Chi-Squared Distribution, t-distribution, CI for means

3.2, 5.5, 7.1

Today's topics

Review: Point estimators, Sample Variance

New terminology: Indicator functions, Order Statistics

New topics:

- Chi-squared Distribution
  - Degrees of freedom
  - Overview, relation to Normal distribution
  - Pdf and relation to Gamma distribution
- t-distribution
  - Definition
  - Uses in statistics
- Confidence Interval for means
  - Example: calculating s^2 and creating a CI for the mean

$\pi f(x_i)$

$L(\theta)$

MLE ← mode

Ω

θ

1c 1d

Review

1st T.m

- MLE
- MOM ← $E[X] \overset{set}{=} \bar{X}$
- Bias

1st S.m

- Sample mean and Sample Variance ← $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

$\bar{x}$

$\hat{\theta}$ function of data

Bias = $E[\hat{\theta}] - \theta$

notes

	W c1	L c2	
R	3	2	5
B	3	2	5
	6	4	$n=10$

#c-1

$df = 1$

$df = (c-1)(r-1)$

χ^2 adding st. normal^2 : # things adding

est. sample var : $\boxed{df=n-1}$

Sample variance, s²

$$\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

"random sample of size n"

Suppose we draw n iid observations from a distribution with mean μ and variance σ^2.

Ideally, we would estimate σ^2 with $\frac{1}{n}\sum(x_i - \mu)^2$.

μ unknown

3, 6, 5, 5, 8

$\bar{x}$

Problem: μ is usually unknown!

We can try replacing μ with the $\bar{x}$, giving $\frac{1}{n}\sum(x_i - \bar{x})^2$.

Unfortunately, this tends to underestimate σ^2.

To compensate, we divide by $n - 1$ instead.

σ^2

"df"

Degrees of Freedom

- The degrees of freedom is the number of values in the final calculation of a statistic or parameter that are free to vary.
- In other words: the number of observations that contain new information.

DF for calculating sample mean

Suppose we have the following sample: $10, 20, 30, 40$

We want to calculate the average.

All four numbers are free to vary, so $\boxed{\text{df} = 4}$

$\bar{x} = 25$

don't overthink this :)

10, 20, 30, 40

mean = 25

30, 30, 30, 10

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

DF for calculating sample variance:

In the sum, we have $(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2$.

If I am going to rely on the sample average to calculate the sample variance, it is going to "cost me" one degree of freedom.

cost 1 df

Using the same sample: 10, 20, 30, 40 $\bar{x} = 25$

Remember: $(x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) + (x_4 - \bar{x}) = 0$

only n−1 are free

using $\bar{x}=25$, $\underbrace{10, 20, 30}_{\text{free}}$ 40 not free

df = 3

$(10-25)^2 + \ldots$

notes

df how much info

mean

$x_1 = 1$

$df = 1$

$(1-4)^2 + (5-4)^2 + (6-4)^2$

$\bar{x} = 1$

Var $x_1 = 1$

$df = 0$

$\bar{x} = 4$

$x_1 = 1$ $x_2 = 5$

$x_3 = 6$

$df = n - 1 = \cancel{3}\ 2$

sample variance

Chi-squared distribution, χ^2 (overview)

The Chi-squared distribution is an important distribution that is frequently used in statistical inference.

Comes from summing the squares of Standard Normal RVs.

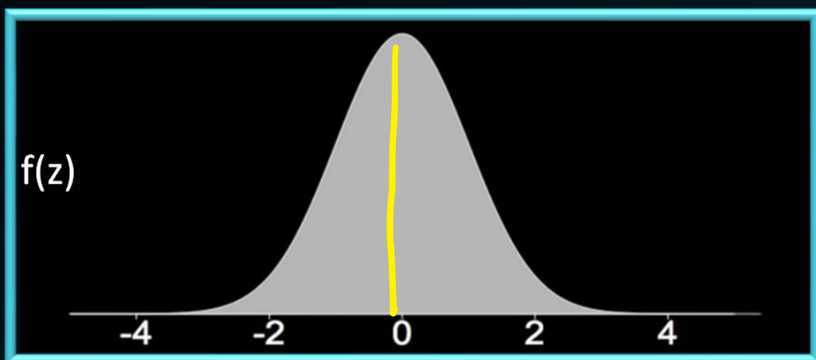
- If $Z \sim N(0,1)$, then Z^2 follows a Chi-Squared distribution with one degree of freedom. $Z^2 \sim \chi^2_{(1)}$

