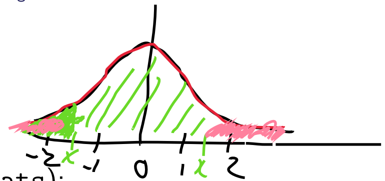
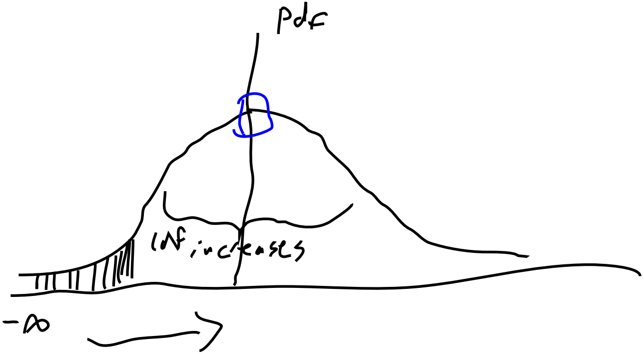
CSCI 3022 Intro to Data Science
Normals and the CLT

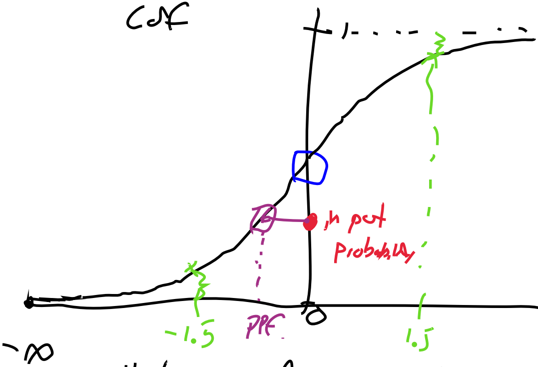


The four big functions (`scipy.stats as stats`):

1. `stats.normal.rvs(params, size=...)` generates random normals. Simulate
2. `stats.normal.pdf(x,params)` returns the pdf of the normal. It's the bell curve itself. It's **symmetric**: the pdf is the same height equal-amount left-right of 0.
3. `stats.normal.cdf(x,params)` returns the cdf of the normal. It's the area to the left of the input x value on the bell curve. It's also **symmetric**, but slightly different: the area to the *left* of an input value x is the same as the area to the *right* of negative x.
4. `stats.normal.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 $p = P(X \leq x)$.

Sketching areas on Normals





"slope of the cdf" = "value of the pdf"

PPF: backwards cdf

- PPF(.25) = First quartile
- PPF(.5) = median

Announcements and Reminders

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

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:

Center

Spread/height

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

spread of X

height of y.

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

Definition: *Standard Normal Distribution:*

The normal distribution with parameter values $\underline{\mu = 0}$ and $\underline{\sigma^2 = 1}$ is called the *standard normal distribution*. has named cdf $\Phi(x)$ or $\Phi(z)$.

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 $\underline{X \sim N(\mu, \sigma^2)}$

Definition: *Standard Normal Distribution:*

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

The Normal Distribution

Definition: *Normal Distribution:*

A continuous r.v. X is said to have a *normal distribution* with parameters _ and __ > 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 ____________

Definition: *Standard Normal Distribution:*

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

Non-Standard Normals

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

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

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:

- mean

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

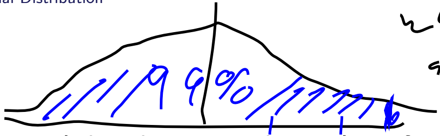
Standard deviation

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

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

is distributed standard normal.

Standard Quantiles



where do I draw a line so that 99% of area is left of that line

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.

$\Phi(z) = .99$

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

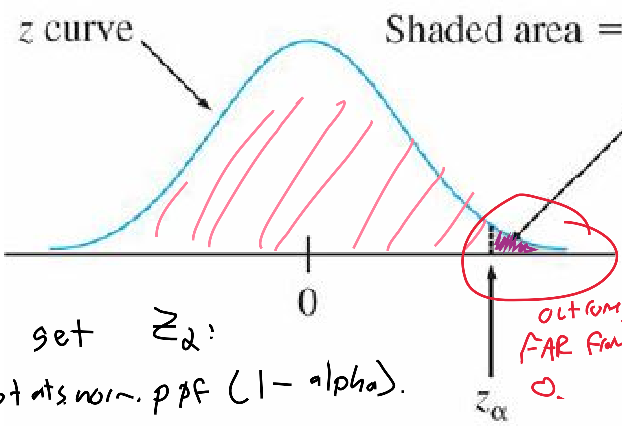
$\Rightarrow \quad z = \Phi^{-1}(.99)$

$= \text{normal.ppf}(.99)$

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: z_α will denote the z value for which α of the area under the z curve lies to the right of α.



