CSCI 3022 Intro to Data Science
Normals

X: units of data/variable of interest

The four big functions (`scipy.stats` as `stats`):

1. `stats.distribution.rvs(params, size=...)` generates random numbers from the named distribution.
2. `stats.distribution.pdf(x,params)` returns the pdf of the distribution at the x value input as the function's first argument. For a discrete random variable, this is $P(X = x)$.
3. `stats.distribution.cdf(x,params)` returns the cdf of the distribution at the x value input as the function's first argument. This is $P(X \leq x)$. (area-to-the-left if conti, or sum-to-the-left)
4. `stats.distribution.ppf(p,params)` returns the *inverse* of cdf of the probability p value input as the function's first argument. This is the value of x that satisfies $p = P(X \leq x)$. "percentile point function".

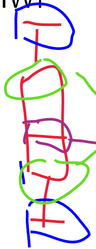
`distribution` arguments we've seen include: `poisson`, `binomial`, `uniform`, `exponential`, and more to come!.

Announcements and Reminders

Proportion = $\frac{\text{count that satisfy property}}{\text{all observations.}}$

- Exam due Friday.
- Practicum posted: it's 2 longer homework problems; due Mar 19. Then we get a week with no HW!

→ First ~20 min of lecture Wednesday to exam-adjacent Q & A.
(ask past exam's questions).



max OR "last data point with 1.5 × IQR of the box"

IQ values: quantiles for "75%" & "25%"

50% / middle.

min

The Normal Distribution

The normal distribution (sometimes called the Gaussian distribution) is probably the most important distribution in all of ~~probability and statistics~~.

everything

Many populations have distributions that can be fit very closely by an appropriate normal (or Gaussian, bell) curve.

Examples: height, weight, and other physical characteristics, scores on various tests, etc.

The Normal Distribution

Definition: *Normal Distribution:* looks like e^{-x^2}

A continuous r.v. X is said to have a *normal distribution* with parameters μ and $\sigma^2 > 0$, if the pdf of X is:

mu ← mean

sigma-squared ← variance

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

Notation: We write $N(\mu, \sigma^2)$

X: input

μ: mean, shifts curve to the right by "μ"

σ^2: integrate to 1 term & also spread

The Normal Distribution

Definition: *Normal Distribution:*

A continuous r.v. X is said to have a *normal distribution* with parameters $\underline{\mu}$ and $\underline{\sigma^2} > 0$, if the pdf of X is:

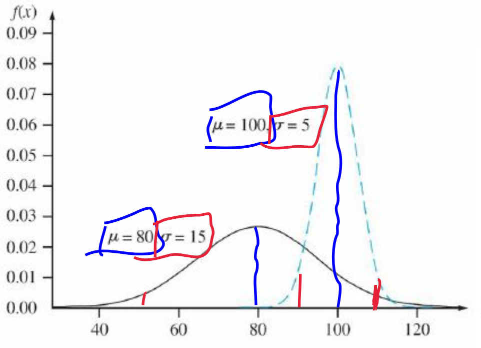
$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

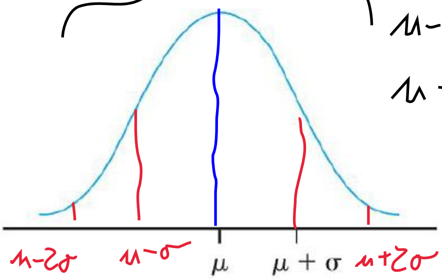
Notation: We write $\underline{X \sim N(\mu, \sigma^2)}$

The Normal Distribution

The figure below presents graphs of f for different parameter pairs:

"most" of the area is from $\mu - 2\sigma$ to $\mu + 2\sigma$





You can play with normals in any statistical software. See for example
`https://academo.org/demos/gaussian-distribution/`

The Standard Normal Distribution

Recall: if we have $X = [x_1, x_2, \cdots x_n]$

replace with: $y[i] = \frac{x[i] - \text{mean}(x)}{\text{sd}(x)}$

Definition: *Standard Normal Distribution:*
The normal distribution with parameter values $\mu = 0$ and $\sigma^2 = 1$ is called the *standard normal distribution*.