- If $Z_1, Z_2, \ldots Z_k, \overset{iid}{\sim} N(0,1)$, then $Z_1^2 + Z_2^2 + \ldots + Z_k^2 \sim \chi^2_{(k)}$

~~Chi~~

$Z \xrightarrow{\text{square}} \chi^2_{(1)}$

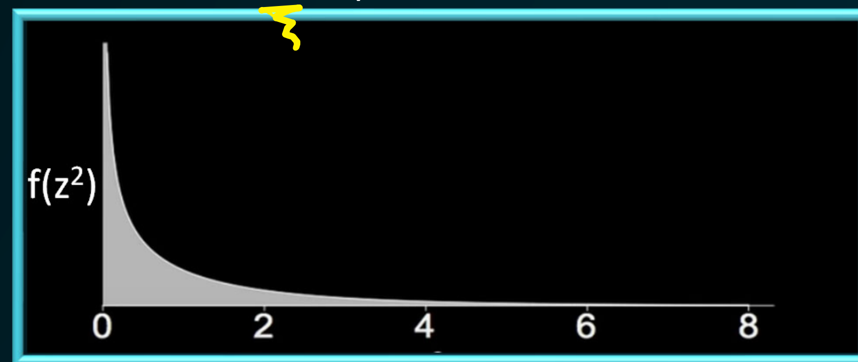
Standard Normal



z

$\sigma^2 = 1$

Chi-Squared, df = 1



z^2

Chi-squared distribution $X \sim \chi^2_{r}$

cont

χ

chi

df

$$f(x) = \frac{1}{\Gamma\left(\frac{r}{2}\right)2^{r/2}} x^{\frac{r}{2}-1} e^{-x/2}, \qquad 0 < x < \infty$$

χ^2_ν

We can see from this pdf that X also follows another distribution! $X \sim$ Gamma $(\alpha = r/2, \theta = 2)$

$$\mu_x = \alpha\theta = \left(\frac{r}{2}\right) \cdot 2 = r$$

$$\sigma_x^2 = \alpha\theta^2 = \left(\frac{r}{2}\right)2^2 = 2r$$

notes

$$\underbrace{Z_1, Z_2, \ldots Z_8}_{} \overset{iid}{\sim} N(0,1)$$

$$W = \sum_{i=1}^{8} \underbrace{Z_i^2}_{}$$

$$Var[W] = ?$$

$$W \sim \chi^2_{8}$$

$$E[W] = 8$$

$$Var[W] = 2 \cdot 8 = \underline{16}$$

$\theta = 2$

Gamma and Chi-Squared

The Chi-Squared distribution is also a special case of the Gamma distribution where $\theta = 2$.

If X $\sim Gamma(\alpha, 2)$, X $\sim \chi^2_{2\alpha}$

χ^2_r

Example:

X $\sim \chi^2_r$

$\sim$ Gamma(α =r/2, θ =2)

notes

Example:

Lets say poisson process with (every minute) rate lambda = ½.

How long will it take for the 6th person to show up? (on average)

$\lambda = 1/2$ $\boxed{\theta = 2}$

$X \sim \text{Gamma}(\alpha = 6, \theta = 2)$

$X \sim \chi^2_{12}$

$\int_{18.6}^{\infty}$

$P\{X > 18.6\}$

χ^2 Table

Let $X \sim \chi^2_{(6)}$

Find P[X < 14.45].

Ans. 0.975

Find P[X < 11]. (give a range)

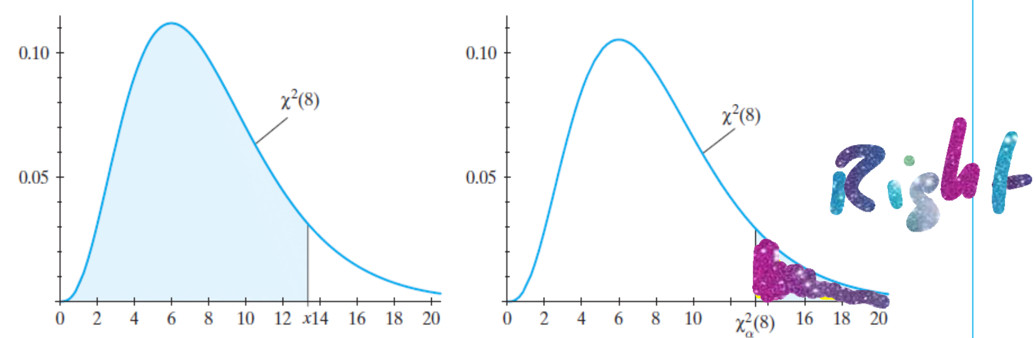
Ans. $0.9 < p < 0.95$

$P[X > 14.45]$

R

0.025

Table IV The Chi-Square Distribution



cdf

$$P(X \le x) = \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1} e^{-w/2} dw$$