$= \alpha$

$= \Phi(z_\alpha)$

$+ = 1$

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

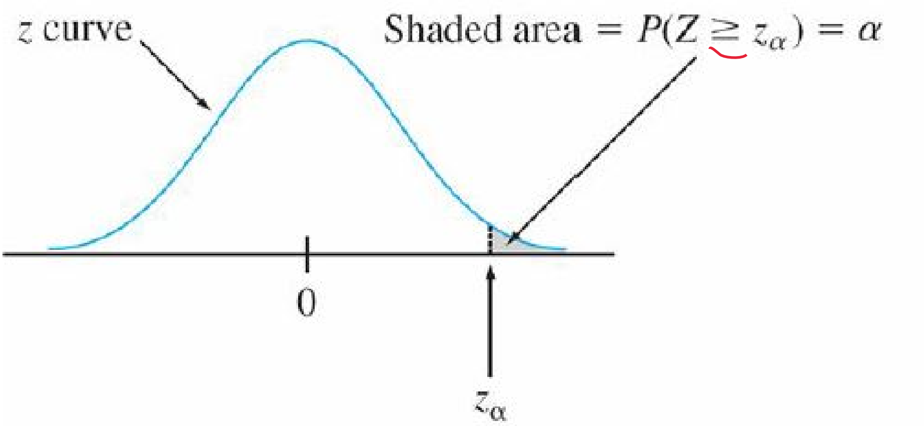
$z_\alpha = \Phi^{-1}(1-\alpha)$

to get z_α:

stats.norm.ppf(1- alpha).

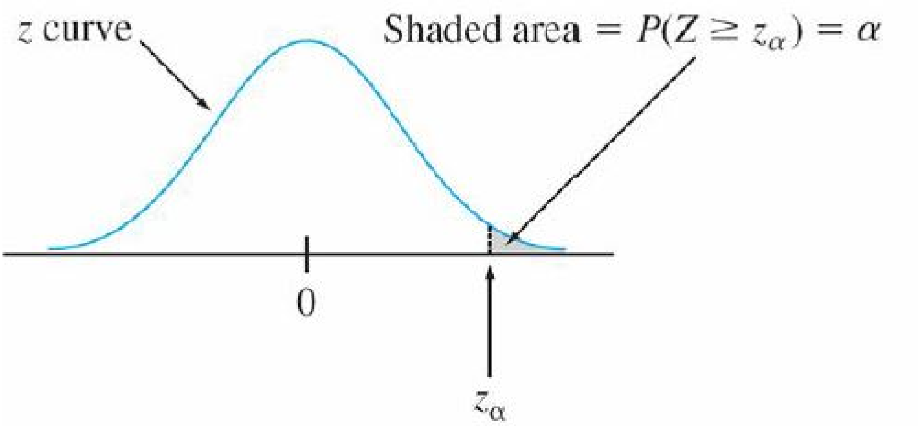
Standard Quantiles

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



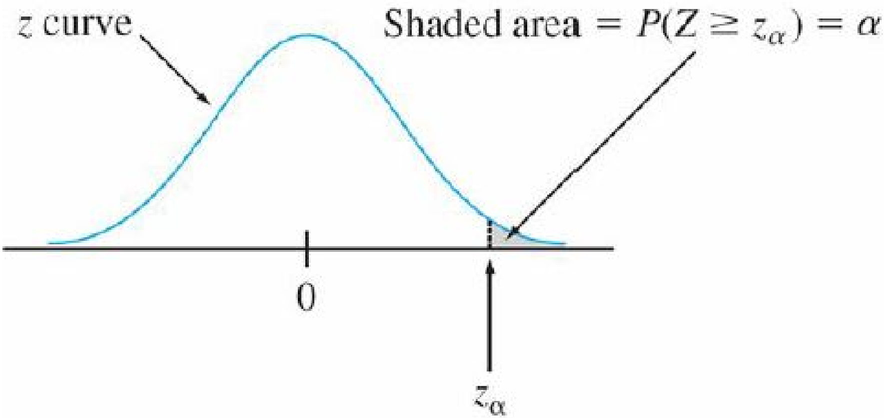
Standard Quantiles

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 $\underline{z_\alpha}$.



