CSCI 3022-002 Intro to Data Science
Statistical Inference

$\underbrace{\text{from sample}}$ $X_1, X_2, \dots, X_n$ $\rightarrow$ X_i mean μ
Var σ^2

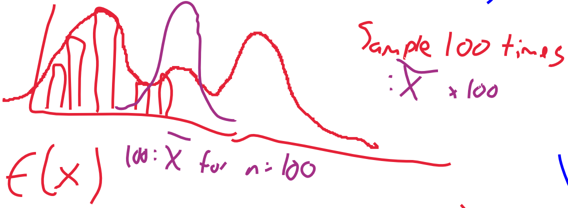
1) a t $\lim_{n\to\infty}$ $\bar{X} \rightarrow \mu$ $\quad \mu = E[X]$

Recap: what did we learn about means?

2) ALSO

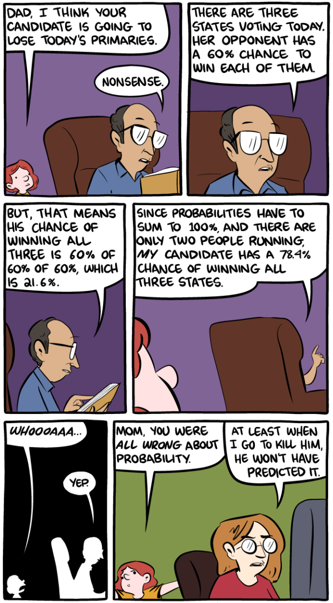
Var($\bar{X}$) decreases compared to Var(X_i) individual values

$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \implies S.D.(\bar{X}) = \sigma/\sqrt{n}$



3) large n: $\bar{X} \sim N(\mu, \sigma^2/n)$

Centering in


DAD, I THINK YOUR CANDIDATE IS GOING TO LOSE TODAY'S PRIMARIES.
NONSENSE.
THERE ARE THREE STATES VOTING TODAY. HER OPPONENT HAS A 60% CHANCE TO WIN EACH OF THEM.
BUT, THAT MEANS HIS CHANCE OF WINNING ALL THREE IS 60% OF 60% OF 60%, WHICH IS 21.6%.
SINCE PROBABILITIES HAVE TO SUM TO 100%, AND THERE ARE ONLY TWO PEOPLE RUNNING, MY CANDIDATE HAS A 78.4% CHANCE OF WINNING ALL THREE STATES.
WHOOOAAA...
YEP.
MOM, YOU WERE ALL WRONG ABOUT PROBABILITY.
AT LEAST WHEN I GO TO KILL HIM, HE WON'T HAVE PREDICTED IT.

smbc-comics.com

- Practicum done!
- Exam next: just pen and paper problems. 3x example exams *are posted!* → following Monday, Oct 26!
- $\boxed{\text{HW 5}}$ (due next monday) Oct 19

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:

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

sample

pop

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

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.

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

$$Var\left[\frac{\Sigma X_i}{n}\right]$$

$$\frac{1}{n^2} Var[\Sigma X_i]$$

$$= Var[X_i]/n$$

Where we at?

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

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

Shape Normal

mean μ

var σ^2/n

Where we at?

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: $n > 30$.
It's all normal? Always has been.

Where we at?

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: $n > 30$.
It's all normal? Always has been.

The Normal cdf
We use the cdf $\Phi(z) = P(Z \leq z)$ to calculate probabilities on normal distributions.

Where we at?

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*:

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

Normal!

The larger the value of n, the better the approximation! Typical rule of thumb: $n > 30$.
It's all normal? Always has been.

The Normal cdf Stats.norm.cdf
We use the cdf $\boxed{\Phi}(z) = P(Z \leq z)$ to calculate probabilities on normal distributions.
The Normal cdf inverse (ppf) Stats.norm.ppf
We use the inverse of the cdf to calculate z values that correspond to specific probabilities.
Notation: $\alpha = P(Z > \boxed{z_\alpha})$, so α is the probability value to the **right** of the $x-$value z_α.

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.

