\mu, \frac{\sigma^2}{n})$ for "large" n

$\bar{X}_n = \frac{1}{n} \sum X_i \Rightarrow S = n\bar{X} \approx n\left(\frac{\sigma}{\sqrt{n}} Z + \mu\right) = \underbrace{(\sigma\sqrt{n}) Z + n\mu}$ $\overset{\text{approx}}{\sim} N(n\mu, n\sigma^2)$

$= \frac{1}{n} S$

(from (a))

$E(\ \ ) = \sigma\sqrt{n}\,\cancel{\underbrace{E(Z)}_{=0}} + n\mu = n\mu$

$V(\ \ ) = (\sigma\sqrt{n})^2 \underbrace{V(Z)}_{=1} = n\sigma^2$

(b) $n\bar{X} = X_1 + \ldots + X_n = S$ can be approximated by $(\sigma\sqrt{n})Z + n\mu \sim \mathrm{Normal}\,(n\mu, n\sigma^2)$ for "large" n.

- How large must n be?
  - Although CLT gives us a useful general approximation, we have no automatic way of knowing how good the approximation is in general. In fact, the goodness of the approximation is a function of the original distribution, and so much be checked case by case.
  - The more the distribution of X (population distribution) is "like" a normal distribution (symmetric, unimodal, continuous, etc.), the smaller the n needed for $\bar{X}$ to be approximated well by a normal distribution.
  - **The rule $n \geq 30$ is a lie!!** But for "school" purposes, we can just use this rule of thumb as our check.
  - With the current availability of cheap, plentiful computing power, the importance of approximation like the CLT is somewhat lessened. However, despite its limitation, it is still a marvelous result.

Examples

1. Let $\bar{X}$ be the mean of a random sample of size 36 from Exp $(\lambda = 1/3)$. Approximate $P(2.5 \leq \bar{X} \leq 4)$.

↳ $E(X) = 1/\lambda = 3$

$V(X) = 1/\lambda^2 = 9$

$n = 36 \geq 30$ ✓

$\Rightarrow$ by CLT $\rightarrow$ $\bar{X} \overset{\text{approx}}{\sim} N\left(\mu = E(X), \sigma^2 = \frac{V(X)}{n}\right)$, with $\mu = 3$, $\sigma^2 = 9/36$

$\rightarrow P(2.5 \leq \bar{X} \leq 4) \approx \text{normalcdf}\left(\text{lower} = 2.5,\ \text{upper} = 4,\ \mu = 3,\ \sigma = \sqrt{9/36}\right) \approx 0.97721$

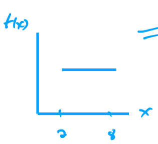
2. Let $X_1, \ldots, X_{20}$ denote a random sample of size 20 from continuous Uniform $(2, 8)$. If $S = X_1 + \ldots + X_{20}$, approximate $P(S < 95)$.

↳ $E(X) = \frac{a+b}{2} = 5$

$V(X) = \frac{(b-a)^2}{12} = \frac{36}{12} = 3$

$n = 20 \not\geq 30$

but distribution is symmetric



$\Rightarrow$ Smaller n is okay

$\Rightarrow$ by CLT $S \overset{\text{approx}}{\sim} N\left(\mu = n\,E(X) = 20(5) = 100,\ \sigma^2 = n\,V(X) = 20(3) = 60\right)$

$\rightarrow P(S < 95) \approx \text{normalcdf}(-10000, 95, 100, \sqrt{60}) \approx 0.2593$

t, Z, and the CLT

- Previously, we learned the following
  1. If $X_1, \ldots, X_n$ are a random sample for a $N(\mu, \sigma^2)$, we know that the quantity

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) = Z \quad \longrightarrow \text{Standardizing } \bar{X}$$

  2. (Building on 1.) However, if σ is unknown, we substitute S then

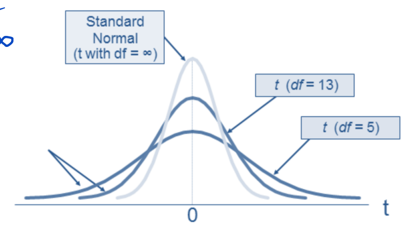
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad \longrightarrow \text{functions of random variables}$$

$$t_r = \frac{Z}{\sqrt{\chi^2_r / r}}$$

  3. (Building on 2.) As $n \to \infty$, $t_{n-1} \to Z$

$df \to \infty$

$$t_{n-1} \xrightarrow{d} Z$$



  4. If $X_1, \ldots, X_n$ are **not normal** random variables, when the sample size is large

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \overset{\text{approx}}{\sim} N(0,1) = Z \quad \longrightarrow \text{CLT} ✓$$

Approximations for discrete distributions

Continuity correction

- Motivation: Now we are going to use the CLT as an approximation tool when sampling from **discrete distributions**.