1) mean $(y) = 0$
2) std $(y) = 1$
3) shape y doesn't shape

A r.v. with this distribution is called a standard normal random variable and is denoted by Z.
Its pdf is:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-0)^2}$$

The Standard Normal Distribution

Definition: *Standard Normal Distribution:*
The normal distribution with parameter values $\underline{\mu = 0}$ and $\underline{\sigma^2 = 1}$ is called the *standard normal distribution*.

A r.v. with this distribution is called a standard normal random variable and is denoted by Z. Its pdf is:

$$f(z) =$$

The Standard Normal Distribution

Definition: *Standard Normal Distribution:*
The normal distribution with parameter values $\underline{\mu = 0}$ and $\underline{\sigma^2 = 1}$ is called the *standard normal distribution*.

A r.v. with this distribution is called a standard normal random variable and is denoted by $\boxed{Z}$.
Its pdf is:

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

Z is $N(0,1)$

The normal cdf

Let's find the cdf of the standard normal distribution!
All we have to to is integrate:

what is the anti-derivative?

$$\int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = F(Z) - F(-\infty)?$$

1) IBP
2) u-sub
3) ~~tears~~
4) ~~symbolab~~

5) idk?
6) it's a trick.

The normal cdf

Let's find the cdf of the standard normal distribution!
All we have to to is integrate:

$$\int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

Should we try a substitution? IBP?... this may not go integreat for us.

The normal cdf

Let's find the cdf of the standard normal distribution!
All we have to to is integrate:

$$\int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

The CDF of the normal distribution has no closed form. But it's really important! So we give it it's own name.

stats.normal.cdf(Z)

(mean=0, std.dev=1).

The normal cdf

For a random variable $Z \sim N(0,1)$, the cdf of Z is given by

$$F(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = \boxed{\Phi(z)}$$

The normal cdf

For a random variable $Z \sim N(0,1)$, the cdf of Z is given by

$$F(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = \boxed{\Phi(z)}$$

Old school statisticians used to carry around giant tables with values of $\Phi(z)$ in them. Actually, many current statisticians do that too, but that's a little silly. We have computers!

<u>$d\Gamma$</u> Stats. normal. cdf (z).

The Standard Normal

Note:

1. The standard normal distribution rarely occurs naturally.
2. Instead, it is a reference distribution from which information about other normal distributions can be obtained via a simple formula.
3. These probabilities can then be found "normal tables".
4. This can also be computed with a single command... (`scipy.stats.norm.cdf`, for example)

The Standard Normal

Note:

1. The standard normal distribution rarely occurs naturally.
2. Instead, it is a reference distribution from which information about other normal distributions can be obtained via a simple formula.
3. These probabilities can then be found "normal tables".
4. This can also be computed with a single command... (`scipy.stats.norm.cdf`, for example)

Recall: one example from HW1: if we take a data set, and *subtract the mean* from each of the data values, then we *divide by the standard deviation*, we ended up with a new data set that was mean of 0 and variance/standard deviation of 1. The new data set had the same **shape** as the original, but now it was "centered" at 0 and "scaled" to be of a known (average) spread.

"standardizing"

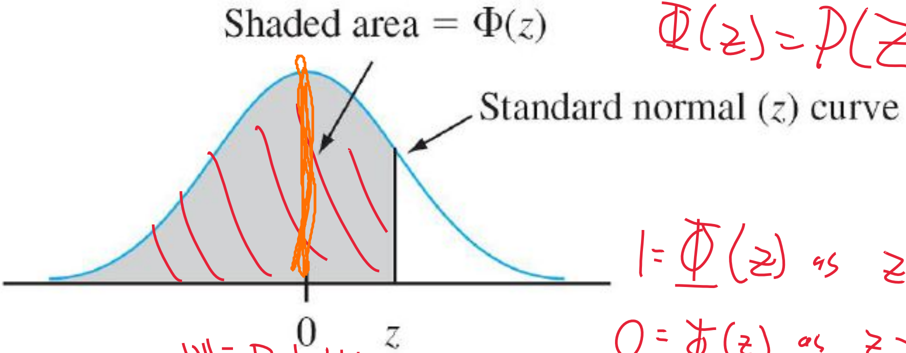
The Standard Distribution

The figure below illustrates the probabilities found in a normal table (such a table can easily be found online):

Φ is a cdf:

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

input



|||= Probability.