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.

2.

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

We use estimators to summarize our i.i.d. sample. → describe population

histogram / plot of pdf.

Examples? data itself

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 $\bar{X}$ is a random variable (since it is based on a random sample).

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

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 → more samples = good

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 → spread out data needs more observations.

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

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

"on average, a sample mean gives us pop mean"

theory μ

data

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

theory average

sample average

$$E[\bar{X}] = E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \frac{1}{n} \cdot E[X_1 + X_2 + \cdots + X_n]$$

lin.

$$= \frac{1}{n}\left(E[X_1] + E[X_2] + \cdots E[X_n]\right)$$

identical

$$= \frac{1}{n}[\mu + \mu + \cdots + \mu]$$

$$= \frac{1}{n}(\mu n) = \mu$$

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

The standard deviation of the sample mean is:

This is also called the standard error of the mean.

Distribution of the Sample Mean

in general, Variance of Sample mean decreases as n increases

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

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

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

$$Var[\bar{X}] = Var\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right]$$

$$= \frac{1}{n^2} Var[X_1 + X_2 + \cdots + X_n]$$

$$= \frac{1}{n^2}\left[Var[X] \cdot n\right]$$

$$= \boxed{\frac{Var[X]}{n}}$$

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:

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

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

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?

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

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

Theorem: That $\bar{X}$ approaches μ as $n \to infty$ is known as *the law of large numbers*.

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?

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

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

Theorem: That $\bar{X}$ approaches μ as $n \rightarrow infty$ is known as *the law of large numbers*.

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?

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

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

data theory

Theorem: That $\bar{X}$ approaches μ as $n \rightarrow infty$ is known as *the law of large numbers*.

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

Distribution of the Sample Mean (Normal Population)

Proposition:

If $\underbrace{X_1, X_2, \ldots X_n}_{n \text{ "normals"}} \overset{iid}{\sim} N(\mu, \sigma^2)$, then $\longrightarrow$ $\bar{X}$ is normal

L.L.N.: $\bar{X}$ has mean μ

$\bar{X}$ has variance σ^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."

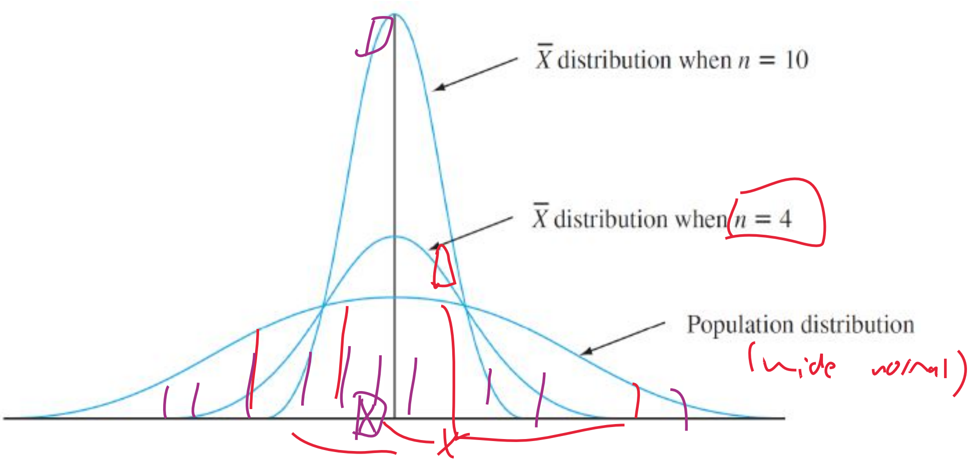
Distribution of the Sample Mean (Normal Population)

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."