  Specifically, we will discuss a way to improve our approximations to account for the discrepancy created from using a continuous distribution / probability methods (integral to calculate area under curve) on originally discrete distributions.

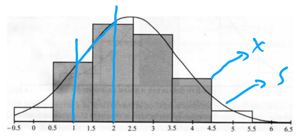
- This is called the (half unit) **continuity correction** and is demonstrated below.

discrete

ex) → $P(X = 1, 2)$

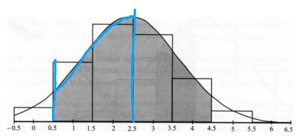
→ without correction

$= P(1 \le S \le 2)$



→ with correction

$= P(0.5 \le S \le 2.5)$



$\leftarrow -0.5$ $\quad \rightarrow +0.5$

- Estimate the following probabilities using the continuity correction:

  (a) $P(1 \le X \le 4) \approx$ $P(0.5 \le S \le 4.5)$

  (b) $P(X = 2) \approx$ $P(1.5 \le S \le 2.5)$

  (c) $P(1 \le X < 4) =$ $P(1 \le S \le 3) \approx P(1.5 \le S \le 3.5)$

  (d) $P(X > 2.5) =$ $P(X \ge 3) \approx P(S > 2.5)$

- In general, if X is the original discrete random variable of interest, S is the corresponding normal random variable based on the CLT, and a, b are some integers $(a \le b)$, then we can summarize the adjustments for the **continuity correction** with:

$$P(X = a) = P(a - 0.5 \le S \le a + 0.5)$$

$$P(a \le X \le b) = P(a - 0.5 \le S \le b + 0.5)$$

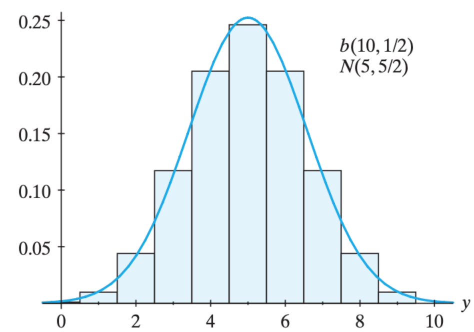
  Just need to take care to decide if we are want to include or exclude a or b (so can rewrite strict inequalities $<$ $>$ as inclusive $\le$ $\ge$ and then use rule).

Normal approximation to the binomial distribution

- The most common scenario when applying the normal approximation is to the binomial distribution.
- Recall if $X \sim \text{Binomial}(n, p)$ $\implies X = \sum_{i=1}^{n} Y_i$, where $Y_i \overset{iid}{\sim} \text{Bernoulli}(p)$ $\implies E(X) = p$, $V(X) = pq$

  ↳ use CLT for sums
- This means for "large n", $X \approx S \overset{approx}{\sim} N\left(\mu = nE(X) = np,\ \sigma^2 = nV(X) = npq\right)$

Condition ☆ "Rule of thumb" is that n is sufficiently large if $np \geq 5$ and $n\underbrace{(1-p)}_{=q} \geq 5$.
We will discuss reasoning behind this after some examples.



- Examples→ Good use[redacted] of the normal approximation

1. [redacted] Suppose that a multiple choice exam has 40 questions, each with 5 possible answers. A student feels that he has a probability of 0.55 of getting any particular question correct, with independence from one question to another.

   Approximate the probability of the student getting at least 25 correct.

→ $X \sim \text{Bin}(n = 40, p = 0.55)$

→ check → $np = 40(0.55) = 22 \geq 5$ ✓

→ $nq = 40(0.45) = 18 \geq 5$ ✓

$\underset{\text{by CLT}}{\implies}$ $X = S \overset{approx}{\sim} N\left(\mu = 40(0.55) = 22,\ \sigma^2 = 40(0.55)(0.45) = 9.9\right)$

(a) With continuity correction:

$P(X \geq 25) \approx P(S \geq 24.5) =$ normalcdf $\rightarrow 0.2134$

(b) Without continuity correction:

$P(X \geq 25) = P(S \geq 25) = 0.1701$ ✗

close ✓

(c) Exact answer using binomial distribution:

$P(X \geq 25) = 1 - P(X \leq 24)$

$= 1 -$ Binomcdf$(n = 40, p = 0.55, x = 24) = 0.2142$

- Bad example: Suppose we change the scenario in example 1 so that there are only 20 questions and the probability of getting any particular question correct is now 0.10.

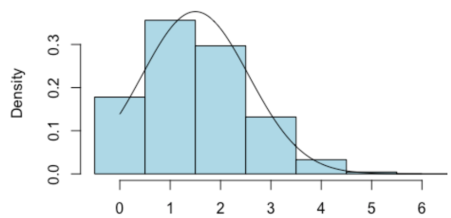
  Compare the approximate answer and exact answer for $P(X \geq 3)$.

