The Normal Distribution and CLT

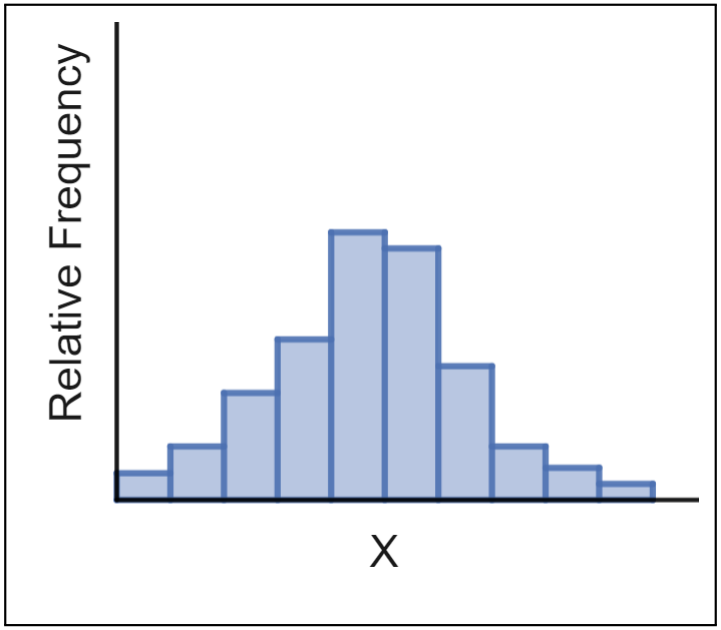
553.211 Intro to Prob. & Stats.

Johns Hopkins University

EN.553.211

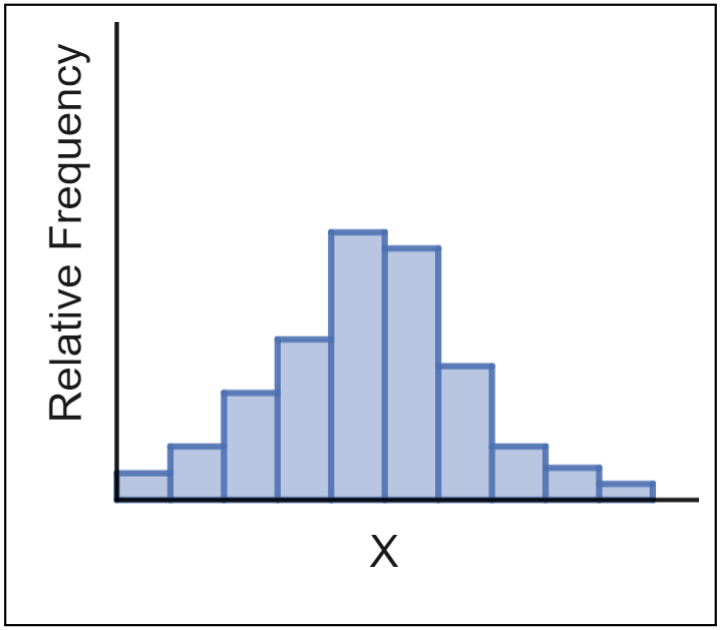
Continuous RVs

Recall that we can represent discrete distributions with probability histograms given by its pmf:



Continuous RVs

Recall that we can represent discrete distributions with probability histograms given by its pmf:



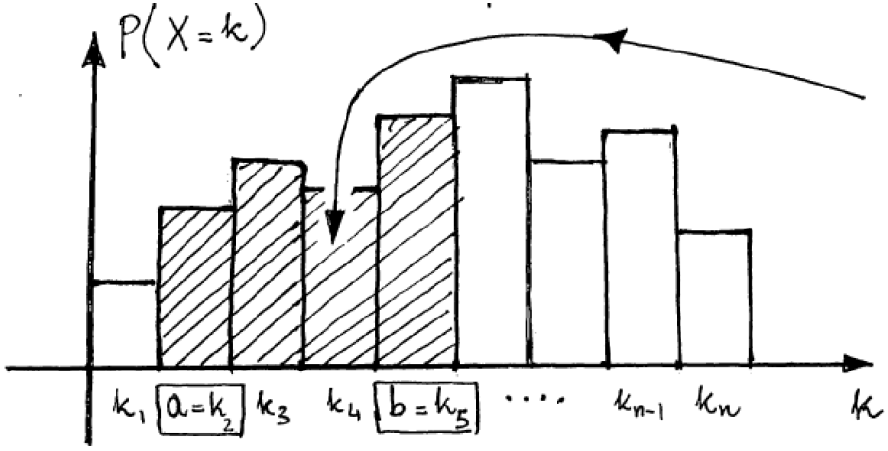
But what happens when X is a continuous RV?

Continuous RVs

- Remember the range of a continuous RV contains an interval, so there are *uncountably* many values which X can assume.
- In this case, it no longer makes sense to think about the function $P(X = x)$.

Probability of Continuous RVs

We saw that for a discrete R.V. X, with range$= \{k_1, k_2, \ldots, k_n\}$, we can illustrate the probability distribution with a histogram:

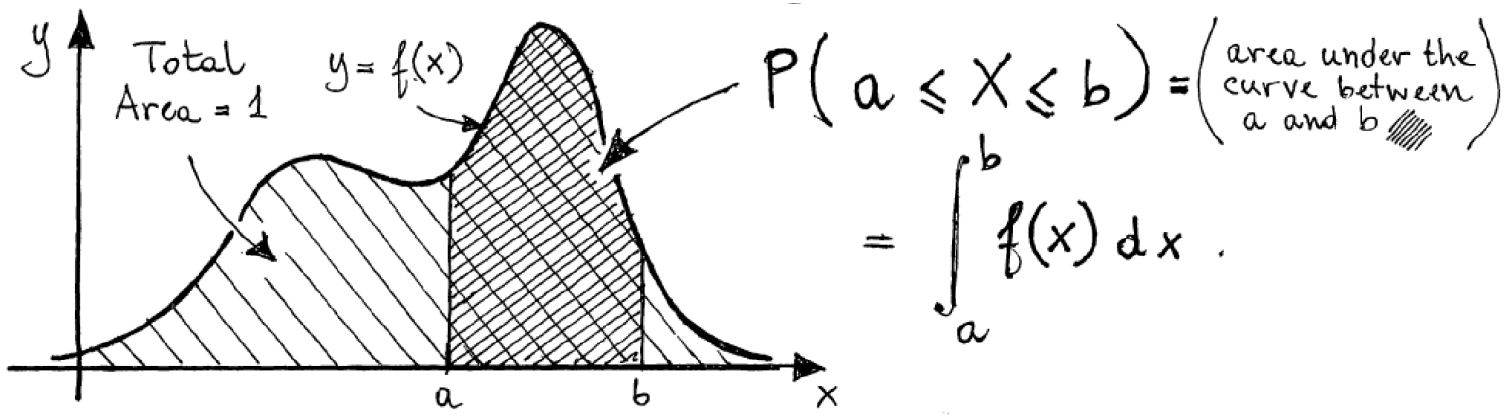


$$P(a \leq X \leq b) = \sum_{k=a}^{b} P(X=k)$$

(which is the area of the histogram in the shaded region)

Probability of Continuous RVs

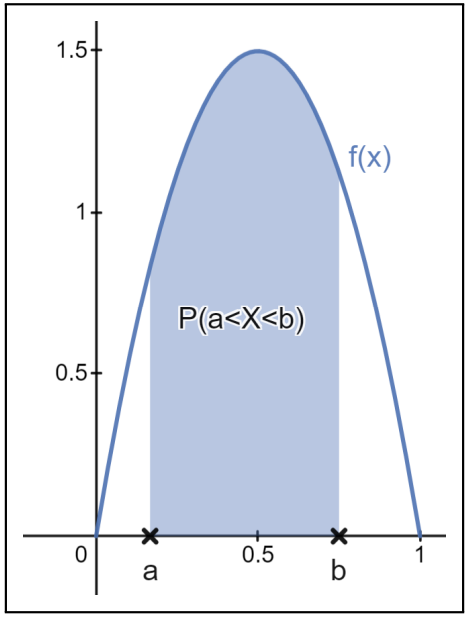
A continuous R.V. X has a continuous range (e.g. an interval $[c, d]$ or the entirety of the real axis $\mathbb{R}$) and instead of histogram, they are described by a **probability density function** $y = f(x)$, which may be interpreted as a continuous histogram.
And $P(a \leq X \leq b)$ is technically computed using an integral (In this course we don't need to compute it using an integral. We will only use probability tables or program)



Probability of Continuous RVs

Definition

The distribution of a continuous RV, X, can be described by its **probability density function** (or **pdf** for short). $f(x)$ is a continuous, non-negative function and for any $a \leq b$, $P(a \leq X \leq b)$ is given by the area under $f(x)$ between a and b, i.e. $P(a \leq X \leq b) = \int_a^b f(x)dx$.



Basic Facts of PDFs

Fact

If $f(x)$ is a pdf of a continuous RV X, then

- $f(x) \geq 0$ *for all values of* x.
- *The total area under the graph of* f *is equal to* 1,
  i.e. $\int_{-\infty}^{\infty} f(x)dx = P(-\infty < X < \infty) = 1$
  (Analogous to the pmf of a discrete RV summing to 1*)*
- $P(X = x) = 0$ *for any* x, *as a line takes up no area!*
  This in turn implies that $P(a \leq X \leq b) = P(a < X < b)$,
  $P(X \leq a) = P(X < a)$, $P(X \geq a) = P(X > a)$, *etc*

Cumulative Distribution Function

Recall

Definition

For any RV, the **cumulative distribution function** (or **cdf** for short) is given by

$$F(x) = P(X \leq x).$$

Note: the definition for cdf is the same for a discrete R.V. and a continuous R.V.