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

$\to$ divide by $\;s.d.(\bar{X}) \Rightarrow \overset{\text{new}}{s.d.} = 1.$

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.

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

→ NEVER!

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)

$\mu = 4.0$ $\sigma = 1.5$

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?

$\hookrightarrow n = 50$:

find $P(3.5 \leq \bar{X} \leq 3.8)$

(maybe small? true $\mu = 4.0$, <u>not</u> in interval).

Example sol:

want: $P(3.5 \le \bar{X} \le 3.8)$ for $\mu = 4.0$, $\sigma = 1.5$

$\bar{X}: \mu_{\bar{X}} = 4 \quad \sigma_{\bar{X}} = \sigma/\sqrt{n}$

Recall: $Z = \dfrac{X-\mu}{\sigma}$ for X normal.

we know: $\bar{X} \sim N\left(4.0, \frac{1.5}{\sqrt{50}}\right)$

want Z!

Normal thing

$$Z = \frac{\bar{X} - 4.0}{1.5/\sqrt{50}}$$

mean

s.e.

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

$$P(3.5 \le \bar{X} \le 3.8)$$

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

$$= P(a \le Z \le b) = \Phi(b) - \Phi(a)$$

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

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

stats.norm.cdf( ) - stats.norm.cdf ( )

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!

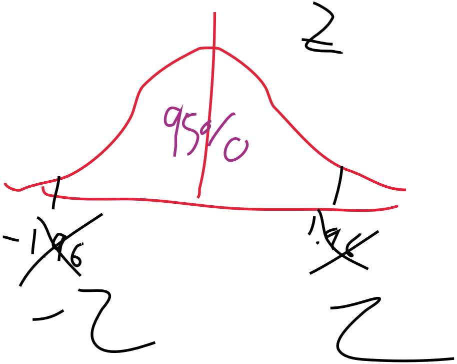
What was the point of all this? We want the extract or infer properties of populations (like μ!) by analyzing samples. To do this, we ask:

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:

$$.95 = P(\underbrace{-1.96 < Z < 1.96}_{\text{95\% of the time...}})$$

↳ Z is between ±2.

This is equivalent to:

$$.95 = P\left(-1.96 < \boxed{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}} < 1.96\right) \rightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ is between } \pm 2.$$

Want result to be about μ
(given $\bar{X}$).

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:

$$.95 = P\left(\overset{\overline{X}}{\downarrow}\hspace{-0.5em}\# < \mu < \overset{\overline{X}}{\downarrow}\hspace{-0.5em}\#\right)$$

Remember: $\overline{X}$ is random
μ is not.

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:

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

multiply

Solving for μ:

The interval:

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

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

Subtract $\bar{X}$

Solving for μ:

The interval:

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

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

multiply by (-1).

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

Solving for μ:

The interval:

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

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

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

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

important thing!

$\bar{X} + \#$

$\bar{X} - \#$

$\#$: (1.96) times $\sigma/\sqrt{n}$

Confidence Interval for the Mean (SD known)

The interval

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

Is called a 95% confidence interval for the mean.

x-values for 95% of area.

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

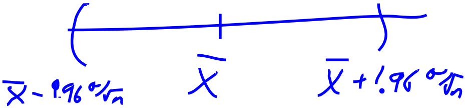
$$.95 = P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

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)



$\bar{X} - 1.96\,\sigma/\sqrt{n}$
$\bar{X}$
$\bar{X} + 1.96\,\sigma/\sqrt{n}$

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

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

Confidence Interval for the Mean (SD known)

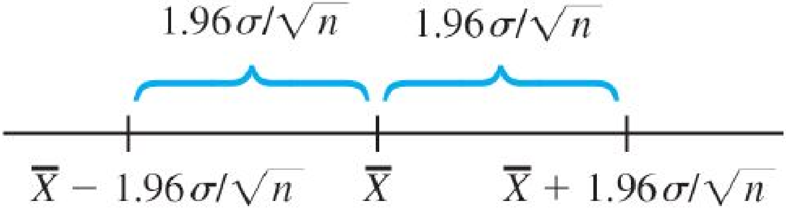
The CI is centered at $\underline{\bar{X}}$ and extends $\underline{1.96 \cdot \sigma/\sqrt{n}}$ to each side in the x direction.