Left

Right

	$P(X \le x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21

P[0.831 < X 12.83] = .95

χ^2 Table

$X \sim \chi^2_{(5)}$, find two constants, a and b, such that P[a < X < b] = 0.95.

Ans. a=0.831, b=12.83

Or

a=0, b = 11.07

Or

a = , b =

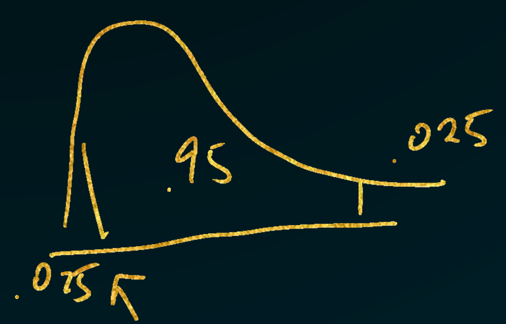
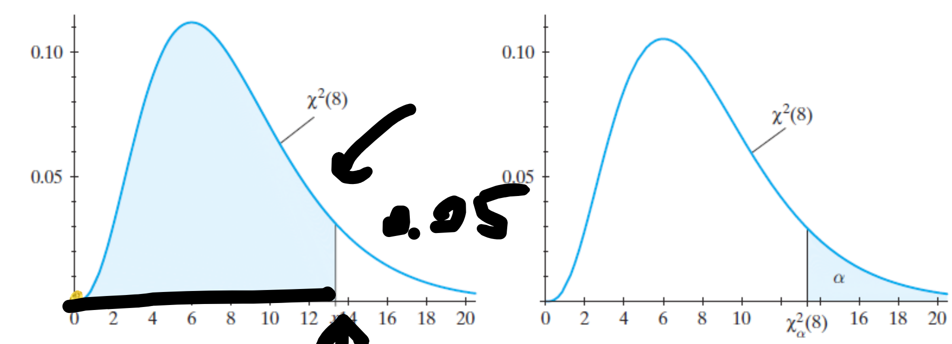


Table IV The Chi-Square Distribution



$$P(X \le x) = \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1} e^{-w/2} dw$$

R

L

	$P(X \le x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	[illegible]	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21

t distribution: When do we need it?

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

Review

If σ is known:

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

$\bar{X}$

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

If σ is unknown: Use s instead of σ.

$T = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$

sample standard dev

t distribution

Guiness

Theorem 5.5-3 **(Student's *t* distribution)** Let

$$T = \frac{Z}{\sqrt{U/r}},$$

where Z is a random variable that is $N(0,1)$, U is a random variable that is $\chi^2(r)$, and Z and U are independent. Then T has a t distribution with pdf

$$f(t) = \frac{\Gamma((r+1)/2)}{\sqrt{\pi r}\,\Gamma(r/2)} \frac{1}{(1+t^2/r)^{(r+1)/2}}, \qquad -\infty < t < \infty.$$

If interested, please refer to textbook for proof. (You are not expected to know it).

σ^2 unknown

$s^2 = 9$ $\quad n = 8$

$\bar{X} = 10$

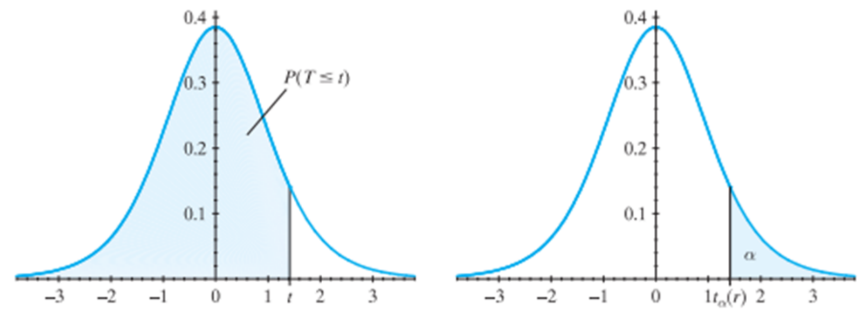
$P[\bar{X} > 10]$

$X \sim t_{8-1}$

$P[X > 1.415] = ?$

$= 0.1$

Table VI The t Distribution



$$P(T \le t) = \int_{-\infty}^{t} \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r}\,\Gamma(r/2)(1 + w^2/r)^{(r+1)/2}}\,dw$$