Cumulative Distribution Function

Recall

Definition

For any RV, the **cumulative distribution function** (or **cdf** for short) is given by

$$F(x) = P(X \leq x).$$

Note: the definition for cdf is the same for a discrete R.V. and a continuous R.V.

So for a continuous RV X with pdf $f(x)$, we have cdf

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Normal Distribution

Normal (a.k.a. Gaussian) distribution is a very important and useful family of distributions for continuous R.Vs.

Motivation of Normal Distribution

Motivation:
Suppose $N = \#$ of successes out of n independent trials, where n is very large, and p is the probability of success at each trail. We have that

$$N \sim \text{Binomial}(n, p)$$

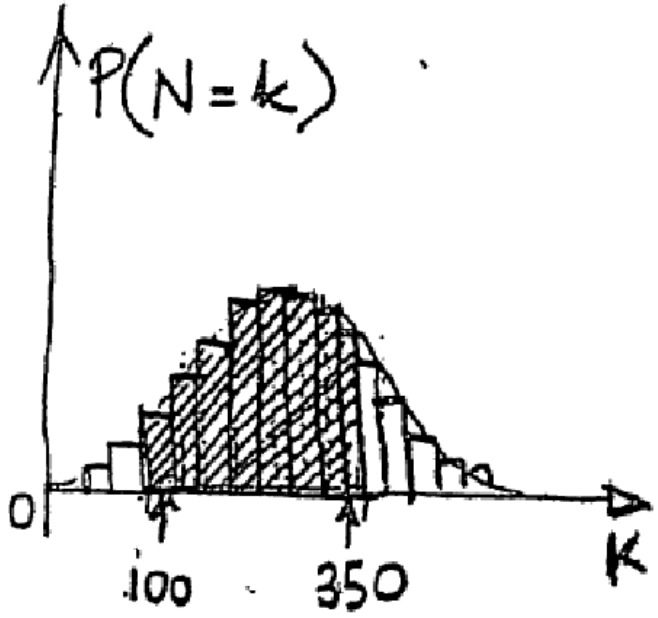
For example, we may have $n = 1000, p = 1/3$. When n is this large, computations may become very complex. If we wanted to compute, say

$$P(100 \leq N \leq 350) = \sum_{k=100}^{350} \binom{1000}{k} (1/3)^k (2/3)^{1000-k},$$

we'd have to calculate 251 binomial coefficients of the form $\binom{1000}{k}$, and then perform a large number of multiplications and sums.

Motivation of Normal Distribution

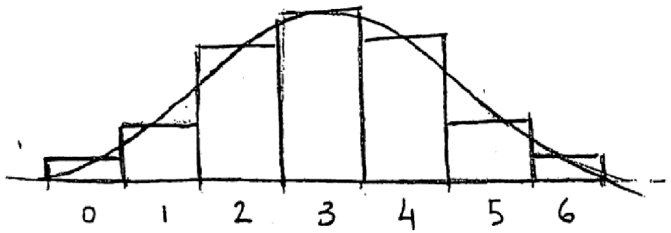
The above motivational example makes it clear that it would be convenient to have a method to compute probabilities for a binomial random variable, at least approximately, in an efficient manner (that does not require that many operations)



Motivation of Normal Distribution

For the above reason, we shall introduce Normal R.V. (also known as "Gaussian" R.V)

We have that the histogram of a Binomial(n, p) may be approximated by a continuous curve:



In fact, one can prove that such curve has equation

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}},$$

where μ and σ are, respectively, the mean and standard deviation of the Binomial R.V.

Normal Distribution

Definition

A continuous RV X has the **normal distribution** with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, if its pdf is given by

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

for any $x \in \mathbb{R}$.
We write

$$X \sim \mathcal{N}(\mu, \sigma)$$

Note: Do *not* memorize this formula! It is interesting, but ultimately not important for solving problems.

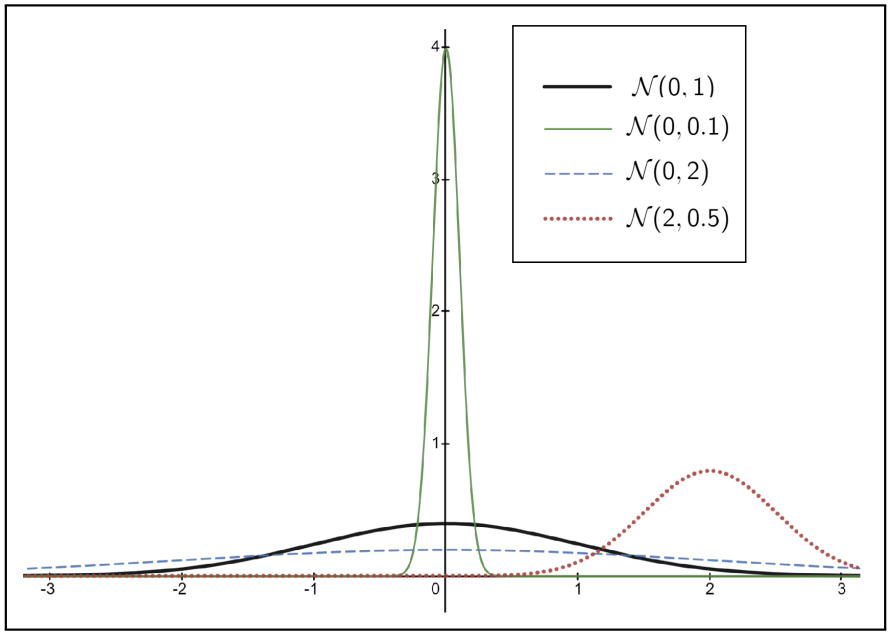
Fact

For $X \sim \mathcal{N}(\mu, \sigma)$, μ is the **expectation**, *while σ is the* **standard deviation**.
(The same notations for mean and standard deviation for Week 1)

Center and Shape

The graph of $y = f(x)$ has the following properties:

- It is centered (and symmetric with respect to) $x = \mu$.
  Changing μ will shift the graph left or right.
- As σ increases, the graph becomes more "spread out" with respect to the mean.
  Changing σ will contract or expand the graph:



Properties of Normal distribution

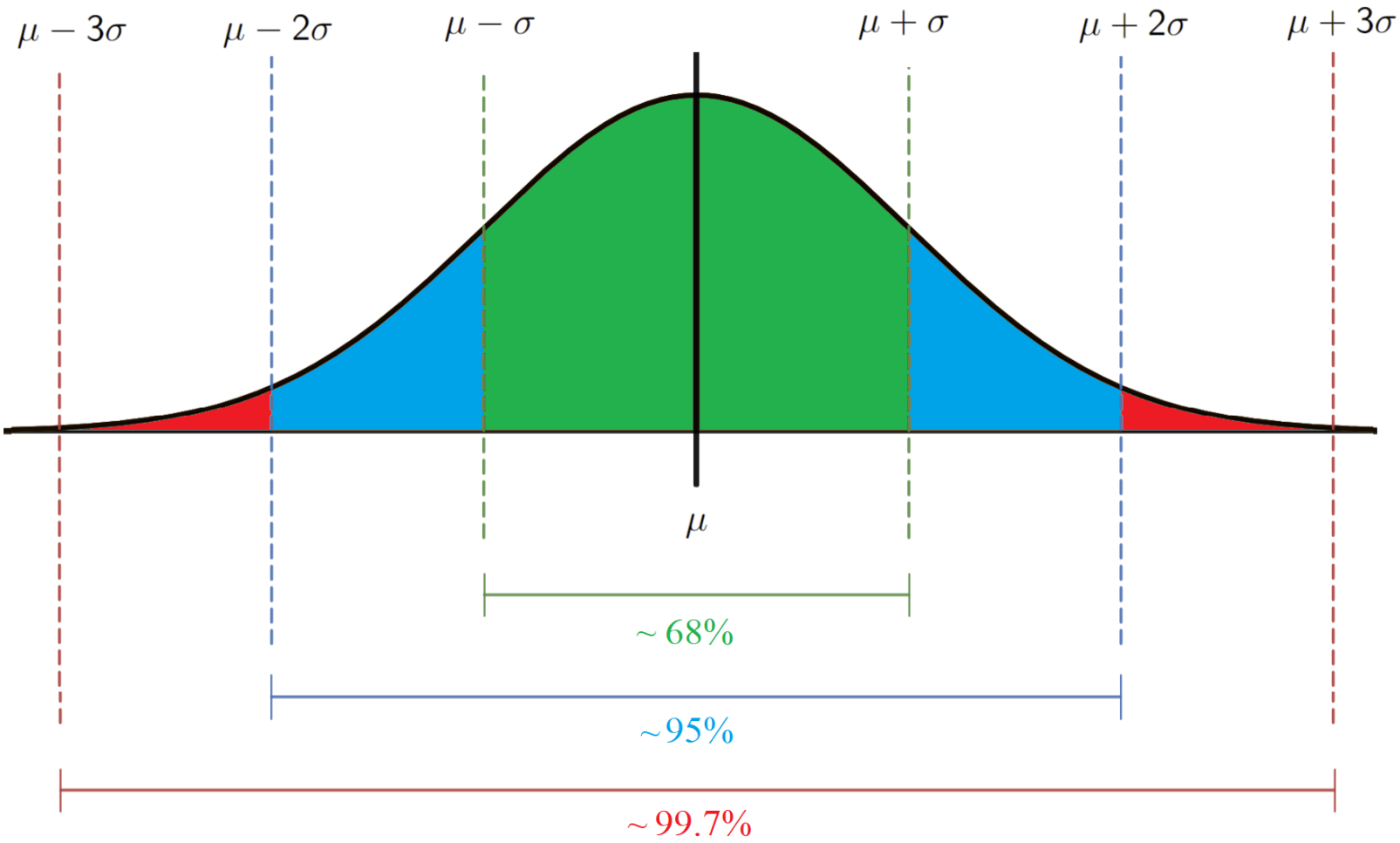
Fact

For $X \sim \mathcal{N}(\mu, \sigma)$

- *Total under the curve*$= P(-\infty < X < \infty) = 1$
- *Area below* μ *and above* μ
  $= P(X < \mu) = P(X > \mu) = 0.5$
- *Area between* $\mu - \sigma$ *and* $\mu + \sigma$
  $= P(\mu - \sigma < X < \mu + \sigma) \approx 0.68$
- *Area between* $\mu - 2\sigma$ *and* $\mu + 2\sigma$
  $= P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95$
- *Area between* $\mu - 3\sigma$ *and* $\mu + 3\sigma$
  $= P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.997$

Recall the empirical rule in Week 1

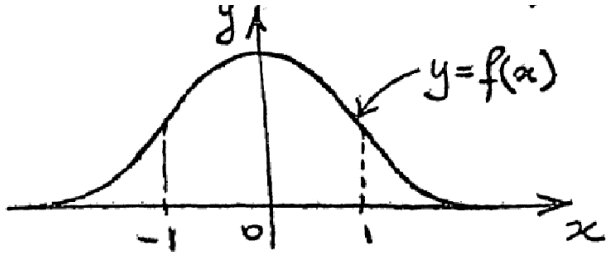
Empirical Rule



Standard Normal Distribution

- A very special and important case is when $\mu = 0$ and $\sigma = 1$
- In this case we say that $Z \sim \mathcal{N}(0, 1)$ is the **standard normal distribution** and we have the **standard normal probability density function:**

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$



- Note that in the standard normal case we normally use the symbol Z to indicate the random variable, instead of the generic X

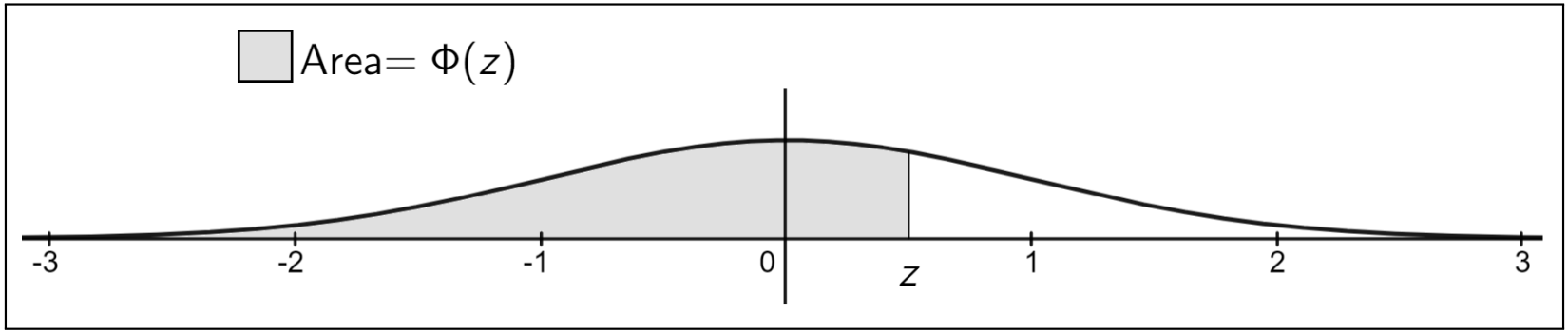
Cdf of Std. Norm. Dist. $\Phi(z)$

How do we compute probabilities in the standard normal case?

Definition

The cdf of a RV $Z \sim \mathcal{N}(0,1)$ distribution is denoted by

$$\Phi(z) = P(Z \leq z)$$



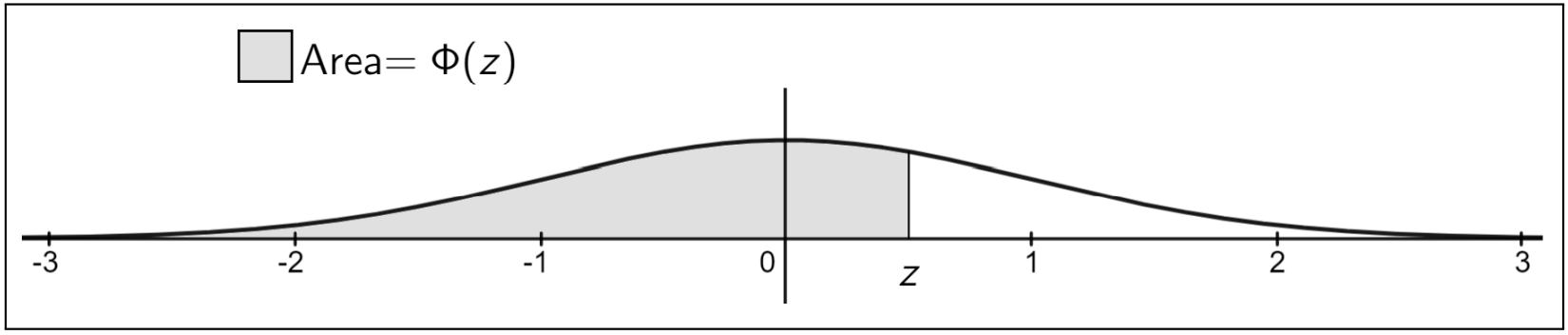
Cdf of Std. Norm. Dist. $\Phi(z)$

How do we compute probabilities in the standard normal case?

Definition

The cdf of a RV $Z \sim \mathcal{N}(0,1)$ distribution is denoted by

$$\Phi(z) = P(Z \leq z)$$



Note Z is the RV, while z is a constant. So $\Phi(z) = P(Z \leq z)$ is the probability that the RV Z is less than the constant z.

Normal Table for $\Phi(z)$

- There is NO closed formula for $\Phi(z)$, crazy right?! It's just an integral after all ($\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} f(x)dx$), but there really is no algebraic formula we can just plug z into.

[1] Posted in Canvas in the Formulas Module

Normal Table for $\Phi(z)$

- There is NO closed formula for $\Phi(z)$, crazy right?! It's just an integral after all ($\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} f(x)dx$), but there really is no algebraic formula we can just plug z into.
- Typically we can either use technology, or the Normal Table[1].

100th's place

	.0	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517

10th's place

$\approx \Phi(0.33)$

[1] Posted in Canvas in the Formulas Module

Properties of $\Phi(z)$

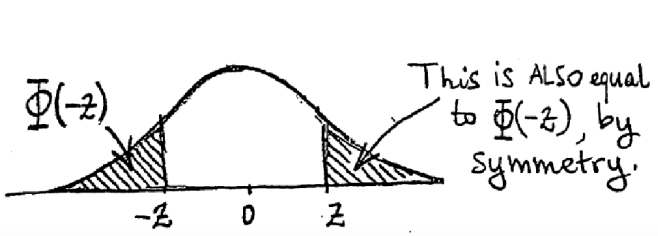
The normal table somtimes tabulates $\Phi(z)$ only for positive values of the constant z. For negative values of z, we use the symmetry of $f(z)$ with respect to the mean $z = 0$ and it follows that

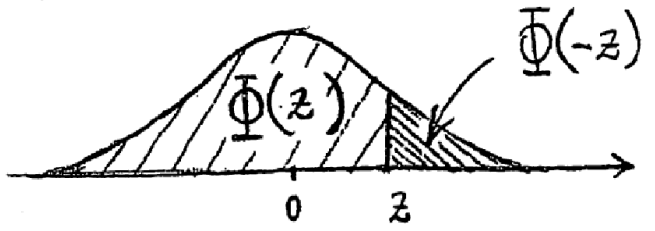
Fact

For the cdf $\Phi(z)$ of a RV $Z \sim \mathcal{N}(0, 1)$

$$\Phi(-z) = 1 - \Phi(z)$$

for any z





E.g.
$P(Z < -0.44) =$

Properties of $\Phi(z)$

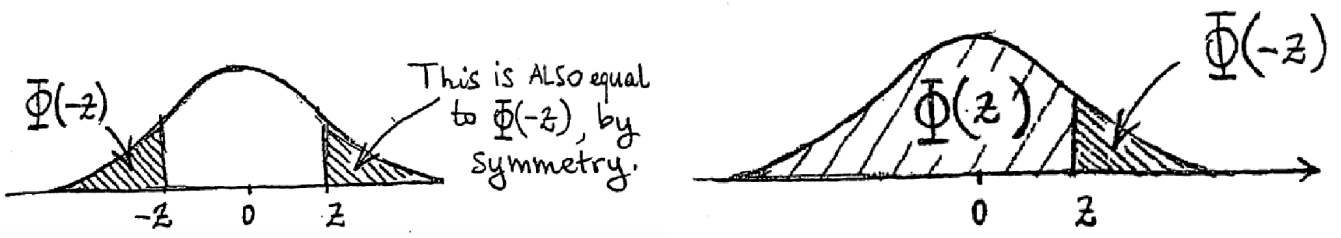
The normal table somtimes tabulates $\Phi(z)$ only for positive values of the constant z. For negative values of z, we use the symmetry of $f(z)$ with respect to the mean $z = 0$ and it follows that

Fact

For the cdf $\Phi(z)$ of a RV $Z \sim \mathcal{N}(0, 1)$

$$\Phi(-z) = 1 - \Phi(z)$$

for any z



E.g.