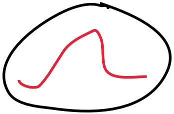
Distribution of the Sample Mean (Normal Population)

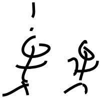


Central Limit Theorem

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

Central Limit Theorem





: wait, there all normals?

: it always been

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:* TL;DR: averaging → bell curve

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*:

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

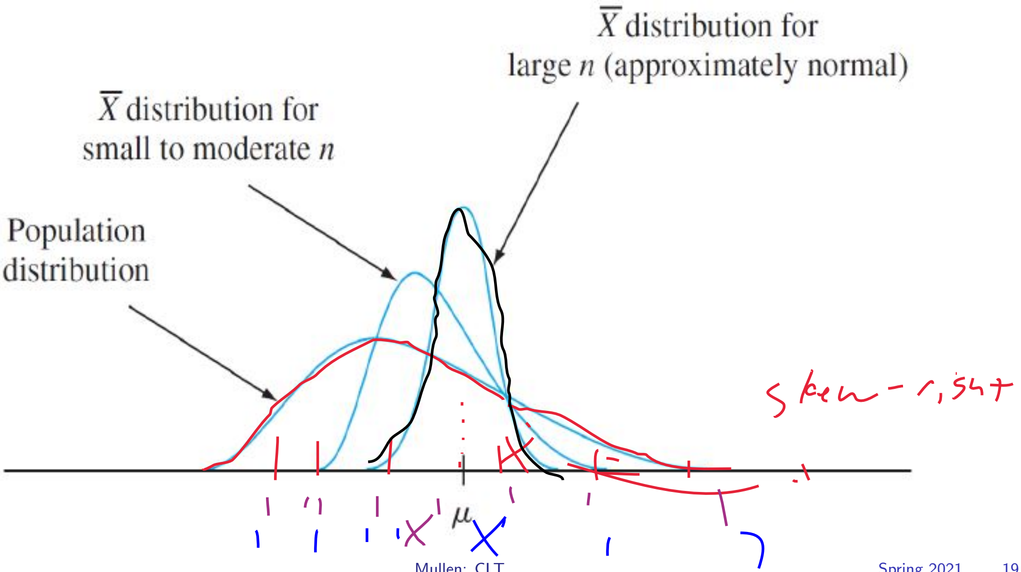
Central Limit Theorem

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: $\boxed{n > 30}$!

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.

So, what?

The CLT tells us that as the sample size n increases, the sample mean $\bar{X}$ is close to normally distributed with expected value of the true population mean μ and with a *smaller* standard deviation $\sigma/\sqrt{n}$.

So, what?

The CLT tells us that as the sample size n increases, the sample mean $\bar{X}$ is close to normally distributed with expected value of the true population mean μ and with a *smaller* standard deviation $\sigma/\sqrt{n}$.
Standarding the sample mean by first subtrating the expected value and then dividing by the standard deviation yields a standard normal random variable.

random

theoretical mean

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

standard ~~deviation~~ error
of the mean.

So, what?

The CLT tells us that as the sample size n increases, the sample mean $\bar{X}$ is close to normally distributed with expected value of the true population mean μ and with a *smaller* standard deviation $\sigma/\sqrt{n}$.
Standarding the sample mean by first subtrating the expected value and then dividing by the standard deviation yields a standard normal random variable.

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

This *always* works if the population is normally distributed and σ, μ are known. If it's not normally distributed, we needed a large enough sample size.

Using the Central Limit Theorem

Example: The amount of impurity in a batch of a chemical product is a random variable with mean value 4.0 g and standard deviation 1.5 g. (unknown distribution)

If 50 batches are independently prepared, what is the (approximate) probability that the average amount of impurity in these 50 batches is between 3.5 and 3.8 g?

Example sol:

Example sol:

We want the probability $P(3.5 < \bar{X} < 3.8)$ for $X \sim N(4.0, 1.5)$. Again we normalize... but $\bar{X}$ has much smaller standard deviation than each one of the individual data values!

$$P(3.5 < \bar{X} < 3.8) = P\left(\frac{3.5 - 4.0}{1.5/\sqrt{50}} < \frac{\bar{X} - 4.0}{1.5/\sqrt{50}} < \frac{3.8 - 4.0}{1.5/\sqrt{50}}\right)$$

$$= P\left(\frac{-1}{3/\sqrt{50}} < Z < \frac{-2}{15/\sqrt{50}}\right)$$

for $Z \sim N(0, 1)$ which is

$$\Phi\left(\frac{-2}{15/\sqrt{50}}\right) - \Phi\left(\frac{-1}{3/\sqrt{50}}\right)$$

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.

and to data!