$1 = \Phi(z)$ as $z \to \infty$

$0 = \Phi(z)$ as $z \to -\infty$

$.5 = \Phi(0)$

The Standard Distribution

$P(Z \leq 1.25) = \Phi(1.25)$, a probability that is tabulated in a normal table. What is this probability?

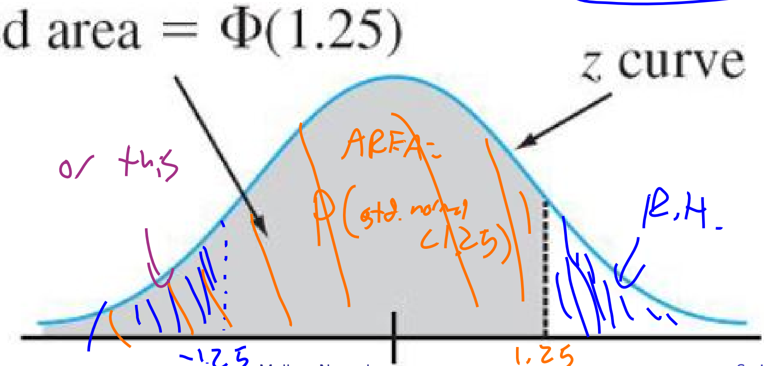
Orange: $\Phi(1.25)$

The figure below illustrates this probability:

is either

$1 - \Phi(1.25)$

$\Phi(-1.25)$



cdf

The Standard Distribution

$\Phi(1.25) = P(Z \leq 1.25)$.

Some quick examples:

1. $P(Z \geq 1.25)$ $= 1 - P(Z < 1.25) = 1 - \Phi(1.25)$

2. Why does $\boxed{P(Z < -1.25) = P(Z > 1.25)}$? What is $\Phi(-1.25)$?

   $1 - \Phi(1.25)$

3. How do we calculate $P(-.38 \leq Z \leq 1.25)$?

$$\int_{-0.38}^{1.25} f(z)\,dz = F(1.25) - F(-0.38)$$
$$= \Phi(1.25) - \Phi(-0.38)$$

The Standard Distribution

Some quick examples:

1. $P(Z \geq 1.25)$
   It's `1-scipy.stats.norm.cdf(1.25)`. Or as a picture:

2. Why does $P(Z < -1.25) = P(Z > 1.25)$? What is $\Phi(-1.25)$?
   Symmetry! Same as above.

3. How do we calculate $P(-.38 \leq Z \leq 1.25)$?
   As an integral, this is $\displaystyle\int_{-.38}^{1.25} f(z)\,dz$. We could split this into 2:
   $\displaystyle\int_{-\infty}^{1.25} f(z)\,dz + \int_{-.38}^{-\infty} f(z)\,dz =$

$$\Phi(1.25) - \Phi(-.38)$$

Standard Quantiles

The 99th *percentile* of the standard normal distribution is that value of z such that the area under the z curve to the left of the value is 0.99.