$P(Z < -0.44) = \Phi(-0.44) = 1 - \Phi(0.44) = 1 - 0.6700 = 0.33$

Properties of $\Phi(z)$

Fact

For the cdf $\Phi(z)$ of a RV $Z \sim \mathcal{N}(0,1)$, we have, for any $a \leq b$

$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a).$$

Properties of $\Phi(z)$

Example (Basic Technique)

Let $Z \sim \mathcal{N}(0, 1)$, compute the probability

$$P(-0.38 \leq Z \leq 1.21).$$

Properties of $\Phi(z)$

Example (Basic Technique)

Let $Z \sim \mathcal{N}(0, 1)$, compute the probability

$$P(-0.38 \leq Z \leq 1.21).$$

1. We first have

$$P(-0.38 \leq Z \leq 1.21) = \Phi(1.21) - \Phi(-0.38)$$

then we obtain $\Phi(0.38)$ and $\Phi(1.21)$, respectively.

Properties of $\Phi(z)$

Example (Basic Technique)

Let $Z \sim \mathcal{N}(0, 1)$, compute the probability

$$P(-0.38 \leq Z \leq 1.21).$$

1. We first have

$$P(-0.38 \leq Z \leq 1.21) = \Phi(1.21) - \Phi(-0.38)$$

then we obtain $\Phi(0.38)$ and $\Phi(1.21)$, respectively.

2. From symmetry we know

$$\Phi(-0.38) = 1 - \Phi(0.38) \approx 1 - 0.6480 = 0.3520,$$

where $\Phi(0.38)$ is taken from the table.

Properties of $\Phi(z)$

Example (Basic Technique)

Let $Z \sim \mathcal{N}(0, 1)$, compute the probability

$$P(-0.38 \leq Z \leq 1.21).$$

1. We first have
$$P(-0.38 \leq Z \leq 1.21) = \Phi(1.21) - \Phi(-0.38)$$
then we obtain $\Phi(0.38)$ and $\Phi(1.21)$, respectively.
2. From symmetry we know
$$\Phi(-0.38) = 1 - \Phi(0.38) \approx 1 - 0.6480 = 0.3520,$$
where $\Phi(0.38)$ is taken from the table.
3. The value $\Phi(1.21) \approx 0.8869$ is also easily found in the table.

Properties of $\Phi(z)$

Example (Basic Technique)

Let $Z \sim \mathcal{N}(0,1)$, compute the probability

$$P(-0.38 \leq Z \leq 1.21).$$

1. We first have
$$P(-0.38 \leq Z \leq 1.21) = \Phi(1.21) - \Phi(-0.38)$$
then we obtain $\Phi(0.38)$ and $\Phi(1.21)$, respectively.
2. From symmetry we know
$$\Phi(-0.38) = 1 - \Phi(0.38) \approx 1 - 0.6480 = 0.3520,$$
where $\Phi(0.38)$ is taken from the table.
3. The value $\Phi(1.21) \approx 0.8869$ is also easily found in the table.
4. Finally we have
$$P(-0.38 \leq Z \leq 1.21) \approx 0.8869 - 0.3520 = 0.5349.$$

Percentiles

Definition

For a constant α between 0 and 1, z_α is the unique solution to the equation

$$\Phi(z_\alpha) = 1 - \alpha.$$

This is the $100 \times (1-\alpha)^{\text{th}}$ **percentile** of the standard normal distribution.

We can find percentiles from the normal table.

Percentiles

Definition

For a constant α between 0 and 1, z_α is the unique solution to the equation

$$\Phi(z_\alpha) = 1 - \alpha.$$

This is the $100 \times (1-\alpha)^{\text{th}}$ **percentile** of the standard normal distribution.

We can find percentiles from the normal table.
E.g. For $z_{0.025} \approx 1.96$, the 97.5^{th} percentile of the standard normal distribution.

What if We Have a Non-Standard Normal Dist?

What if $X \sim N(\mu, \sigma)$ with $\mu \neq 0$ and/or $\sigma \neq 1$?

What if We Have a Non-Standard Normal Dist?

What if $X \sim N(\mu, \sigma)$ with $\mu \neq 0$ and/or $\sigma \neq 1$?

First let's review a fact which is true for *any* RV:

Fact

For fixed constants a and b we have

$$E(aX + b) = aE(X) + b$$

and

$$V(aX + b) = a^2 V(X)$$

What if We Have a Non-Standard Normal Dist?

What if $X \sim N(\mu, \sigma)$ with $\mu \neq 0$ and/or $\sigma \neq 1$?

First let's review a fact which is true for *any* RV:

Fact

For fixed constants a and b we have

$$E(aX + b) = aE(X) + b$$

and

$$V(aX + b) = a^2 V(X)$$

This fact lets us "translate" any given normal distribution $X \sim \mathcal{N}(\mu, \sigma)$ into a standard normal distribution!

Non-Standard Normal Distribution

Theorem

If $X \sim \mathcal{N}(\mu, \sigma)$, *and* $Y = aX + b$, *where* a *and* b *are constants, then*

$$Y \sim \mathcal{N}(a\mu + b, |a|\sigma)$$

Non-Standard Normal Distribution

Theorem

If $X \sim \mathcal{N}(\mu, \sigma)$, and $Y = aX + b$, where a and b are constants, then

$$Y \sim \mathcal{N}(a\mu + b, |a|\sigma)$$

Transformations of the type $Y = aX + b$ are very common

Example

C =temperature measured in $^\circ\mathrm{C}$
F =temperature measured in $^\circ\mathrm{F}$
then we have

$$F = \frac{9}{5}C + 32$$

The theorem simply says that if C is normal with $\mu_c = 5$ and $\sigma_c = 10$, i.e. $C \sim \mathcal{N}(5, 10)$, then $F \sim$

Non-Standard Normal Distribution

Theorem

If $X \sim \mathcal{N}(\mu, \sigma)$, and $Y = aX + b$, where a and b are constants, then

$$Y \sim \mathcal{N}(a\mu + b, |a|\sigma)$$

Transformations of the type $Y = aX + b$ are very common

Example

C =temperature measured in $^\circ\mathrm{C}$
F =temperature measured in $^\circ\mathrm{F}$
then we have

$$F = \frac{9}{5}C + 32$$

The theorem simply says that if C is normal with $\mu_c = 5$ and $\sigma_c = 10$, i.e. $C \sim \mathcal{N}(5, 10)$, then $F \sim \mathcal{N}(\frac{9}{5} \cdot 5 + 32, \frac{9}{5} \cdot 10)$

Z-Score

Fact ($\star\star\star$ Important Consequence! $\star\star\star$)

If $X \sim \mathcal{N}(\mu, \sigma)$ *then*

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

Theorem

If $X \sim \mathcal{N}(\mu, \sigma)$*, and* a *and* b *are two constants, then*

1. $P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$
2. $P(X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right)$
3. $P(X \geq a) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right)$

Z-Score

Definition

If $X \sim \mathcal{N}(\mu, \sigma)$ and c is a constant, then

$$\frac{c - \mu}{\sigma}$$

is called the **z-score of** c.

Non-Standard Normal Distribution

Example

Suppose $X \sim \mathcal{N}(4, 3)$ so $\mu = 4, \sigma = 3$.
We want to compute $P(3 \leq X \leq 7.5)$.

Non-Standard Normal Distribution

Example

Suppose $X \sim \mathcal{N}(4, 3)$ so $\mu = 4, \sigma = 3$.
We want to compute $P(3 \leq X \leq 7.5)$.

The z-sore of 3: $\frac{3-4}{3} = -0.333$
The z-sore of 7.5: $\frac{7.5-4}{3} = 1.167$

Non-Standard Normal Distribution

Example

Suppose $X \sim \mathcal{N}(4, 3)$ so $\mu = 4, \sigma = 3$.
We want to compute $P(3 \leq X \leq 7.5)$.

The z-sore of 3: $\frac{3-4}{3} = -0.333$
The z-sore of 7.5: $\frac{7.5-4}{3} = 1.167$
Compute the cdf:

$$\Phi(-0.333) = 1 - \Phi(0.333) \approx 1 - 0.6293$$
$$\Phi(1.167) \approx 0.8790$$

By the previous Theorem:

$$P(3 \leq X \leq 7.5) = \Phi\left(\frac{7.5-4}{3}\right) - \Phi\left(\frac{3-4}{3}\right) = \Phi(1.167) - \Phi(-0.333)$$
$$\approx 0.5083$$

Non-Standard Normal Distribution

Example (ponderal index)

The ponderal index is a measure of overall size, similar to the body mass index. The ponderal index of full-term newborns follows a normal distribution with a mean of 2.4 kg/m^3 (kilograms per cubic meter) and a standard deviation of 0.45 kg/m^3.