What was the point of all this? We want the extract or infer properties of populations (like μ!) by analyzing samples. To do this, we ask:

and to data!

What was the point of all this? We want the extract or infer properties of populations (like μ!) by analyzing samples. To do this, we ask:

1. Is the sample mean $\bar{x}$ a good approximation of the population mean μ?
2. Is the sample proportion $\hat{p}$ a good approximation of the population proportion p?
3. Are two samples coming from populations with different means?

and to data!

What was the point of all this? We want the extract or infer properties of populations (like μ!) by analyzing samples. To do this, we ask:

1. Is the sample mean $\bar{x}$ a good approximation of the population mean μ?
2. Is the sample proportion $\hat{p}$ a good approximation of the population proportion p?
3. Are two samples coming from populations with different means?
4. **If Yes,** how sure or confident are we?

and to data!

What was the point of all this? We want the extract or infer properties of populations (like μ!) by analyzing samples. To do this, we ask:

1. Is the sample mean $\bar{x}$ a good approximation of the population mean μ?
2. Is the sample proportion $\hat{p}$ a good approximation of the population proportion p?
3. Are two samples coming from populations with different means?
4. **If Yes,** how sure or confident are we?
5. How much data would we need to be sure or confident?

Confidence Interval for the Mean (SD known)

Because the area under the standard normal curve between –1.96 and 1.96 is 0.95, we know:

This is equivalent to:

Confidence Interval for the Mean (SD known)

Because the area under the standard normal curve between –1.96 and 1.96 is 0.95, we know:

$$.95 = P(-1.96 < Z < 1.96)$$

This is equivalent to:

$$.95 = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right)$$

Confidence Interval for the Mean (SD known)

Because the area under the standard normal curve between –1.96 and 1.96 is 0.95, we know:

$$.95 = P(-1.96 < Z < 1.96)$$

This is equivalent to:

$$.95 = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right)$$

We want to know things about μ, however!

Confidence Interval for the Mean (SD known)

Because the area under the standard normal curve between –1.96 and 1.96 is 0.95, we know:

This is equivalent to:

We want to know things about μ, however!
The 95% confidence interval for μ is the values of X that satisfy this inequality.

Solving for μ:

The interval:

$$.95 = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right)$$

Solving for μ:

The interval:

$$.95 = P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right)$$

$$.95 = P\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Solving for μ:

The interval:

$$.95 = P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right)$$

$$.95 = P\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$.95 = P\left(1.96\frac{\sigma}{\sqrt{n}} > \mu-\bar{X} > -1.96\frac{\sigma}{\sqrt{n}}\right)$$

Solving for μ:

The interval:

$$.95 = P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right)$$

$$.95 = P\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$.95 = P\left(1.96\frac{\sigma}{\sqrt{n}} > \mu-\bar{X} > -1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$.95 = P\left(\bar{X}-1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right)$$

Confidence Interval for the Mean (SD known)

The interval

$$.95 = P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Is called a 95% confidence interval for the mean.

This interval varies from sample to sample, as the sample mean varies. So, the interval itself is a random interval.

Which parts of the interval are random?

Confidence Interval for the Mean (SD known)

The interval

$$.95 = P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Is called a 95% confidence interval for the mean.

This interval varies from sample to sample, as the sample mean varies. So, the interval itself is a random interval.

Which parts of the interval are random? The two copies of $\bar{X}$

Confidence Interval for the Mean (SD known)

The CI is centered at __ and extends __________ to each side in the x direction.

That width of __________ is not random; only the location of the interval (its midpoint $\bar{X}$) is random.

Confidence Interval for the Mean (SD known)

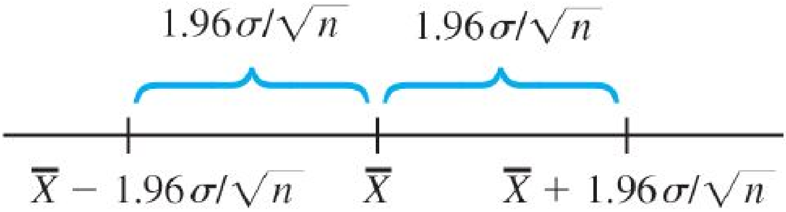
The CI is centered at $\underline{\bar{X}}$ and extends $\underline{1.96 \cdot \sigma/\sqrt{n}}$ to each side in the x direction.