$$P(T \le -t) = 1 - P(T \le t)$$

	$P(T \le t)$						
	0.60	0.75	0.90	0.95	0.975	0.99	0.995
r	$t_{0.40}(r)$	$t_{0.25}(r)$	$t_{0.10}(r)$	$t_{0.05}(r)$	$t_{0.025}(r)$	$t_{0.01}(r)$	$t_{0.005}(r)$
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012
14	0.258	0.692	1.345	1.761	2.145	2.624	2.997
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947

Calculating t and χ^2 properties using R

Find probability to the left:

- *T distribution:* *pt(x, df)*
- *Chi Squared:* *pchisq(x, df)*

Finding critical value given a probability:

- *T distribution:* *qt(p, df)*
- *Chi Squared:* *qchisq(p, df)*

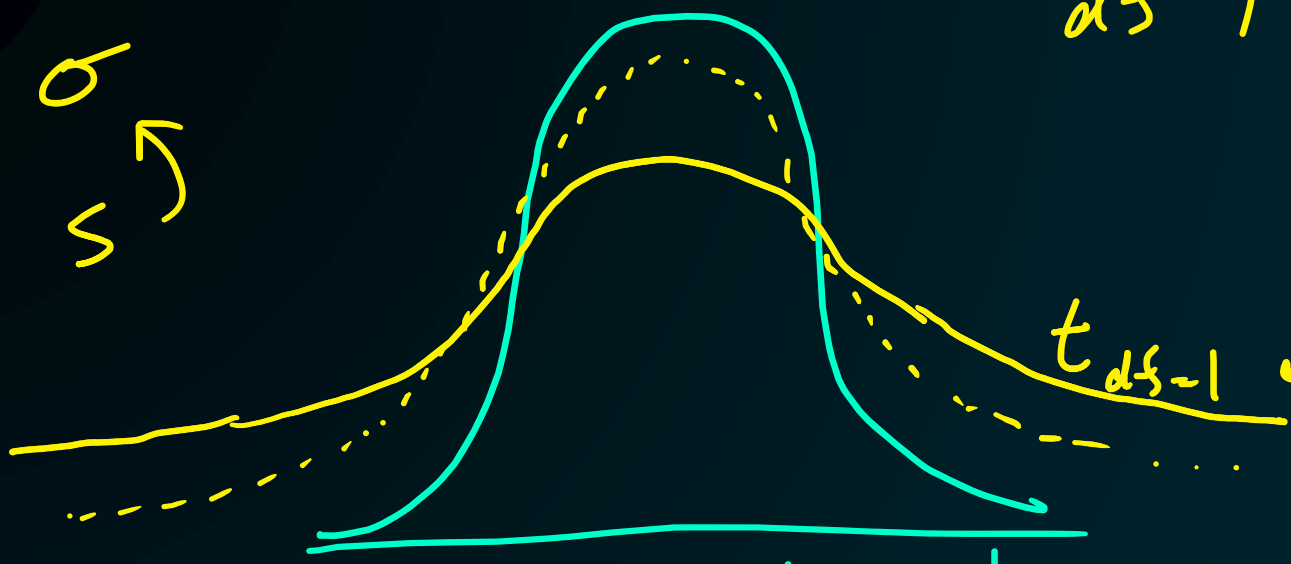
notes

σ

s

$df \uparrow$

$t \to Z$



thicker tails

STA400

- Prob
- Statistics
  - Estimation
    - Point
      - MLE
      - MOM
    - Interval (confidence Intervals)
  - Testing (Hyp)

Estimation

[Confidence Level $\rightarrow$ 95%

$\alpha = 0.05$

$100(\underbrace{1-0.05}_{.95})\%$

Confidence Intervals

A $\boxed{100(1-\alpha)\%}$ **confidence interval** is a range of numbers believed to include an unknown population parameter.