1. What percentage of newborns are less than 1.75 kg/m^3?
2. What is the ponderal index for which approximately 10% of all newborns are less than?

Non-Standard Normal Distribution

Part 1:

$X \sim N(2.4, 0.45)$. And we want to compute $P(X < 1.75)$.
We have

$$P(X < 1.75) = P\Big(\frac{X - 2.4}{0.45} < \frac{1.75 - 2.4}{0.45}\Big) = P\Big(Z < \frac{1.75 - 2.4}{0.45}\Big)$$
$$= \Phi\Big(\frac{1.75 - 2.4}{0.45}\Big)$$

where $Z \sim N(0, 1)$.
The z-score for 1.75 kg/m^3 is

$$\frac{1.75 - 2.4}{0.45} = -1.\overline{444}.$$

Then from the tables

$$\Phi(-1.44) = 1 - \Phi(1.44) \approx 1 - 0.9251 = 0.0749.$$

Non-Standard Normal Distribution

Part 2:

$X \sim N(2.4, 0.45)$. And we want to find the constant x such that $P(X < x) = 0.1$.
We have

$$0.1 = P(X < x) = P\Big(\frac{X - 2.4}{0.45} < \frac{x - 2.4}{0.45}\Big) = P\Big(Z < z\Big) = \Phi(z)$$

where $Z \sim N(0, 1)$ and $z = \frac{x-2.4}{0.45}$ (z-score of x).

Non-Standard Normal Distribution

We need to find the number z such that

$$\Phi(z) = 0.1.$$

(It is $z_{0.9}$)

	.0	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

From the table we find $\Phi(1.28) \approx 0.9$. Thus $\Phi(-1.28) \approx 1 - 0.9 = 0.1$.

Non-Standard Normal Distribution

So we have found the z-score $z = -1.28$, and now we need the corresponding observation. So we solve for x:

$$-1.28 = \frac{x - 2.4}{0.45}$$

to get $x = 1.824$ kg/m^3.

Approximating Binomial Distribution

Recall normal distribution approximates binomial distribution

- Let $X \sim$Binomial(n, p).

Approximating Binomial Distribution

Recall normal distribution approximates binomial distribution

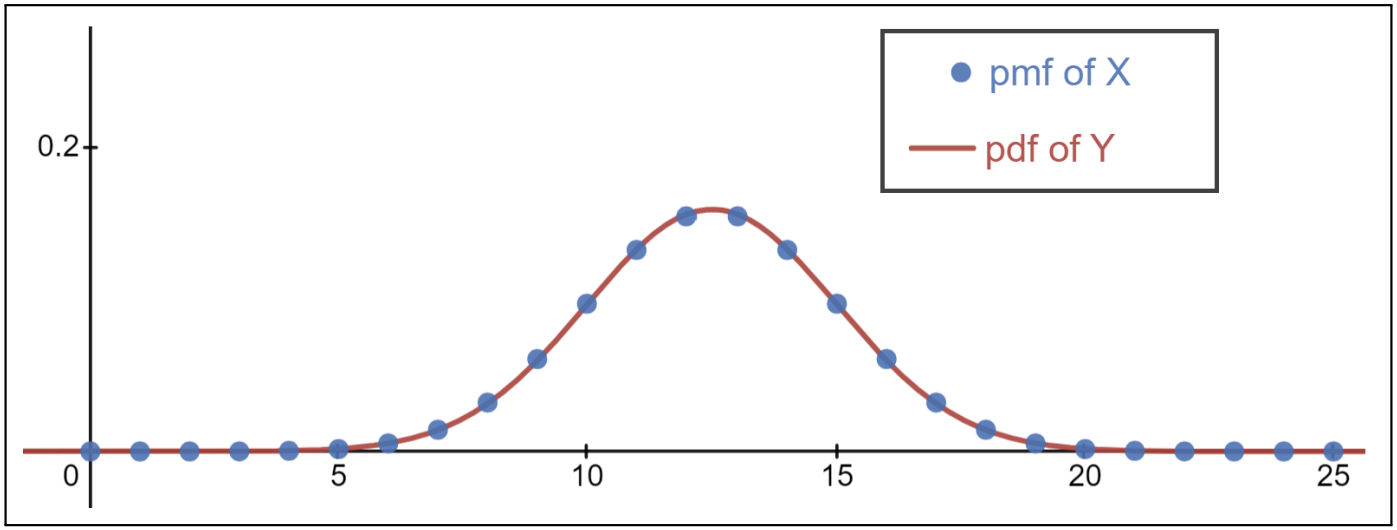
- Let $X \sim$Binomial(n, p).
- Then X has expectation np and standard deviation $\sqrt{npq}$.

Approximating Binomial Distribution

Recall normal distribution approximates binomial distribution

- Let $X \sim$Binomial(n, p).
- Then X has expectation np and standard deviation $\sqrt{npq}$.
- Let $Y \sim \mathcal{N}(np, \sqrt{npq})$, then the pdf of Y and the pmf of X fit very closely together:

Approximating Binomial Distribution



Example with $n = 25$ and $p = 0.5$. In this case, the pmf of X is

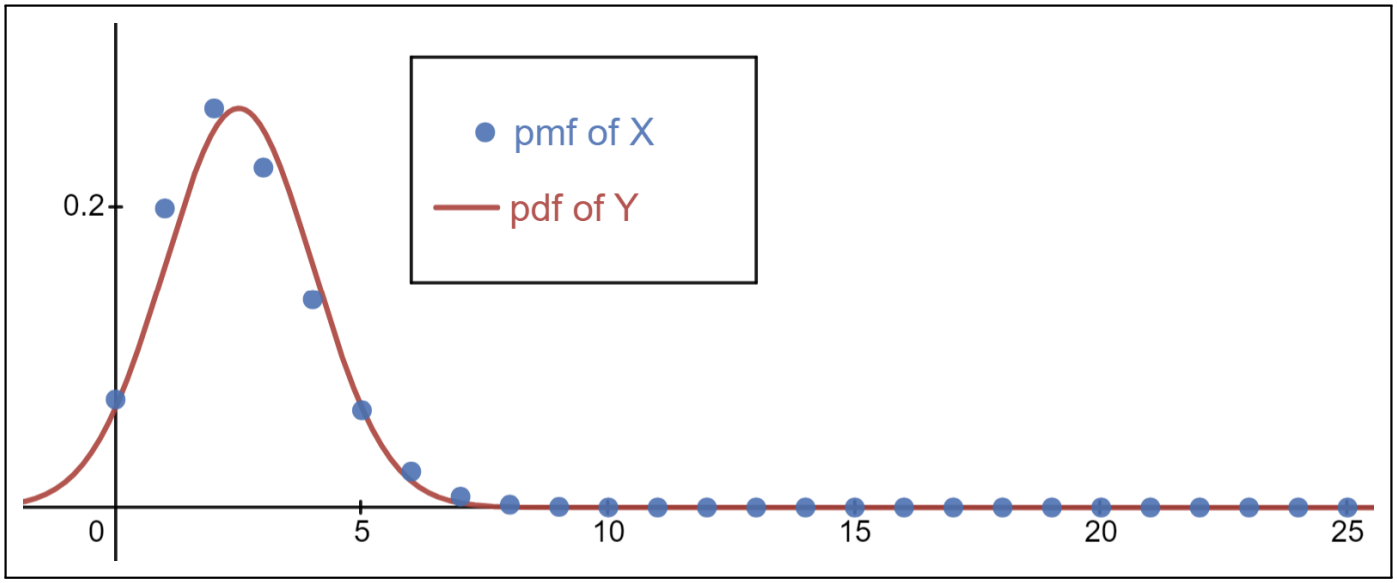
$$P(X = k) = \binom{25}{k} \frac{1}{2^n},$$

while the pdf of Y is

$$f(y) = \frac{1}{2.5\sqrt{2\pi}} e^{(y-12.5)^2/12.5}$$

Their graphs are so close together

When the Approximation Breaks



Here's an example with $n = 25$ and $p = 0.1$. It turns out the approximation is not as good when np or nq is too small. We make this more precise in the following fact.

Calculating Bin. Using Norm.