That width of $\underline{1.96 \cdot \sigma/\sqrt{n}}$ is not random; only the location of the interval (its midpoint $\bar{X}$) is random.

Confidence Interval for the Mean (SD known)

The CI is centered at $\underline{\bar{X}}$ and extends __________ to each side in the x direction.

That width of $\underline{1.96 \cdot \sigma/\sqrt{n}}$ is not random; only the location of the interval (its midpoint $\bar{X}$) is random.



Confidence Interval for the Mean (SD known)

As we showed, for a given sample, the CI can be expressed as

$$.95 = P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

A couple of concise expressions for the interval are

where the left endpoint is the lower limit and the right endpoint is the upper limit.

Confidence Interval for the Mean (SD known)

As we showed, for a given sample, the CI can be expressed as

$$.95 = P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

A couple of concise expressions for the interval are

$$[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}]$$

where the left endpoint is the lower limit and the right endpoint is the upper limit.

Confidence Interval for the Mean (SD known)

As we showed, for a given sample, the CI can be expressed as

$$.95 = P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

A couple of concise expressions for the interval are

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

where the left endpoint is the lower limit and the right endpoint is the upper limit.

Interpreting CIs

We are "95% confident" that the true parameter is in this interval.

What does that mean??

A correct interpretation of "95% confidence" relies on the long-run relative frequency interpretation of probability.

Interpreting CIs

We are "95% confident" that the true parameter is in this interval.

What does that mean??

A correct interpretation of "95% confidence" relies on the long-run relative frequency interpretation of probability.

In **repeated** sampling, 95% of the confidence intervals obtained from all samples will actually contain . The other 5% of the intervals will not.

Interpreting CIs

We are "95% confident" that the true parameter is in this interval.

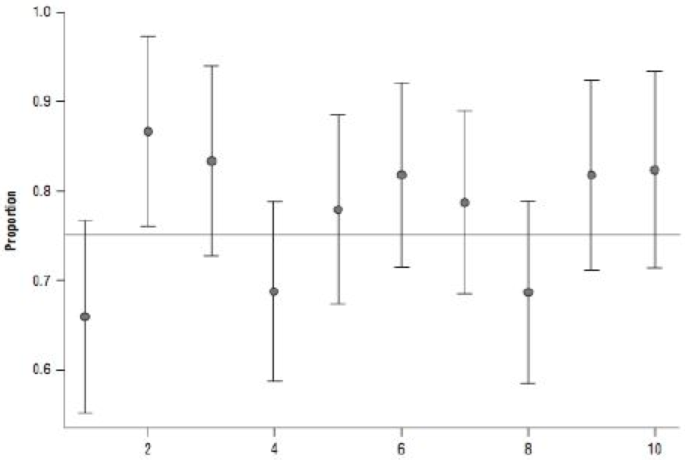
What does that mean??

A correct interpretation of "95% confidence" relies on the long-run relative frequency interpretation of probability.

The confidence level is not a statement about any particular interval instead it pertains to what would happen if a very large number of like intervals were to be constructed using the same CI formula.

Interpreting CIs

Figure 1: Confidence Interval



Note: Suppose that the true proportion of believers in climate change among French citizens is 0.75, as represented by the horizontal black line near the middle. This figure shows ten 90% confidence intervals used to estimate the