- α refers to the likelihood that the true population parameter lies outside the **confidence interval**.

90% $\alpha = 0.1$

- Its complement, $(1-\alpha)$, is called the **confidence coefficient.**
  - It is a measure of the confidence we have that the interval contains the parameter of interest.

99% $\alpha = 0.01$

$\sigma_x = \sigma$

$\sigma_{\bar{x}} = \sigma/\sqrt{n}$

Interval

95% $\alpha = 0.05$

A 100(1-α)% CI for μ when σ is known

$\bar{X} \pm$ ___

critical value

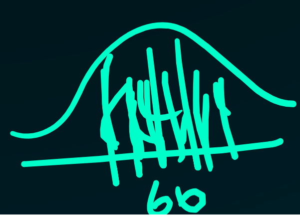
$$\left(\bar{X} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\,,\ \bar{X} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$$

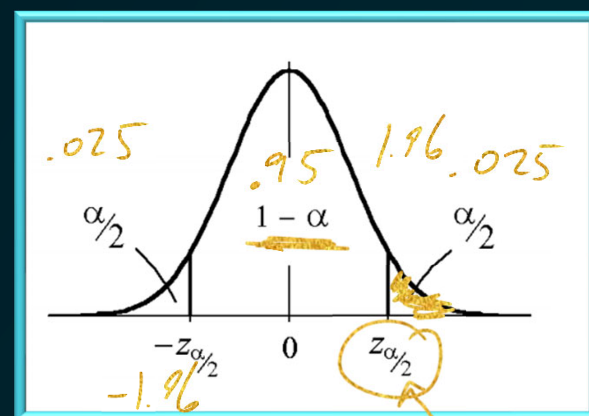
0%

$\bar{X}$ $\boxed{60}$

95%

(52, 68)





95%

$-6 < -5 < -4$

$6 > 5 > 4$

Z

A 100(1-α)% CI for μ when σ is known (derivation)

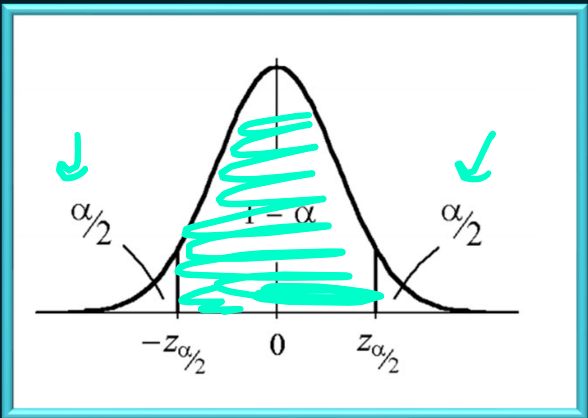
$$P\left(-z_{\alpha/2} \le \boxed{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}} \le z_{\alpha/2}\right) = 1-\alpha.$$

$$-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2},$$

$$-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le \overline{X}-\mu \le z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),$$

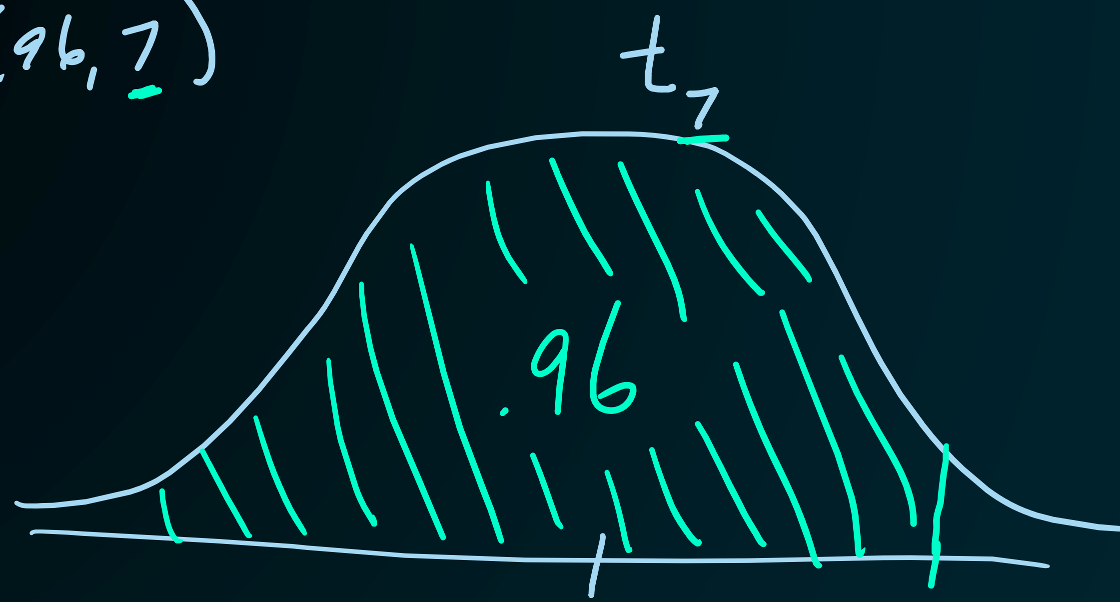
$$-\overline{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le -\mu \le -\overline{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),$$