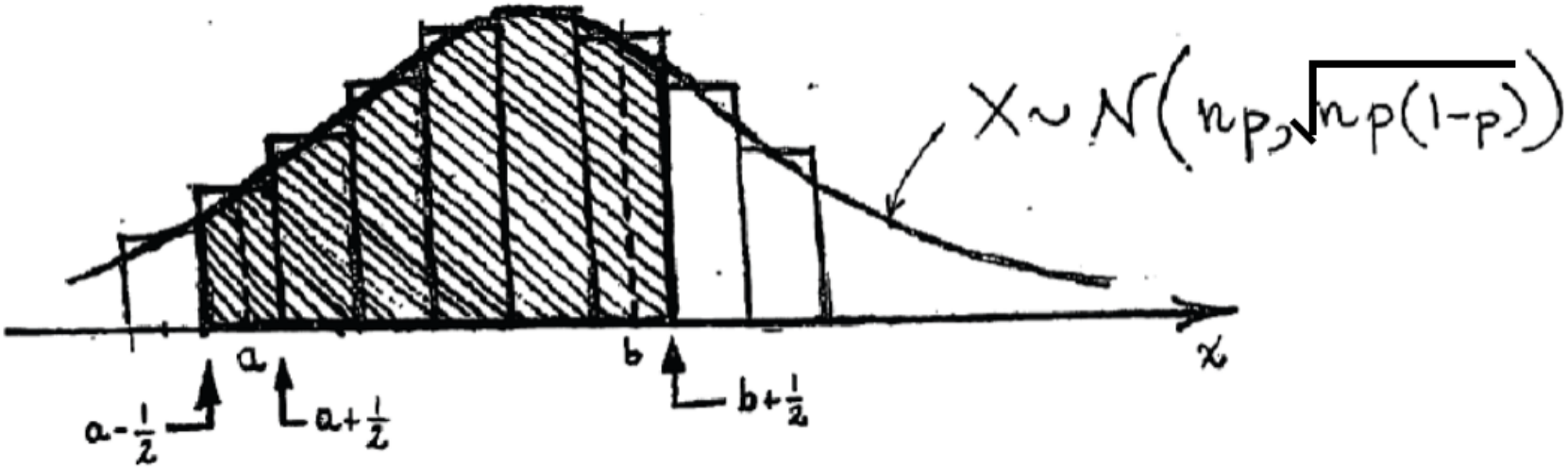
Fact (Continuity Correction)

If $X \sim Binomial(n, p)$ *with both* $np > 5$ *and* $nq > 5$. *Then*

1. $$P(a \leq X \leq b) \approx \Phi\left(\frac{b + 0.5 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - 0.5 - np}{\sqrt{npq}}\right)$$
2. $$P(X \leq a) \approx \Phi\left(\frac{a + 0.5 - np}{\sqrt{npq}}\right)$$
3. $$P(X \geq a) \approx 1 - \Phi\left(\frac{a - 0.5 - np}{\sqrt{npq}}\right)$$

What's with the ± 0.5?

Why Correct for Continuity?



We add the 0.5 to compensate for the area that we missed. This will give a better estimate of the probability.

Calculating Bin. Using Norm.

Example (Fuse Reliability)

The reliability of an electrical fuse is the probability that a fuse, chosen at random from production, will function under its designed conditions. A random sample of 1000 fuses was tested. Calculate the approximate probability of observing between 25 and 30 defectives, assuming that the fuse reliability is 0.98.

Calculating Bin. Using Norm.

Let X be the number of defective fuses in our sample of 1000. Then $X \sim$Binomial$(1000, 0.02)$.

Calculating Bin. Using Norm.

Let X be the number of defective fuses in our sample of 1000. Then $X \sim$Binomial$(1000, 0.02)$.

Since $1000 \times 0.02 = 20 > 5$ and $1000 \times 0.98 = 980 > 5$, we can use the normal distribution with

$$\mu = 1000 \cdot 0.02 = 20$$

and

$$\sigma = \sqrt{1000 \cdot 0.02 \cdot 0.98} \approx 4.4272$$

to estimate $P(25 \leq X \leq 30)$.

Calculating Bin. Using Norm.

The z-scores for 25 and 30 are

$$\frac{30.5 - 20}{4.4272} \approx 2.37 \quad \text{and} \quad \frac{24.5 - 20}{4.4272} \approx 1.02$$

We have

$$\begin{aligned} P(25 \leq X \leq 30) &\approx P(24.5 \leq Y \leq 30.5) \\ &= \Phi(2.37) - \Phi(1.02) \\ &\approx 0.9911 - 0.8461 \\ &= 0.145 \end{aligned}$$

where $Y \sim N(20, 4.4272)$

Calculating Bin. Using Norm.

The z-scores for 25 and 30 are

$$\frac{30.5 - 20}{4.4272} \approx 2.37 \quad \text{and} \quad \frac{24.5 - 20}{4.4272} \approx 1.02$$

We have

$$\begin{aligned} P(25 \leq X \leq 30) &\approx P(24.5 \leq Y \leq 30.5) \\ &= \Phi(2.37) - \Phi(1.02) \\ &\approx 0.9911 - 0.8461 \\ &= 0.145 \end{aligned}$$

where $Y \sim N(20, 4.4272)$

Note: This is pretty close to the true probability of about 0.142, we would be even more accurate if we rounded the z-scores to 4 decimal places instead of 2. We rounded to 2 places simply because we're using the tables to look up the probability.

Calculating Bin. Using Norm.

- Remember, a Binomial(n, p) RV (call it T) can be thought of as the sum of n independent Bernoulli(p) RVs:

$$T = X_1 + \cdots + X_n,$$

where $X_i \sim$Bernoulli(p) and $X_1, \ldots, X_n$ are independent. So X_i's have the expectation $E(X_i) = p$, and standard deviation $SD(X_i) = \sqrt{V(X_i)} = \sqrt{pq}$.

Calculating Bin. Using Norm.

- Remember, a Binomial(n, p) RV (call it T) can be thought of as the sum of n independent Bernoulli(p) RVs:

$$T = X_1 + \cdots + X_n,$$

where $X_i \sim$Bernoulli(p) and $X_1, \ldots, X_n$ are independent. So X_i's have the expectation $E(X_i) = p$, and standard deviation $SD(X_i) = \sqrt{V(X_i)} = \sqrt{pq}$.

- We just learned that T has *approximately* $\mathcal{N}(n \cdot p, \sqrt{n} \cdot \sqrt{pq})$.

Calculating Bin. Using Norm.

- Remember, a Binomial(n, p) RV (call it T) can be thought of as the sum of n independent Bernoulli(p) RVs:

$$T = X_1 + \cdots + X_n,$$

where $X_i \sim$Bernoulli(p) and $X_1, \ldots, X_n$ are independent. So X_i's have the expectation $E(X_i) = p$, and standard deviation $SD(X_i) = \sqrt{V(X_i)} = \sqrt{pq}$.

- We just learned that T has *approximately* $\mathcal{N}(n \cdot p, \sqrt{n} \cdot \sqrt{pq})$.
- Thus if T is the sum of n RVs $X_1, \ldots, X_n$, and we know $X_1, \ldots, X_n$ are independent and they have the same distribution–Bernoulli distribution. We then know T is approximately normal.
- Question: If X_i's are independent but follow other distributions (may be not Bernoulli distribution), can we still say $T = X_1 + \cdots + X_n$ is approximately normal?

⋆⋆⋆ Central Limit Theorem ⋆⋆⋆

Theorem (CLT)

If

- $X_1, \ldots, X_n$ *are independent random variables*
- *They have the same probability distribution with mean* $E(X_i) = \mu$ *and standard deviation* $SD(X_i) = \sigma$

We define $T = X_1 + X_2 + \cdots + X_n$

Then, for n large enough, we have

$$T \overset{approx.}{\sim} \mathcal{N}\left(n\mu, \sqrt{n}\sigma\right)$$

Central Limit Theorem

Remark

- *The theorem applies to any probability distribution for X_i's*
- *Notice T is approximately normal (not exactly normal)*
- *How large is "large enough"? Conservatively, $n \geq 30$.*
  If the distribution of X_i's is not too jagged, the approximation is good even for small n.
  For Binomial distribution, we always check np, nq, not just n.

Sampling Distribution for Sample Total and Sample Mean

- Think back to Week 1, we take a sample of size n from a population of size N. Typically, $n \ll N$

Sampling Distribution for Sample Total and Sample Mean

- Think back to Week 1, we take a sample of size n from a population of size N. Typically, $n \ll N$
- Because the sample is picked at random from the population, the individuals in the sample is random variables. The sample is indicated with

$$X_1, X_2, \ldots, X_n$$

Sampling Distribution for Sample Total and Sample Mean

- Think back to Week 1, we take a sample of size n from a population of size N. Typically, $n \ll N$
- Because the sample is picked at random from the population, the individuals in the sample is random variables. The sample is indicated with

$$X_1, X_2, \ldots, X_n$$

- Once we have the sample, we can estimate the population mean μ and population variance σ^2 by computing the sample average, sample variance.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$X_1, X_2, \ldots, X_n$ are RVs, so those statistics about the sample are also RVs.