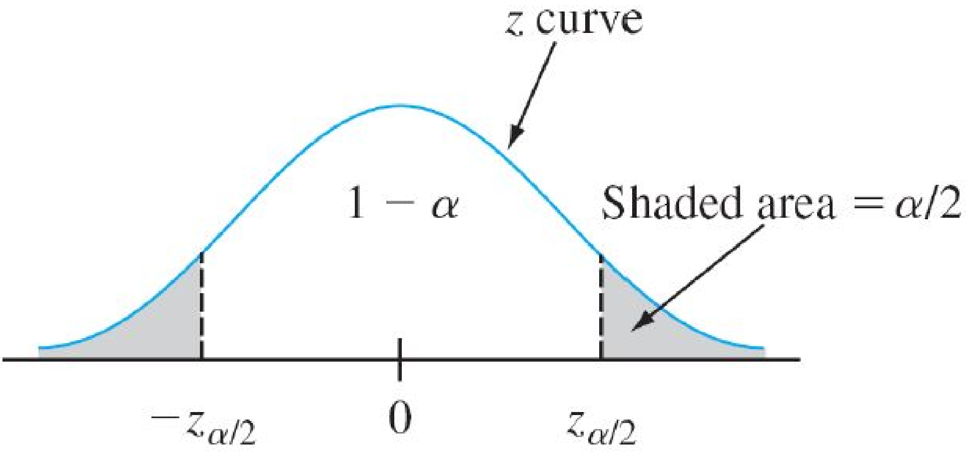
Interpreting CIs

Some reading on the common misinterpretations of CIs:

http://www.ejwagenmakers.com/inpress/HoekstraEtAlPBR.pdf

Other Levels of Confidence

A confidence level of $1-\alpha$ can be achieved by using another $z_{\alpha/2}$ in place of $z_{0.025} = 1.96$:



Other Levels of Confidence

A $100(1-\alpha)\%$ confidence interval for the mean when the value of α is known is given by:

Or, equivalently, by:

Other Levels of Confidence

A $100(1-\alpha)\%$ confidence interval for the mean when the value of α is known is given by:

$$1-\alpha = P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Or, equivalently, by:

Other Levels of Confidence

A $100(1-\alpha)\%$ confidence interval for the mean when the value of α is known is given by:

Or, equivalently, by:

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Confidence Interval for the Mean (SD known)

Example:

A sample of 40 units is selected and diameter measured for each one. The sample mean diameter is 5.426 mm, and the standard deviation of measurements is 0.1mm.

1. Calculate a confidence interval for true average hole diameter using a confidence level of 90%.

2. What about the 99% confidence interval?

3. What are the advantages and disadvantages to a wider confidence interval?

Confidence Interval for the Mean (SD known)

Example:

A sample of 40 units is selected and diameter measured for each one. The sample mean diameter is 5.426 mm, and the standard deviation of measurements is 0.1mm.

1. Calculate a confidence interval for true average hole diameter using a confidence level of 90%.

$$5.426 \pm z_{.05}\frac{0.1}{\sqrt{40}}$$

2. What about the 99% confidence interval?

3. What are the advantages and disadvantages to a wider confidence interval?

Confidence Interval for the Mean (SD known)

Example:

A sample of 40 units is selected and diameter measured for each one. The sample mean diameter is 5.426 mm, and the standard deviation of measurements is 0.1mm.

1. Calculate a confidence interval for true average hole diameter using a confidence level of 90%.

$$5.426 \pm z_{.05} \frac{0.1}{\sqrt{40}}$$

2. What about the 99% confidence interval?

$$5.426 \pm \texttt{scipy.stats.ppf(.995)} \frac{0.1}{\sqrt{40}}$$

3. What are the advantages and disadvantages to a wider confidence interval?

Sample Size Calculations

For a desired confidence level and interval width, we can determine the necessary sample size.

Example: For a given computer model, memory fetch response time is normally distributed with standard deviation of 25 milliseconds. A new computer has been purchased, and we wish to estimate the true average response time. What sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10 units?

Sample Size Calculations

For a desired confidence level and interval width, we can determine the necessary sample size.

Example: For a given computer model, memory fetch response time is normally distributed with standard deviation of 25 milliseconds. A new computer has been purchased, and we wish to estimate the true average response time. What sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10 units?
The width is $W = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. We want:

Sample Size Calculations

For a desired confidence level and interval width, we can determine the necessary sample size.

Example: For a given computer model, memory fetch response time is normally distributed with standard deviation of 25 milliseconds. A new computer has been purchased, and we wish to estimate the true average response time. What sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10 units?

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < 10$$

$$\implies \; z_{\alpha/2}\frac{\sigma}{10} < \sqrt{n}$$

$$\implies \; \left(z_{\alpha/2}\frac{\sigma}{10}\right)^2 < n$$

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