(assume check ✓) $\Rightarrow X = S \overset{approx}{\sim} N\left(\mu = 20(0.1) = 2,\ \sigma^2 = 20(0.1)(0.9) = 1.8\right)$

$\rightarrow$ approx $\rightarrow$ $P(X \geq 3) \approx P(S \geq 2.5) = 0.3545$

$\rightarrow$ Exact $\rightarrow$ $P(X \geq 3) = 0.323$ ✗ even with continuity correction

- Why is the approximation bad??
  - Lets take a look at the histogram and overlaid normal pdf for a similar scenario:



$X \sim Binomial(n = 6, p = 0.25)$

  - This illustrates the mismatch between the skewed probability histogram for and the symmetric pdf of the normal distribution. In order to do a good job of approximating the binomial distribution, the normal curve must have the bulk of its own distribution between legitimate outcomes for the Binomial distribution $[0, n]$.
  - How do we apply / check this: Based on the empirical rule, the central 95% of any normal distribution lies within two standard deviations of its mean.

left side

$\mu - 2\sigma > 0$

$np - 2\sqrt{npq} > 0$

$n^2p^2 > 4npq$

$np > 4q = 4(1-p) = 4 - 4p$

$np > 4 - 4p$

right side

$\mu + 2\sigma < n$

$np + 2\sqrt{npq} < n$

$2\sqrt{npq} < n - np = n(1-p) = nq$

$4npq < n^2q^2$

$4p < nq$

$4(1-q) = 4 - 4q < nq$

  - Thus, as long as we ensure that $np \geq 5$ & $nq \geq 5$, the normal approximation to the a binomial distribution will be good.
  - ☆ Contextually, this condition means that we must expect (expected value) the number of success (np) and failures (nq) to be at least 5.

- This relationship between a large sample binomial and normal is important for confidence intervals and hypothesis tests of population proportions which we will cover next.

Normal approximation to the Poisson distribution

- A Poisson distribution with large enough mean can also be approximated with the use of a normal distribution.

  Let $X \sim \text{Poisson}(\lambda)$, with $E(X) = V(X)$, where [redacted] λ [redacted] $= 1, 2, \ldots$ (in general, we just need $\lambda > 0$, but for demonstration lets assume λ is a positive integer).

- We can rewrite X as a sum of Poisson random variables:

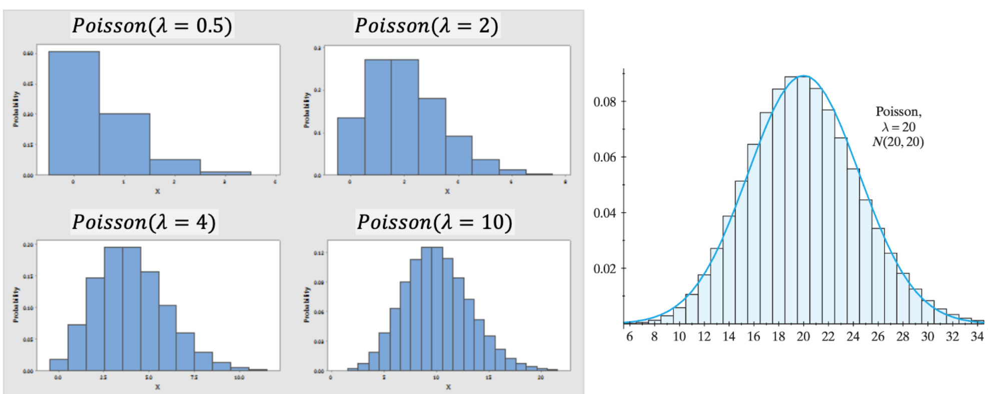
  $X = \sum_{i=1}^{\lambda} Y_i$, where $Y_i \overset{iid}{\sim} \text{Poisson}(\lambda_Y = 1)$

  Theorem

  $\sum \text{Poisson}(\lambda) = \text{Poisson}(\sum \lambda)$

- This means for "large n", X [redacted] $S \overset{approx}{\sim} N\left(\mu = nE(Y) = \lambda(1) = \lambda,\ \sigma^2 = nV(Y) = \lambda(1) = \lambda\right)$

  "Rule of thumb" is that n is sufficiently large if λ [redacted] ≥ 10 (doesn't need to be an integer).



Poisson probability histograms

- Example

  Let X equal the number of alpha particles emitted by barium-133 per second and counted by a Geiger counter. Assume that $X \sim \text{Poisson}(\lambda = 49)$.