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.


$1.96\sigma/\sqrt{n}$
$1.96\sigma/\sqrt{n}$
$\overline{X} - 1.96\sigma/\sqrt{n}$
$\overline{X}$
$\overline{X} + 1.96\sigma/\sqrt{n}$

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

$$\left[\bar{X} - 1.96\,\sigma/\sqrt{n},\ \bar{X} + 1.96\,\sigma/\sqrt{n}\right]$$

$$\left[\bar{X} \pm 1.96\,\sigma/\sqrt{n}\right]$$

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.

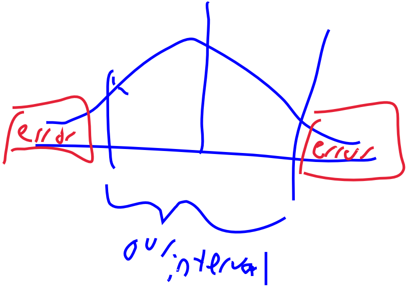
Confidence Interval for the Mean (SD known)

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

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



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 μ. 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

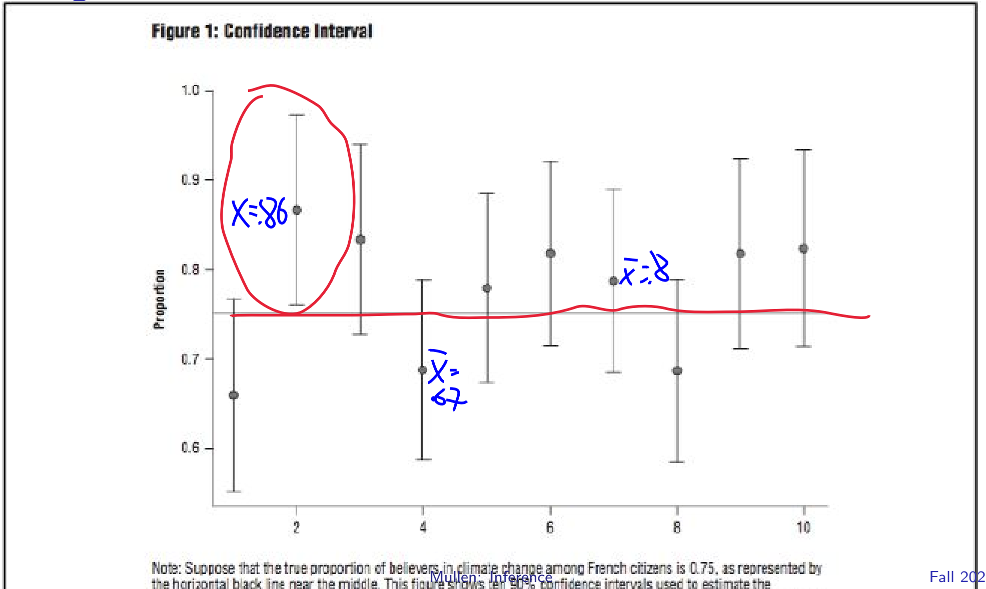


Figure 1: Confidence Interval

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

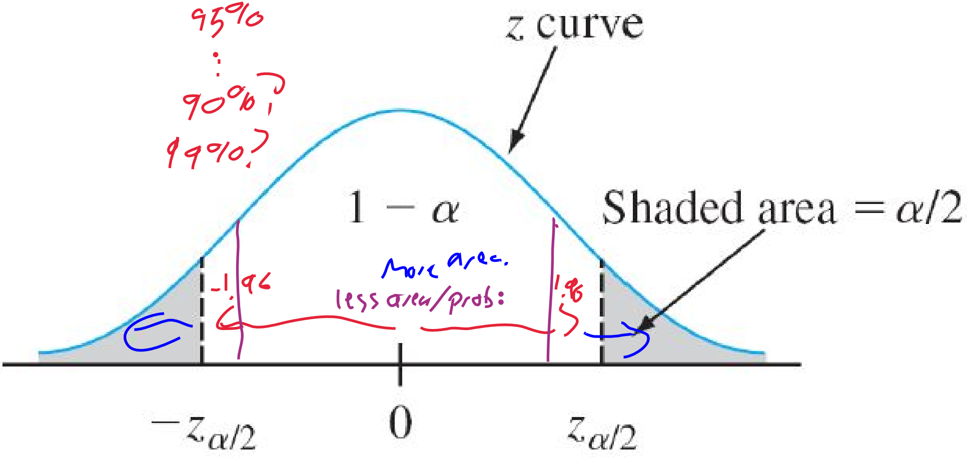
A confidence level of $1-\alpha$ can be achieved by using another $z_{\alpha/2}$ in place of $z_{0.025} = 1.96$:

instead of 95%

⋮

90%?

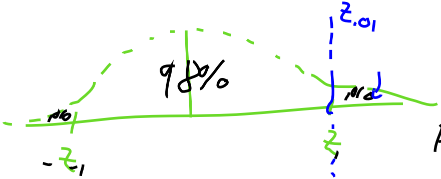
99%?



Other Levels of Confidence

98%: α error = .02

$P(Z \le -z_1) = .01$



$P(Z > z_1) = .01$

$P(Z > z_{.01}) = .01$

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

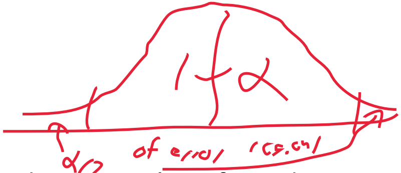
90%...

99%

Or, equivalently, by:

Other Levels of Confidence

check: $1 - \alpha + \frac{\alpha}{2} + \frac{\alpha}{2} = 1$ ✓



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:

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

recall: $\alpha = 5\%$ $z_{\alpha/2} = z_{.025} = 1.96$

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?

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

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

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

Large Sample Confidence Intervals

$$\bar{X} \pm z_{\alpha/2} \overset{\text{s.e. of } \bar{X}}{\left(\sigma/\sqrt{n}\right)}$$

A difficulty in using our previous equation for confidence intervals is that it uses the value σ of which will rarely be known.

Also, we may want a CI for a mean from some other non-normal distribution.

Large Sample Confidence Intervals

In this instance, we need to work with the sample standard deviation s. Remember from our first lesson that the standard deviation is calculated as:

population st.dev $\sigma \approx$

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}}$$

With this, we instead work with the standardized random variable:

Large Sample Confidence Intervals

In this instance, we need to work with the sample standard deviation s. Remember from our first lesson that the standard deviation is calculated as:

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}}$$

With this, we instead work with the standardized random variable:

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

compared to

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Large Sample Confidence Intervals

Previously, there was randomness only in the numerator of Z by virtue of the estimator $\bar{X}$.

In the new standardized variable, both $\bar{X}$ and s vary in value from one sample to another.

When n is large, the substitution of s for σ adds little extra variability, so nothing needs to change.

$s \approx \sigma$

When n is smaller, the distribution of this new variable should be wider than the normal to reflect the extra uncertainty. (We talk more about this in a week or two.)

Large Sample Confidence Intervals

Previously, there was randomness only in the numerator of Z by virtue of the estimator $\underline{\bar{X}}$.

In the new standardized variable, both $\underline{\bar{X}}$ and $\underline{s}$ vary in value from one sample to another.

When n is large, the substitution of s for σ adds little extra variability, so nothing needs to change.

When n is smaller, the distribution of this new variable should be wider than the normal to reflect the extra uncertainty. (We talk more about this in a week or two.)

Large Sample Confidence Intervals

Large Sample CI: If n is sufficiently large $(n \geq 30)$, the standardized random variable

has approximately a standard normal distribution. This implies that

is a large-sample confidence interval for μ with confidence level approximately $100(1-\alpha)\%$. This formula is valid regardless of the population distribution for sufficiently large n.