Sampling Distribution for Sample Total and Sample Mean

Definition

The probability distribution of a statistics (i.e. of a function) of the data sample $X_1, X_2, \ldots, X_n$ is called a **sampling distribution**

- This naturally leads to the question, what is the sampling distribution for those statistics?
- Today we focus on the sampling distribution of sample mean and sample total

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$T = \sum_{i=1}^{n} X_i$$

Central Limit Theorem for Sample Total and Sample Mean

- The sampling distribution of X_i is the same with the population distribution

Theorem (Sampling distribution of X_i)

We perform random sampling from a population.
Then each observation X_i of the sample $X_1, X_2, \ldots, X_n$ has the same probability distribution with the population distribution.
Notice $E(X_i) = \mu, SD(X_i) = \sigma$

$\Rightarrow E(T) = n\mu$ and $E(\overline{X}) = \mu$

Central Limit Theorem for Sample Total and Sample Mean

- The sampling distribution of X_i is the same with the population distribution

Theorem (Sampling distribution of X_i)

We perform random sampling from a population.
Then each observation X_i of the sample $X_1, X_2, \ldots, X_n$ has the same probability distribution with the population distribution.
Notice $E(X_i) = \mu, SD(X_i) = \sigma$

$\Rightarrow E(T) = n\mu$ and $E(\overline{X}) = \mu$

- Recall that we get the sample without replacement, but when $n \ll N$ (we always suppose $n \ll N$), it's "as if" we were sampling with replacement. Thus we can assume "almost" independence of $X_1, X_2, \ldots, X_n$

Central Limit Theorem for Sample Total and Sample Mean

- The sampling distribution of X_i is the same with the population distribution

Theorem (Sampling distribution of X_i)

We perform random sampling from a population.
Then each observation X_i of the sample $X_1, X_2, \ldots, X_n$ has the same probability distribution with the population distribution.
Notice $E(X_i) = \mu, SD(X_i) = \sigma$

$\Rightarrow E(T) = n\mu$ and $E(\overline{X}) = \mu$

- Recall that we get the sample without replacement, but when $n \ll N$ (we always suppose $n \ll N$), it's "as if" we were sampling with replacement. Thus we can assume "almost" independence of $X_1, X_2, \ldots, X_n$
  $\Rightarrow$ It follows that $SD(T) = \sqrt{n}\sigma$ and $SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$

Central Limit Theorem for Sample Total and Sample Mean

- The sampling distribution of X_i is the same with the population distribution

Theorem (Sampling distribution of X_i)

We perform random sampling from a population.
Then each observation X_i of the sample $X_1, X_2, \ldots, X_n$ has the same probability distribution with the population distribution.
Notice $E(X_i) = \mu, SD(X_i) = \sigma$

$\Rightarrow E(T) = n\mu$ and $E(\overline{X}) = \mu$

- Recall that we get the sample without replacement, but when $n \ll N$ (we always suppose $n \ll N$), it's "as if" we were sampling with replacement. Thus we can assume "almost" independence of $X_1, X_2, \ldots, X_n$
  $\Rightarrow$ It follows that $SD(T) = \sqrt{n}\sigma$ and $SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$
- What is the shape of the distributions of T and $\overline{X}$?

⋆⋆⋆ Central Limit Theorem for Sample Total and Sample Mean⋆⋆⋆

Since $X_1, X_2, \ldots, X_n$ are independent with identical distribution, we can apply CLT

Theorem (CLT for sample total and sample mean)

Let $\{X_1, \ldots, X_n\}$ be a random sample from a population with mean μ and std. dev. σ (with $n \ll N$). Let T and $\overline{X}$ be the sample total and sample average, respectively.
If n is sufficiently large (typically, $n \geq 30$), then we have

$$T \overset{\text{approx.}}{\sim} \mathcal{N}\left(n\mu, \sqrt{n}\sigma\right)$$

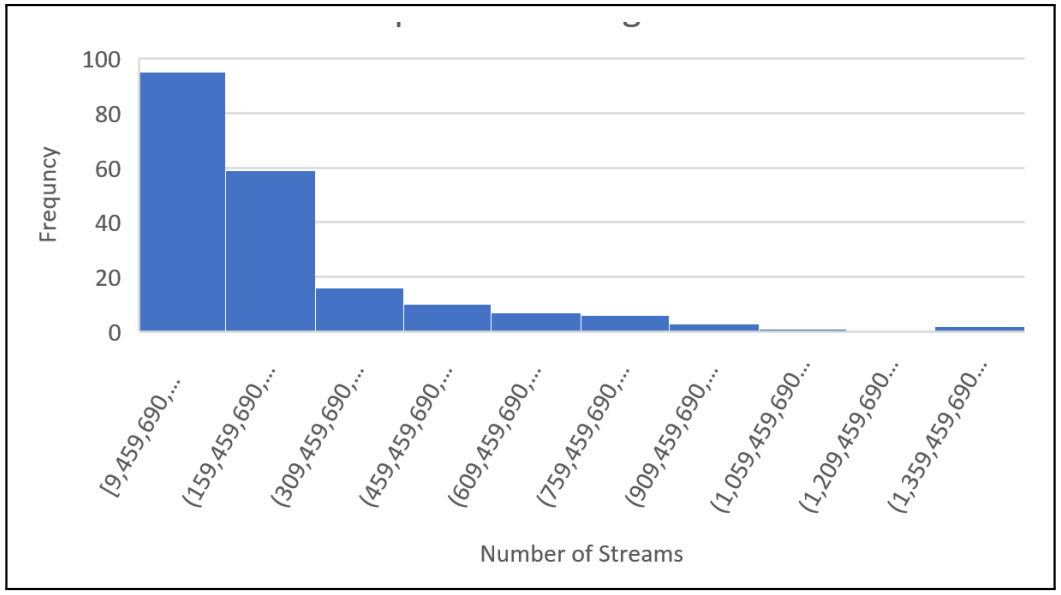
and

$$\overline{X} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

regardless of the population distribution.

Central Limit Theorem for Sample Total and Sample Mean

In the `T_Swift` Excel file in Canvas, you can find the estimated total Spotify streams for each of the top 200 most steamed Taylor Swift songs as of Aug. 2023. As you can see, the number of streams follows the following distribution:



Central Limit Theorem for Sample Total and Sample Mean

If we take a random sample of 40 songs from our list and find the average number of streams of those songs, we might get a number like 273,692,641.

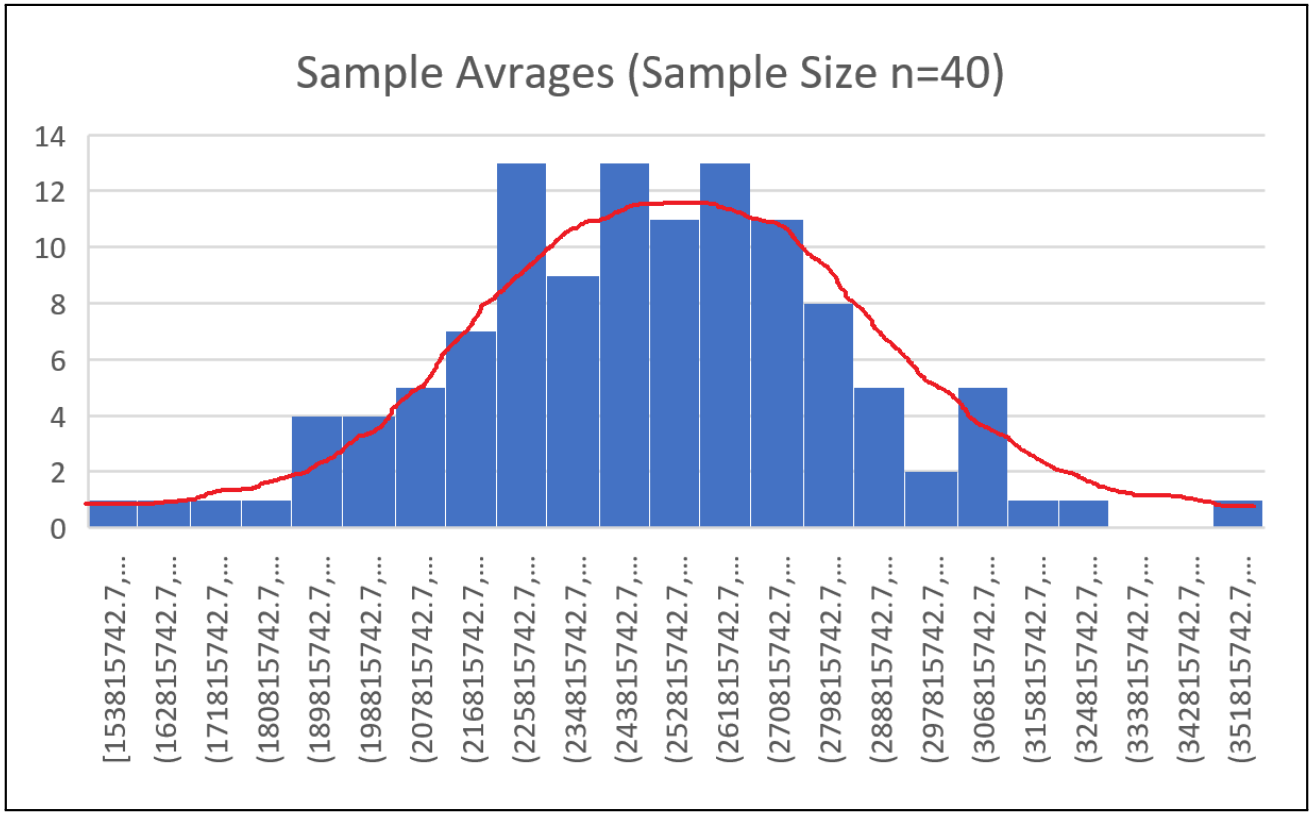
If we repeat this process of taking a random sample of 40 songs and finding the sample average many many times[2], then we'd have a list of these averages:

Sample #	**Sample Avg.**
1	273,692,641
2	267,909,152
⋮	⋮
117	220,153,729

[2] In the `T_Swift` Excel file we took 117 random samples, this number was picked arbitrarily.