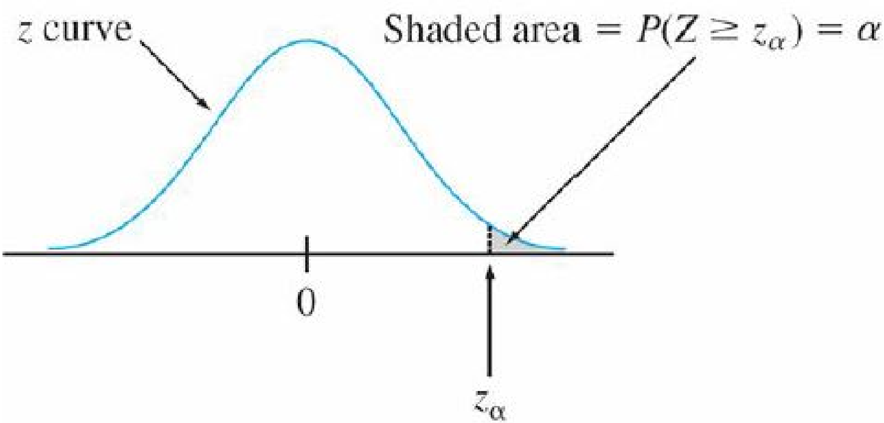
Tables and cdf functions give, for fixed z, the area under the standard normal curve to the left of z; now we have the area and want the value of z.

This is the “inverse” problem to $P(Z \leq z) = ?$

How can the table be used for this?

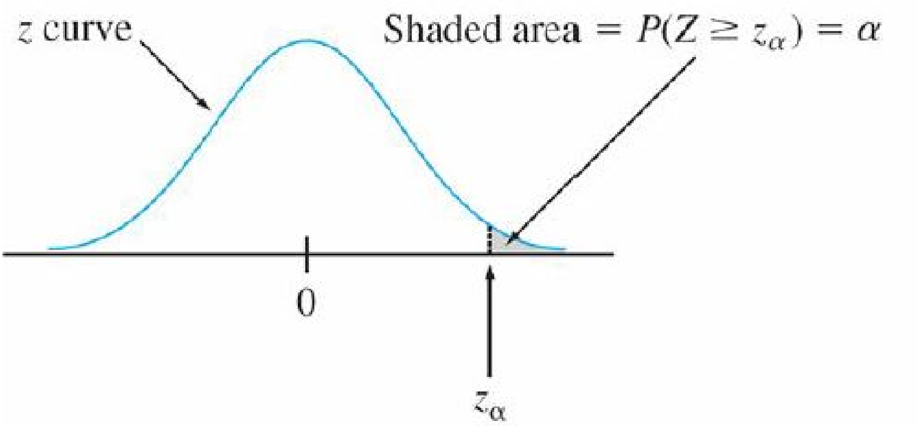
Standard Quantiles

In statistical inference, we need the z values that give certain tail areas under the standard normal curve. There, this notation will be standard: __ will denote the z value for which _ of the area under the z curve lies to the right of __.



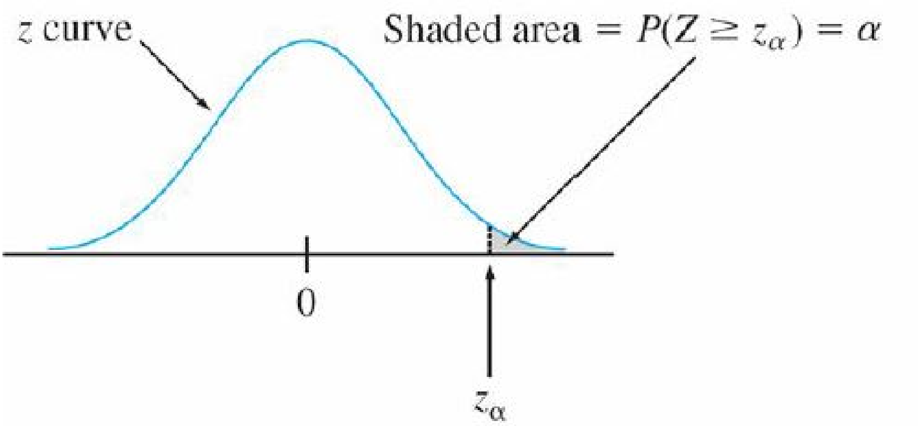
Standard Quantiles

In statistical inference, we need the z values that give certain tail areas under the standard normal curve. There, this notation will be standard: $\underline{z_\alpha}$ will denote the z value for which $\underline{\alpha}$ of the area under the z curve lies to the right of $\underline{z_\alpha}$.