$$\overline{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \ge \mu \ge \overline{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).$$



notes

qt(.96, 7)



qt(.96, 7)

2.046

$\sigma^2 = 16$

$n = 5$

$\sqrt{\frac{\sigma^2}{n}}$

$$\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

CI example, σ known

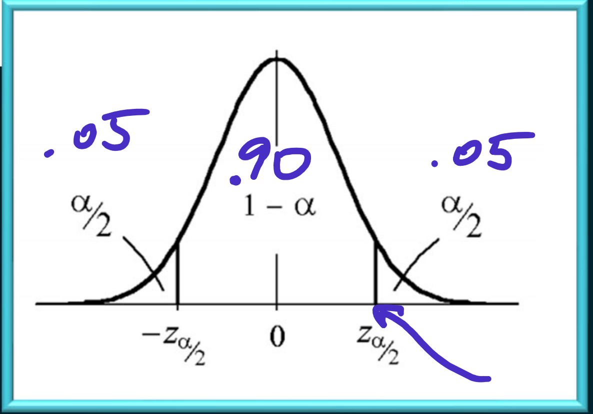
Example 7.1-2 Let $\bar{x}$ be the observed sample mean of five observations of a random sample from the normal distribution $N(\mu, 16)$. A 90% confidence interval for the unknown mean μ is

$$\left[\bar{x} - 1.645\sqrt{\frac{16}{5}},\ \bar{x} + 1.645\sqrt{\frac{16}{5}}\right].$$

$\bar{x} =$

90%　　$\alpha = 0.1$

$[\bar{x}$



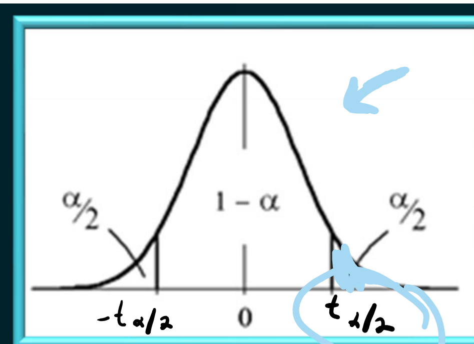
A 100(1-α)% CI for μ when σ is unknown (use s)

$$\boxed{\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}.}$$

- where the distribution, t, has $\boxed{\text{df} = n - 1}$

$qt(.96, 7)$ $n=8$ s^2 has $df=7$

Example: Confidence Interval



Assume come a $\sim N(\mu, \sigma^2)$

Given the following sample: {16, 12, 18, 13, 21, 15, 8, 17}

Construct a 92% confidence interval for the true mean.

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}.$$

$\bar{x} = 15$, $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{112}{7} = 16$, s=4

$\alpha = 0.08$, $\alpha/2 = 0.04$

df = n -1 = 7, $\boxed{t_{7,\,0.04} = 2.046}$ right tail

CI: $\mathbf{15 \pm 2.046 \cdot \frac{4}{\sqrt{8}}}$ = $\boxed{(12.107\ ,\ 17.893)}$

$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{8}}$

x	$x-\bar{x}$	$(x-\bar{x})^2$
16	1	1
12	−3	9
18	3	9
13	−2	4
21	6	36
15	0	0
8	−7	49
17	2	4
	0	112

Standard error

standard deviation

σ (of X)

SE of $\bar{X}$ $= \frac{\sigma}{\sqrt{n}}$

(SD)