# 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

# Review

- MLE
- MOM
- Bias
- Sample mean and Sample Variance

# notes

# Sample variance, $s^2$

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

Suppose we draw $n$ iid observations from a distribution with mean $\mu$ and variance $\sigma^2$.

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

Problem: $\mu$ is usually unknown!

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

Unfortunately, this tends to underestimate $\sigma^2$.

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

# 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 df = 4

$\bar{x}$ = 25

don't overthink this :)

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

# 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.

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

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

using $\bar{x} = 25$, $\underbrace{10, 20, 30}_{\text{free}}$ $\underset{\text{not free}}{40}$ only $n-1$ are free

# notes

# Chi-squared distribution, $\chi^2$ (overview)

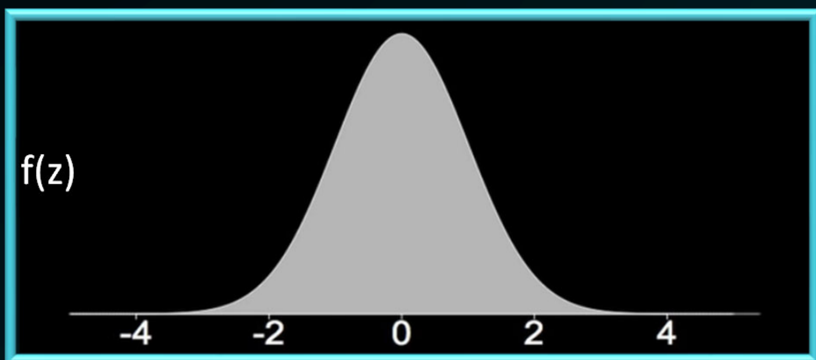
The Chi-squared distribution is an important distribution that is frequently used in statistical inference.

Comes from summing the squares of Standard Normal RVs.

- If Z ~ 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 $Z_1^2 + Z_2^2 + \ldots + Z_k^2 \sim \chi^2_{(k)}$

$Z$

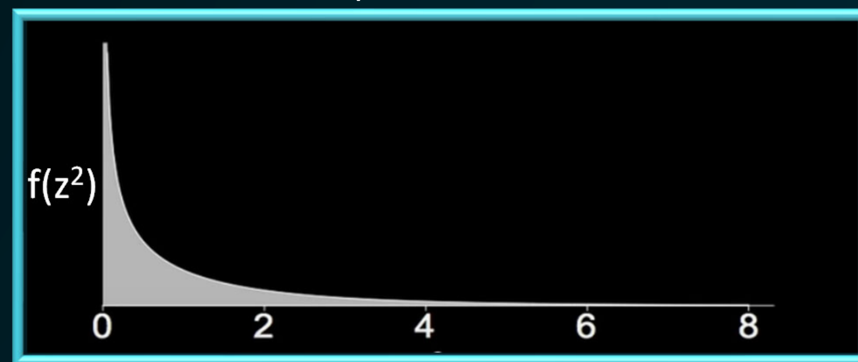
Standard Normal




$z$

$\chi^2_{(1)}$

Chi-Squared, df = 1




$z^2$

# Chi-squared distribution $X \sim \chi^2_r$

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

We can see from this pdf that X also follows another distribution! X ∼ $(\alpha = \quad , \theta = \quad )$

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

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

# notes

# 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 $\sim \chi^2_{2\alpha}$

Example:

X $\sim \chi^2_r$

$\sim$ Gamma($\alpha$ =r/2, $\theta$ =2)

# notes

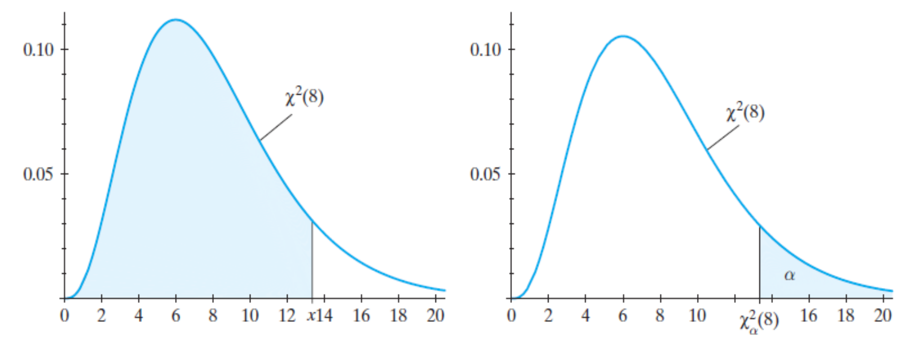
# $\chi^2$ Table

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

Find P[X < 14.45].
*Ans. 0.975*

Find P[X < 11]. (give a range)
*Ans.  0.9 < p < 0.95*

**Table IV** The Chi-Square Distribution




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

| | $P(X \le x)$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 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 |