Large Sample Confidence Intervals

Large Sample CI: If n is sufficiently large $(n \geq 30)$, the standardized random variable

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

has approximately a standard normal distribution. This implies that

Same!

$$\bar{X} \pm z_{\alpha/2}\frac{\boxed{s}}{\sqrt{n}} \quad \sim \boxed{\sigma}$$

is a large-sample confidence interval for μ with confidence level approximately $100(1-\alpha)\%$. This formula is valid regardless of the population distribution for sufficiently large n.

Types of Confidence Intervals

	$n \geq 30$	$n < 30$
Underlying Normal Distribution	σ known	σ known
	σ unknown	σ unknown
Underlying Non-Normal Distribution	σ known	σ known
	σ unknown	σ unknown

$$Z = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

or $\sigma/\sqrt{n}$

Sum of normals is normal

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

↳ small samples!

Method:

Z or approximately Z by Central Limit Theorem

Special Cases: Populations

Let p denote the proportion of ”successes” in a population (e.g., individuals who graduated from college, computers that do not need warranty service, etc.). A random sample of n individuals is selected, and X is the number of successes in the sample.

Then, X can be modeled as a ________ rv with mean of __ and

variance of _________

Special Cases: Populations

Let p denote the proportion of ”successes” in a population (e.g., individuals who graduated from college, computers that do not need warranty service, etc.). A random sample of n individuals is selected, and X is the number of successes in the sample.

Then, X can be modeled as a Binomial rv with mean of $\underline{np}$ and

variance of $\underline{np(1-p)}$

Special Cases: Populations

Let p denote the proportion of ”successes” in a population (e.g., individuals who graduated from college, computers that do not need warranty service, etc.). A random sample of n individuals is selected, and X is the number of successes in the sample.

Then, X can be modeled as a Binomial rv with mean of $\underline{np}$ and

variance of $\underline{np(1-p)}$

If both $np > 10$ and $n(1-p) > 10$, X has approximately a normal distribution.

Special Cases: Populations

The estimator of p is: $\hat{p}$= ____

Standardizing the estimator yields:

and a resulting CI is:

Special Cases: Populations

The estimator of p is: $\hat{p}= \underline{X/n}$

Standardizing the estimator yields:

and a resulting CI is:

Special Cases: Populations

The estimator of p is: $\hat{p}$= $\underline{X/n}$

This estimator is approximately normally distributed and:

$$E[\hat{p}] = p \qquad Var[\hat{p}] = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Standardizing the estimator yields:

and a resulting CI is:

Special Cases: Populations

The estimator of p is: $\hat{p}= \underline{X/n}$

This estimator is approximately normally distributed and:

$$E[\hat{p}] = p \qquad Var[\hat{p}] = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Standardizing the estimator yields:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

and a resulting CI is:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Special Cases: Populations

Example:
The EPA considers indoor radon levels above 4 picocuries per liter (pCi/L) of air to be high enough to warrant amelioration efforts. Tests in a sample of 200 homes found 127 (63.5%) of these sampled households to have indoor radon levels above 4 pCi/L. Calculate the 99% confidence interval for the proportional of homes with indoor radon levels above 4 pCi/L.

Special Cases: Populations

Example:
The EPA considers indoor radon levels above 4 picocuries per liter (pCi/L) of air to be high enough to warrant amelioration efforts. Tests in a sample of 200 homes found 127 (63.5%) of these sampled households to have indoor radon levels above 4 pCi/L. Calculate the 99% confidence interval for the proportional of homes with indoor radon levels above 4 pCi/L.

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}};$$

`stats.norm.ppf(0.995) = 2.57;`

use $\hat{p}$ where we must;

$$= 0.635 \pm 2.57\sqrt{\frac{0.635(1-0.635)}{200}}$$

$$= [0.548, 0.722]$$

Now what?

We used the central limit theorem to compute **confidence intervals**, which gave us an idea as to how close our *sample* mean is to the *population* mean. This worked:

1. If the population was normal.
2. If the population was non-normal with large n.
3. If the population was binomial with “enough” failures and successes.

Those are not all the things we care about! Next time, we use the idea of confidence to ask questions like "are A and B different?" We eventually also want to start building up what to do with smaller samples, and estimating things that aren't normally distributed (like estimating σ^2)!