  Approximate $P(45 \leq X < 60)$.

check $\rightarrow$ $\lambda = 49 \geq 10$ ✓ $\underset{\text{by CLT}}{\Longrightarrow}$ $X = S \overset{approx}{\sim} N(\mu = 49, \sigma^2 = 49)$

$\rightarrow P(45 \leq X < 60) \approx P(44.5 \leq S \leq 59.5) = 0.6730$

Summary of normal approximation to the binomial and Poisson distributions

- Suppose n is large and $a = 0, 1, \ldots, n$
  (a) CLT:

  If $X \sim \text{Binomial}\,(n, p) \Longrightarrow X \approx S \sim \text{Normal}\,(\mu = np, \sigma = \sqrt{npq})$

  If $X \sim \text{Poisson}\,(\lambda) \Longrightarrow X \approx S \sim \text{Normal}\,(\mu = \lambda, \sigma = \sqrt{\lambda})$

  (b) Continuity correction:

  $P(X \leq a) \approx \text{Normalcdf(lower} = 0, \text{upper} = a + 0.5)$

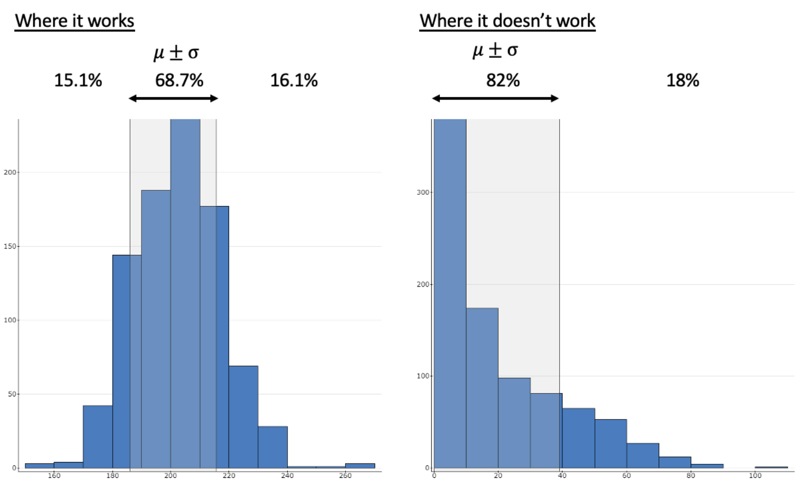
  $P(X < a) \approx \text{Normalcdf(lower} = 0, \text{upper} = a - 0.5)$

- Final note: In practice, if you have technology / software, just compute discrete probabilities exactly. However, it is important to learn how to apply the central limit theorem.

Central interval probabilities

Empirical rule

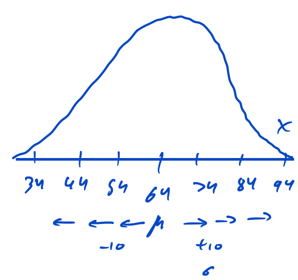
- Motivation: Because the normal distribution can be used in so many scenarios due to the CLT, there are common generalizations that are made about **central interval** probabilities for distributions that are approximately bell-shaped.
- First, lets calculate these exactly for the standard normal curve. These will of course apply to any normal distribution X with mean μ and standard deviation σ because we can standardize to get Z.
  1. $P(-1 \leq Z \leq 1) =$ $P(|Z| \leq 1) = P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.6832$
  2. $P(|Z| \leq 2) =$ 0.95445
  3. $P(|Z| \leq 3) =$ 0.99730
- Not all data is exactly normally distributed of course, but because of the CLT many distributions can be approximated by a normal distribution. So we can use the exact probabilities above to make generalizations about these distributions that have a similar shape.
- The **empirical rule** states that for approximately normal distribution:
  1. Approximately 68% of data falls within 1 standard deviation of the mean. $\mu \pm \sigma$
  2. Approximately 95% of data falls within 2 standard deviations of the mean. $\mu \pm 2\sigma$
  3. Approximately 99.7% (nearly all) of data falls within 3 standard deviations of the mean. $\mu \pm 3\sigma$



- Example: Suppose that the scores on an achievement test are known to have, approximately, a normal distribution with mean $\mu = 64$ and standard deviation $\sigma = 10$.

  (a) Find the scores probability scores are between 54 and 74.

  $(54, 74) = \mu \pm \sigma \implies 68\%$

  (b) Find which two values lies the central 95%?

  $95\% \implies \mu \pm 2\sigma \implies (44, 84)$

  (c) Find the percent of scores above 94.

  $94 = \mu + 3\sigma \implies \underbrace{100 - 99.7\% = 0.3\%}_{\text{outside}} \quad \underbrace{0.3\%/2}_{\text{ONLY right}} = 0.15\%$

- Thus, knowledge of the mean and the standard deviation gives us a fairly good picture of the frequency distribution of scores when the bell-shape is present (or assumed).