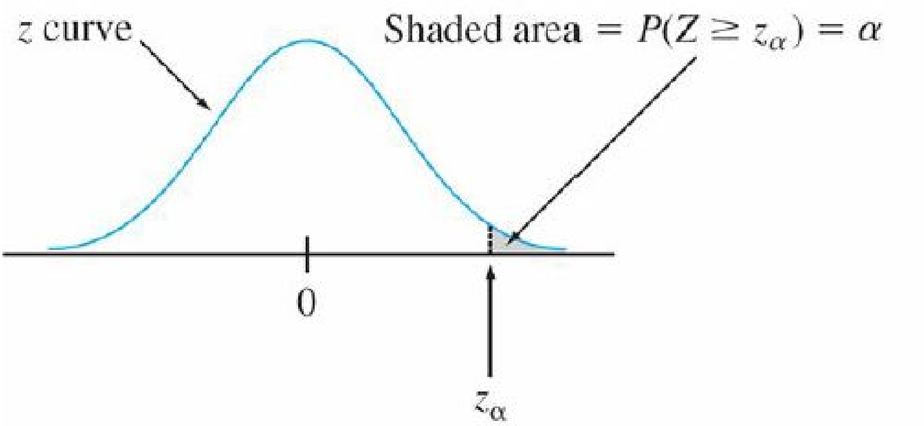
Standard Quantiles

In statistical inference, we need the z values that give certain tail areas under the standard normal curve. There, this notation will be standard: __ will denote the z value for which _ of the area under the z curve lies to the right of __.



Standard Quantiles

In statistical inference, we need the z values that give certain tail areas under the standard normal curve. There, this notation will be standard: $\underline{z_\alpha}$ will denote the z value for which $\underline{\alpha}$ of the area under the z curve lies to the right of $\underline{z_\alpha}$.



Non-Standard Normals

When $X \sim N(\mu, \sigma^2)$, probabilities involving X are computed by "standardizing." The standardized variable is:

X subtract mean, divide by std. dev.

Proposition: If X has a normal distribution with mean μ and standard deviation σ, then

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

is distributed standard normal.

Non-Standard Normals

When $X \sim N(\mu, \sigma^2)$, probabilities involving X are computed by "standardizing." The standardized variable is:

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

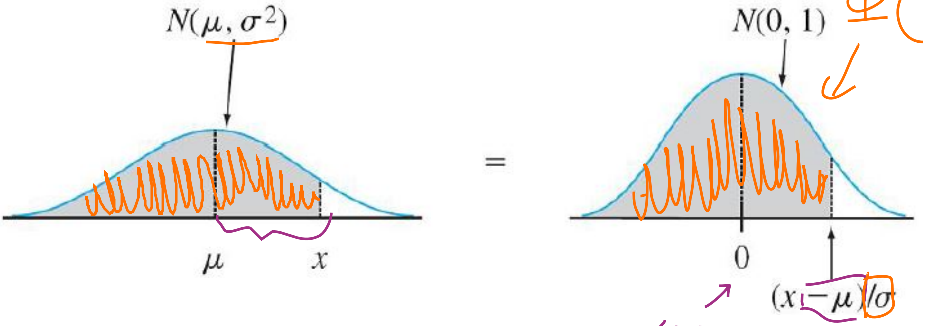
Proposition: If X has a normal distribution with mean μ and standard deviation σ, then

is distributed standard normal.

Non-Standard Normals

Why do we standardize normal random variables?

$\Phi\left(\frac{x-\mu}{\sigma}\right)$



center is zero

Equality of nonstandard and standard normal curve areas

Using Normals

Example:
The time that it takes a driver to react to the brake lights on a decelerating vehicle is critical in helping to avoid rear-end collisions.

Research suggests that reaction time for an in-traffic response to a brake signal from standard brake lights can be modeled with a normal distribution having mean value 1.25 sec and standard deviation of 0.46 sec.

reaction X is $N(\mu = 1.25,\ \sigma^2 = (.46)^2)$

What is the probability that reaction time is between 1.00 sec and 1.75 sec?

good at: $\boxed{\Phi}$; $P(a \le Z \le b)$. $\quad P(1 \le X \le 1.75)$

if X has $\mu = 1.25$, $\sigma = 0.46$, then $Z = \dfrac{X - 1.25}{.46} \Rightarrow \mu=0,\ \sigma=1$

Solution:

Example: For a normal distribution having mean value 1.25 sec and standard deviation of 0.46 sec.

x_1 x_2

What is the probability that reaction time is between 1.00 sec and 1.75 sec?