# $\chi^2$ Table

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

*Ans. a=0.831, b=12.83*

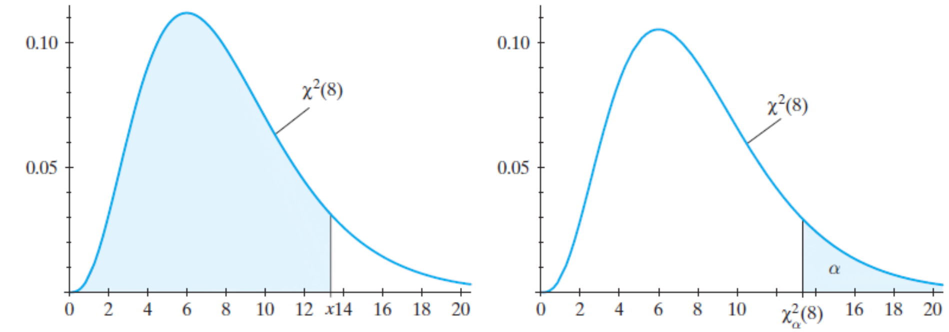
*Or*

*a=0, b = 11.07*

*Or*

a = , b =

**Table IV** The Chi-Square Distribution




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

| | $P(X \le x)$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 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 |

# t distribution: When do we need it?

If $\sigma$ is known:

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

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

If $\sigma$ is unknown: Use $\boldsymbol{s}$ instead of $\sigma$

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

# t distribution

**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).

# Calculating t and $\chi^2$ 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

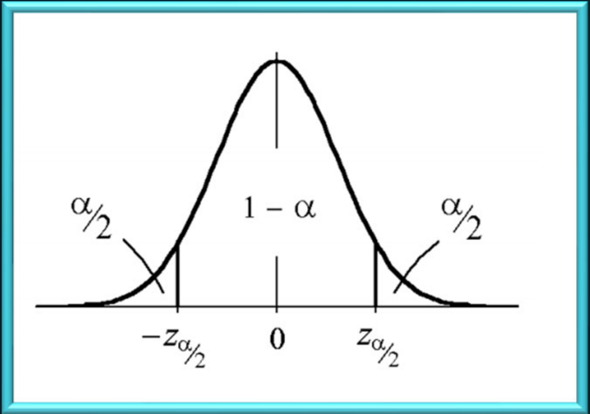
# Confidence Intervals

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

- $\alpha$ refers to the likelihood that the true population parameter lies outside the **confidence interval**.
- 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.

## A 100(1-$\alpha$)% CI for $\mu$ when $\sigma$ is known

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



## A 100(1-$\alpha$)% CI for $\mu$ when $\sigma$ is known (derivation)

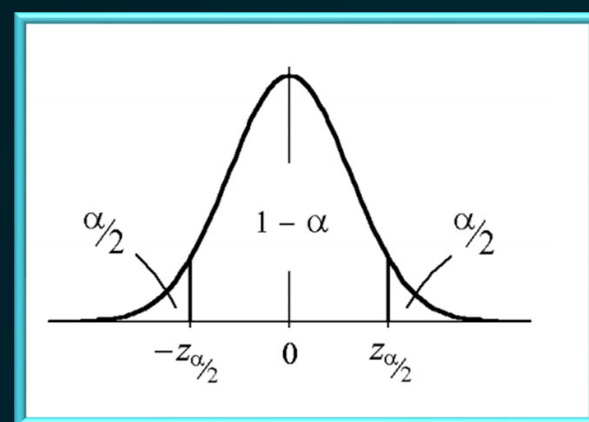
$$P\left(-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha.$$

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

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

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

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



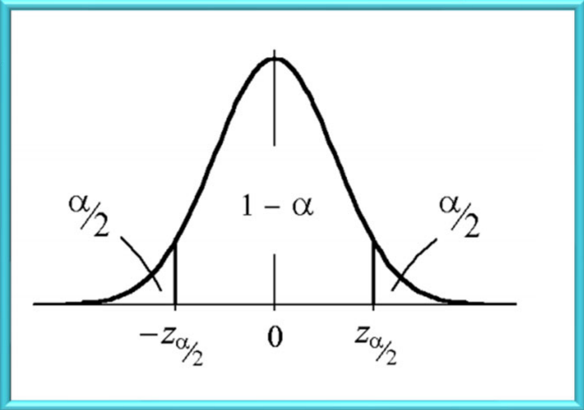
# notes

# CI example, $\sigma$ known

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

**Example 7.1-2** Let $\overline{x}$ 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 $\mu$ is

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

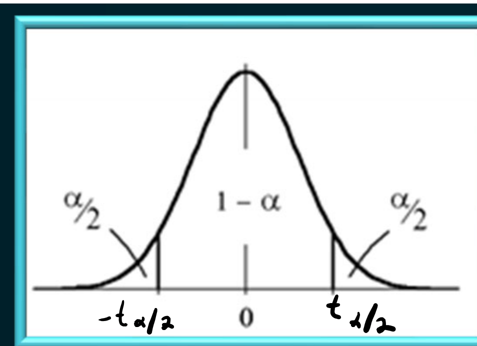


A 100(1-$\alpha$)% CI for $\mu$ when $\sigma$ is unknown (use s)

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

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

# Example: Confidence Interval




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

Construct a 92% confidence interval for the true mean.

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

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

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

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

| $x$ | $x-\bar{x}$ | $(x-\bar{x})^2$ |
|---|---|---|
| 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 |

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