Central Limit Theorem for Sample Total and Sample Mean

The distribution of these *averages* of random samples start to look pretty normal



Sum of Independent Normal RVs

What if $X_1, X_2, \ldots, X_n$ are normal to start with? This further assumption allows for a stronger result

Theorem (Normal Sum)

If

- $X_1, X_2, \ldots, X_n$ *be independent RVs*
- $X_i \sim \mathcal{N}(\mu_i, \sigma_i)$ *for each* $i = 1, 2, \ldots, n$

We define $T = X_1 + X_2 + \cdots + X_n$
Then, for any n, we have

$$T \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right)$$

Notice here T is exactly normal

Sum of Independent Normal RVs

Remark

In this case, as long as $X_1, X_2, \ldots, X_n$ are normal

- *$X_1, X_2, \ldots, X_n$ don't have to have the same mean and variance*
- *n doesn't have to be large. We can have $n = 2$*

⋆⋆⋆ Sampling Distribution of Sample Total and Sample Mean for Normal Population⋆⋆⋆

Theorem

Let $\{X_1, \ldots, X_n\}$ be a random sample from a population with normal distribution $\mathcal{N}(\mu, \sigma)$. Let T and $\overline{X}$ be the sample total and sample average, respectively. Then, for any n we have

$$T \sim \mathcal{N}\left(n\mu, \sqrt{n}\sigma\right)$$

and

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Notice here T and $\overline{X}$ are exactly normal

Standard Error

Definition

The standard deviation of a statistic used as an estimator of a population parameter is called the **standard error of the estimator** (which we'll write as **SE** for short) because it refers to the precision of the estimator

E.g. $\overline{X}$ is the corresponding statistic for the population parameter μ. Because $E(\overline{X}) = \mu$, μ is the center of the distribution of $\overline{X}$. Then $SE(\overline{X}) = SD(\overline{X}) = \sigma/\sqrt{n}$ refers the precision of $\overline{X}$ as the estimator of μ

Standard Error

Definition

The standard deviation of a statistic used as an estimator of a population parameter is called the **standard error of the estimator** (which we'll write as **SE** for short) because it refers to the precision of the estimator

E.g. $\overline{X}$ is the corresponding statistic for the population parameter μ. Because $E(\overline{X}) = \mu$, μ is the center of the distribution of $\overline{X}$. Then $SE(\overline{X}) = SD(\overline{X}) = \sigma/\sqrt{n}$ refers the precision of $\overline{X}$ as the estimator of μ

Note: Usually, the deviation of the particular sample is *not* exactly the SE! SE describes the variability of the sample distribution of the statistic.

Standard Error of Sample Mean

Remark

- *The standard error of the estimator $E(\overline{X})$ is its standard deviation*

$$SE(\overline{X}) = SD(\overline{X}) = \frac{\sigma}{\sqrt{n}},$$

which decreases as the sample size n increases. If $SE(\overline{X})$ is small, which can be obtained by increasing n, $\overline{X}$ is close to μ with large probability.

E.g. in order to halve the standard error $SE(\overline{X})$, we need to quadruple the sample size.

Sampling Distribution of Sample Total and Sample Mean

Example (Alzheimer's Disease)

The duration of Alzheimer's disease from the onset of symptoms until death ranges from 3 to 20 year; the average is 8 years with a standard deviation of 4 years. The administrator of a large medical center randomly selects the medical records of 30 deceased Alzheimer's patients from the medical center's database, and records the average duration. Find the probability that:

- The average duration is less than 7 years.
- The average duration lies within 1 year of the mean.
- The sample size needed to ensure that the probability in 2 is at least 0.95.

Sampling Distribution of Sample Total and Sample Mean

Part 1:

In this case, we're given $\mu = 8$ and $\sigma = 4$ and $n = 30 \geq 30$. By CLT, the sample average $\overline{X}$ will have an approximately $\mathcal{N}(\mu, \sigma/\sqrt{n})$, and

$$SE(\overline{X}) = SD(\overline{X}) = \frac{4}{\sqrt{30}} \approx 0.73$$

Sampling Distribution of Sample Total and Sample Mean

Part 1:

In this case, we're given $\mu = 8$ and $\sigma = 4$ and $n = 30 \geq 30$. By CLT, the sample average $\overline{X}$ will have an approximately $\mathcal{N}(\mu, \sigma/\sqrt{n})$, and

$$SE(\overline{X}) = SD(\overline{X}) = \frac{4}{\sqrt{30}} \approx 0.73$$

Now, the z-score of 7 years is

$$\frac{7-8}{0.73} \approx -1.37$$

and from the tables we find

$$P(\overline{X} \leq 7) \approx \Phi(-1.37) \approx 0.0853$$

Sampling Distribution of Sample Total and Sample Mean

Part 2:

The next question asks the prob. that $\overline{X}$ is within $\mu \pm 1$, i.e. Between 7 and 9 years.

Sampling Distribution of Sample Total and Sample Mean

Part 2:

The next question asks the prob. that $\overline{X}$ is within $\mu \pm 1$, i.e. Between 7 and 9 years.

We already found $P(\overline{X} \leq 7)$, so let's find the z-score of 9 years:

$$\frac{9-8}{0.73} \approx 1.37$$

Sampling Distribution of Sample Total and Sample Mean

Part 2:

The next question asks the prob. that $\overline{X}$ is within $\mu \pm 1$, i.e. Between 7 and 9 years.

We already found $P(\overline{X} \leq 7)$, so let's find the z-score of 9 years:

$$\frac{9-8}{0.73} \approx 1.37$$

So we have

$$\begin{aligned} P(7 \leq \overline{X} \leq 9) &\approx \Phi(1.37) - \Phi(-1.37) \\ &\approx 0.9147 - 0.0853 \\ &= 0.8294. \end{aligned}$$

Sampling Distribution of Sample Total and Sample Mean

Part 3:

Here we take advantage of the fact that the larger the sample, the smaller the SE. We want to find n so that

$$P(7 \leq \overline{X} \leq 9) \geq 0.95$$

How do we do this?

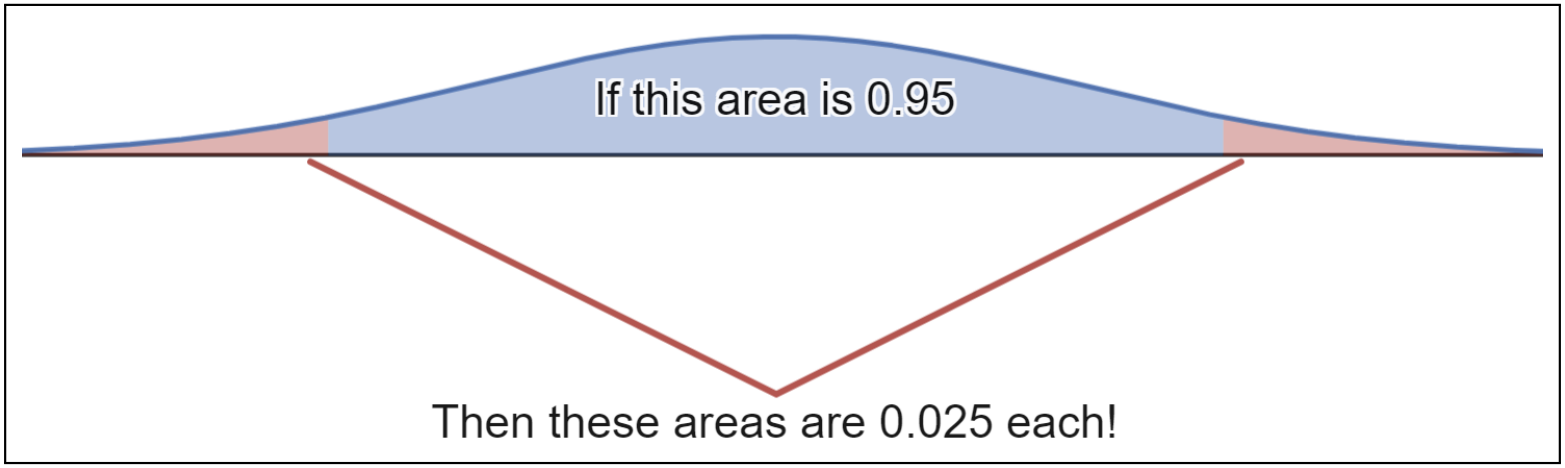
Sampling Distribution of Sample Total and Sample Mean

Part 3:

Here we take advantage of the fact that the larger the sample, the smaller the SE. We want to find n so that

$$P(7 \leq \overline{X} \leq 9) \geq 0.95$$

How do we do this? We'll use symmetry!



Standard Error

From the tables we find $z_{0.025} = 1.96$. So we solve

$$\frac{9-8}{4/\sqrt{n}} \geq 1.96$$

to get

$$n \geq (1.96 \times 4)^2 \approx 61.4656$$

which we would round up to 62.

Takeaways

Sampling distribution of Sample Total and Sample Mean

Given a population of size N with population mean μ and population standard deviation σ. We gather a random sample of size n. Suppose $n \ll N$.

- If the population distribution is normal, we have

$$T \sim \mathcal{N}\left(n\mu, \sqrt{n}\sigma\right), \overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Or, if the sample size $n \geq 30$, regardless of the population distribution by CLT we have

$$T \overset{\text{approx.}}{\sim} \mathcal{N}\left(n\mu, \sqrt{n}\sigma\right), \overline{X} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Then we can compute the probabilities or percentiles for the RV T and $\bar{X}$ with normal distribution.

Sampling distribution of Sample Mean

Remark

- $\overline{X}$ *is the estimator of* μ.
- *Since* $E(\overline{X}) = \mu$, *we say that "*$\overline{X}$ *is an unbiased estimator of* μ*". Because* $\overline{X}$ *is, on average, equal to* μ.

Definition

An estimator of a parameter is said to be **unbiased** if the mean of its distribution is equal to the true value of the parameter. Otherwise, the estimator is said to be **biased.**