.5 away from avg. or $\frac{.5}{.46}$ std. devs. larger than average

1.00 is .25 less than 1.25.

this $^{25}/_{46}$ "standard deviation"

$$\frac{x_1 - 1.25}{.46} = z_1 \qquad \frac{x_2 - 1.25}{.46} = z_2$$

Solution:

Example: For a normal distribution having mean value 1.25 sec and standard deviation of 0.46 sec.

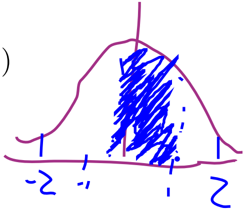
$$X \sim N(1.25, .46)$$

What is the probability that reaction time is between 1.00 sec and 1.75 sec?
We want $P(1 < X < 1.75)$... but we can't compute these probabilities unless the r.v. in the middle of the inequality is *standard* normal. So we normalize!

Solution:

Example: For a normal distribution having mean value 1.25 sec and standard deviation of 0.46 sec.

What is the probability that reaction time is between 1.00 sec and 1.75 sec?
We want $P(1 < X < 1.75)$... but we can't compute these probabilities unless the r.v. in the middle of the inequality is *standard* normal. So we normalize!



$$P(1 < X < 1.75) = P(1 - 1.25 < X - 1.25 < 1.75 - 1.25)$$

$$= P(\frac{-.25}{.46} < \frac{X - 1.25}{.46} < \frac{.5}{.46}) = P(\frac{-.25}{.46} < Z < \frac{.5}{.46})$$

$$= \Phi(\frac{-.25}{.46}) - \Phi(\frac{.5}{.46})$$

$\approx -.5$ ≈ 1

iid

Definition: *Random Sample:*
The r.v.'s $X_1, X_2, \ldots, X_n$ are said to form a (simple) random sample of size n if:

1. i : independent

2. id : identically distributed.

We say that these X_i's are:

iid

Definition: *Random Sample:*
The r.v.'s $X_1, X_2, \ldots, X_n$ are said to form a (simple) random sample of size n if:

1. $X_1, X_2, \ldots X_n$ are independent.

2.No value in the population has a higher chance of being included than any other.

We say that these X_i's are: *independent* and *identically distributed.*
and we write:

$$X_1, X_2, \ldots X_n \stackrel{iid}{\sim} f(x;\theta)$$

Estimators and Their Distributions

$\bar{X}$; S

We use $\boxed{\text{estimators}}$ to summarize our i.i.d. sample.

we used λ, $\underbrace{n}$ exp pois $\underbrace{n \; \& \; p,}_{\text{binom}}$ $\underbrace{\mu \; \& \; \sigma^2}_{\text{normal}}$

Examples?

distribution

$\bar{X}$ & μ are means

Estimators and Their Distributions

We use estimators to summarize our i.i.d. sample.

Examples?

1. *Sample* Mean might estimate a population mean.
2. *Sample* Variances estimate population variance.
3. *Sample* Quantiles
4. $\hat{p}$ for p
5. etc., etc.

Estimators and Their Distributions

We use estimators to summarize our i.i.d. sample.

Examples?

1. *Sample* Mean might estimate a population mean.
2. *Sample* Variances estimate population variance.
3. *Sample* Quantiles
4. $\hat{p}$ for p
5. etc., etc.

Why use one estimator over another?

Estimators and Their Distributions

We use estimators to summarize our i.i.d. sample. Any estimator, including the sample mean ___ is a random variable (since it is based on a random sample).

This means that ___ has a distribution of it's own, which is referred to as sampling distribution of the sample mean. This sampling distribution depends on:

Definition: The standard deviation of this distribution is called the *standard error* of the estimator.

Estimators and Their Distributions

We use estimators to summarize our i.i.d. sample. Any estimator, including the sample mean $\underline{\bar{X}}$ is a random variable (since it is based on a random sample).

This means that $\underline{\bar{X}}$ has a distribution of it's own, which is referred to as sampling distribution of the sample mean. This sampling distribution depends on:

1. n

Definition: The standard deviation of this distribution is called the *standard error* of the estimator.

Estimators and Their Distributions

We use estimators to summarize our i.i.d. sample. Any estimator, including the sample mean ___ is a random variable (since it is based on a random sample).

This means that ___ has a distribution of it's own, which is referred to as sampling distribution of the sample mean. This sampling distribution depends on:

1. n
2. population distribution

Definition: The standard deviation of this distribution is called the *standard error* of the estimator.

Estimators and Their Distributions

We use estimators to summarize our i.i.d. sample. Any estimator, including the sample mean ___ is a random variable (since it is based on a random sample).

This means that ___ has a distribution of it's own, which is referred to as sampling distribution of the sample mean. This sampling distribution depends on:

1. n
2. population distribution
3. method of sampling

Definition: The standard deviation of this distribution is called the *standard error* of the estimator.

Distribution of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with known mean value and standard deviation . Then:

$$E[\bar{X}] =$$

$$Var[\bar{X}] =$$

The standard deviation of the sample mean is:

This is also called the standard error of the mean.

Distribution of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with known mean value and standard deviation . Then:

$$E[\bar{X}] = \mu$$

$$Var[\bar{X}] = \frac{\sigma^2}{n}$$

The standard deviation of the sample mean is:

This is also called the standard error of the mean.

Distribution of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with known mean value and standard deviation . Then:

$$E[\bar{X}] =$$

$$Var[\bar{X}] =$$

The standard deviation of the sample mean is:

$$s.e.(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

This is also called the standard error of the mean.

Distribution of the Sample Mean

What does this mean? Why is it true?

$$E[\bar{X}] =$$

$$Var[\bar{X}] =$$

Also, what do we know about the *distribution* of the sample mean?

Distribution of the Sample Mean

What does this mean? Why is it true?

$$E[\bar{X}] = E\left[\frac{\sum X_i}{n}\right] = \frac{\sum E[X_i]}{n} = \frac{n\mu}{n} = \mu$$

$$Var[\bar{X}] =$$

Also, what do we know about the *distribution* of the sample mean?

Distribution of the Sample Mean

What does this mean? Why is it true?

$$E[\bar{X}] = E\left[\frac{\sum X_i}{n}\right] = \frac{\sum E[X_i]}{n} = \frac{n\mu}{n} = \mu$$

$$Var[\bar{X}] = Var[\sum X_i/n] = \frac{1}{n^2}\sum Var[X_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Also, what do we know about the *distribution* of the sample mean?

Distribution of the Sample Mean (Normal Population)

Proposition:
If $X_1, X_2, \ldots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then

We know everything there is to know about the distribution of the sample mean when the population distribution is normal.
This happens to be a result of that "a sum of normal random variables is still normal."

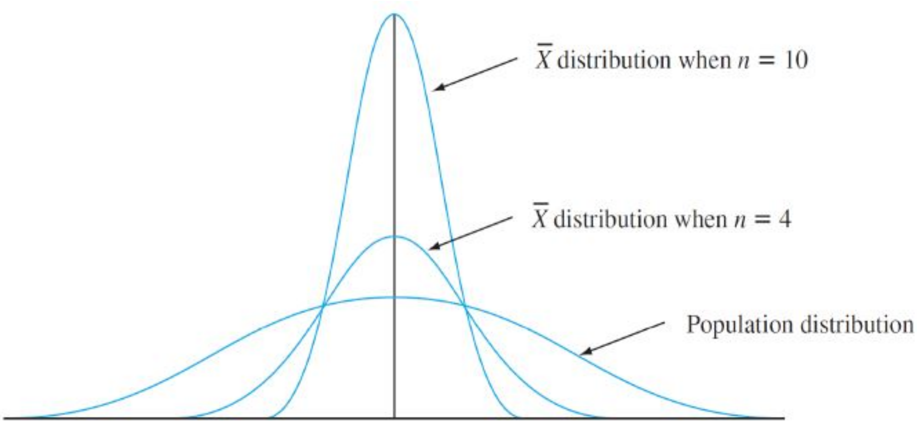
Distribution of the Sample Mean (Normal Population)

Proposition:
If $X_1, X_2, \ldots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then

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

We know everything there is to know about the distribution of the sample mean when the population distribution is normal.
This happens to be a result of that "a sum of normal random variables is still normal."

Distribution of the Sample Mean (Normal Population)



Central Limit Theorem

But what if the underlying distribution of the X_i's is not normal?

Central Limit Theorem

Important: When the population distribution is nonnormal, averaging produces a distribution more bellshaped than the one being sampled.

A reasonable conjecture is that if n is large, a suitable normal curve will approximate the actual distribution of the sample mean.

The formal statement of this result is one of the most important theorems in probability: *Central Limit Theorem!*

Central Limit Theorem

Theorem: *Central Limit Theorem:*

Central Limit Theorem

Theorem: *Central Limit Theorem:*
Let $X_1, X_2, \ldots X_n$ be iid from a distribution with mean μ and variance σ^2. Then, for n *large enough*:

Central Limit Theorem

Theorem: *Central Limit Theorem:*
Let $X_1, X_2, \ldots X_n$ be iid from a distribution with mean μ and variance σ^2. Then, for n *large enough*:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

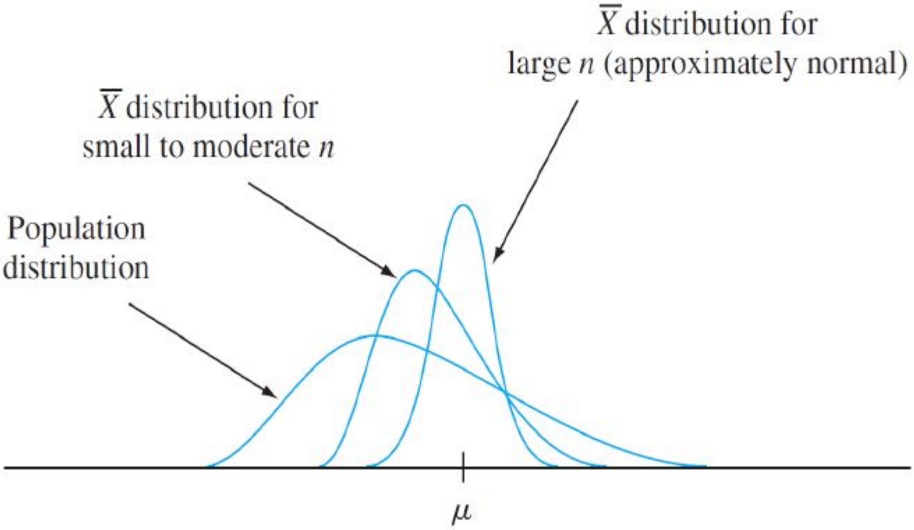
Central Limit Theorem

Theorem: *Central Limit Theorem:*
Let $X_1, X_2, \ldots X_n$ be iid from a distribution with mean μ and variance σ^2. Then, for n *large enough*:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The larger the value of n, the better the approximation! Typical rule of thumb: $n > 30$.

Central Limit Theorem



Central Limit Theorem

The CLT provides insight into why many random variables have probability distributions that are approximately normal.

For example, the measurement error in a scientific experiment can be thought of as the sum of a number of underlying perturbations and errors of small magnitude.

A practical difficulty in applying the CLT is in knowing when n is sufficiently large. The problem is that the accuracy of the approximation for a particular n depends on the shape of the original underlying distribution being sampled.

Daily Recap

Today we learned

1. The Normal Distribution... and why we care!

Moving forward:

- nb day Friday!

Next time in lecture:

- Using Normals to estimate *population* means based on *sample* means