Chapter 8

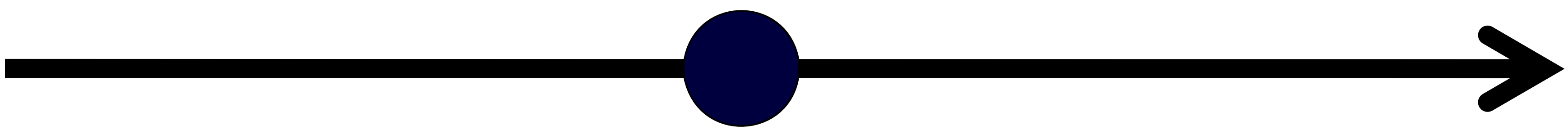
Statistical Inference : Confidence Intervals (One Population)

Overview

- Confidence Intervals for the Proportion
- Confidence Intervals for the Mean
- Sample Size Determination

Point Estimate vs. Interval Estimate

- A **point estimate** is a ***single number*** that is our “best guess” for the parameter

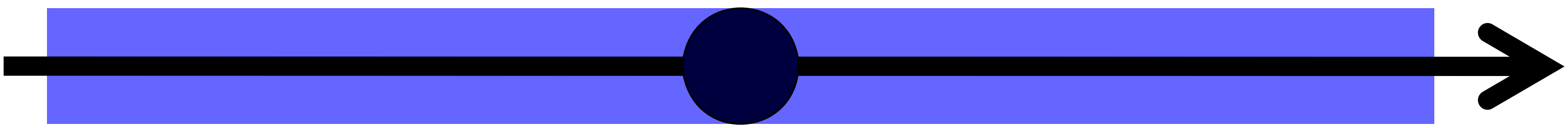


- A *point estimate* alone is not sufficiently informative.

  It doesn’t tell us how close the estimate is likely to be to the parameter.

Point Estimate vs. Interval Estimate

- An **interval estimate** is an ***interval of numbers*** within which the parameter value is believed to fall.



- An *interval estimate* is more useful.

  It incorporates a **margin of error** which helps us to gauge the accuracy of the point estimate.

Point Estimator

- The best point estimate of the population proportion *p* is the sample proportion, $\hat{p}$.

- The best point estimate of the population mean *μ* is the sample mean $\bar{Y}$.

Properties of a Good Estimator

1. The estimator should be an **unbiased estimator**.
   E(estimator) = parameter

2. The estimator should be a **relatively efficient estimator**;
   that is, of all the statistics that can be used to estimate a parameter, the relatively efficient estimator has the smallest variance.

Interval Estimate

- An **interval estimate** of a parameter is an interval or a range of values used to estimate the parameter.
- The interval is constructed around the point estimate.

Interval = point estimate ± a margin of error

- This estimate may or may not contain the value of the parameter being estimated.

Confidence Level of the Interval Estimate

- The **confidence level** of an interval estimate of a parameter is the probability that the interval estimate will contain the parameter, assuming that a large number of samples are selected and that the estimation process on the same parameter is repeated.
- Common confidence levels are 90%, **95%** and 99%.

Confidence Level of the Interval Estimate

- Written as $(1-\alpha)100\%$
  where α is the **error probability**.

-

Confidence Level, $(1-\alpha)$	Error Probability, α
0.99	0.01
0.95	0.05
0.90	0.10

- For 95% confidence*:*
  - *With probability 0.95*, a sample statistic value occurs such that the confidence interval *contains* the population parameter.
  - *With probability 0.05*, the method produces a confidence interval that *misses* the parameter.

Confidence Interval (CI)

- A **confidence interval** is
  a specific interval estimate of a parameter determined by using data obtained from a sample and by using the specific confidence level of the estimate.

  $(1-\alpha)100\%$ CI for p
  $(1-\alpha)100\%$ CI for μ

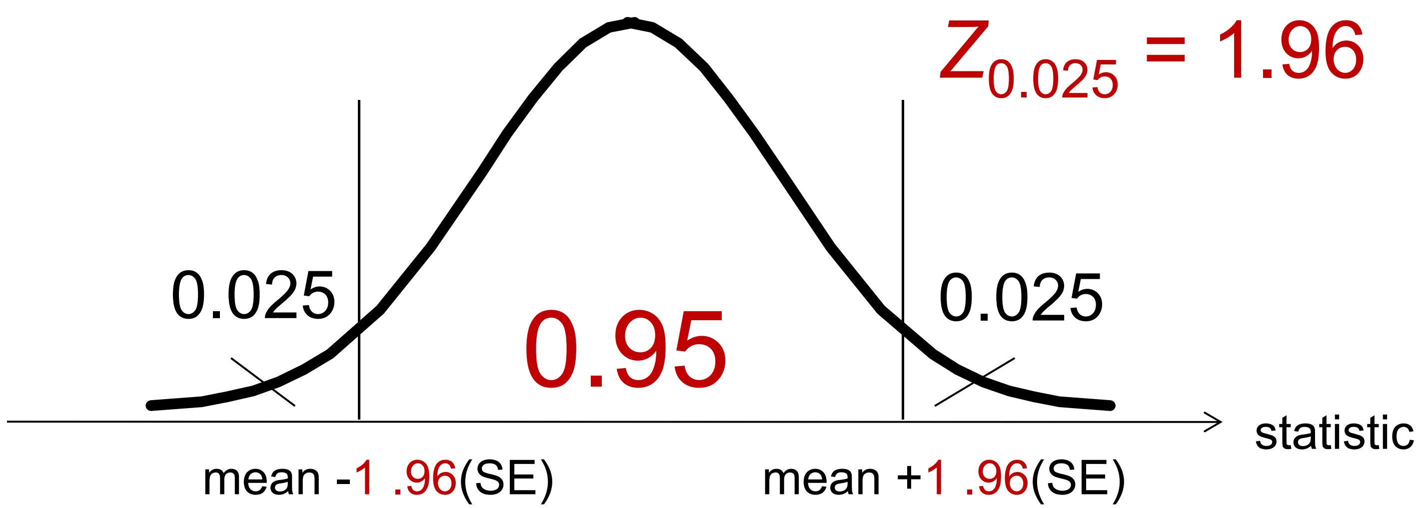
- What is the logic behind constructing a confidence interval?

Recall

The Sampling Distribution

- Gives the possible values for the sample statistics and their probabilities.
- Has a **mean** equal to the population parameter.
- Has a standard deviation called the **standard error (SE)**.
- Is approximately a **normal** distribution for large random samples.

The Sampling Distribution



- Fact: Approximately 95% of a normal distribution falls within 1.96 (or 2) standard deviations of the mean
- That means: With probability 0.95, the sample statistic falls within about 1.96 standard errors of the population parameter.

Confidence Interval

- A confidence interval is constructed by adding and subtracting a **margin of error** from a given point estimate

$$(1-\alpha)100\%\ \mathrm{CI} = \text{point estimate} \pm \text{ margin of error}$$

$$= \text{point estimate} \pm Z_{\alpha/2}(\mathrm{SE})$$

$$95\%\ \mathrm{CI} = \text{point estimate} \pm 1.96(\mathrm{SE})$$

when $np \geq 15$ and $nq \geq 15$

- A 95% confidence interval has margin of error equal to 1.96 standard errors

For a 90% confidence interval: $z_{\alpha/2} = 1.65$

For a 95% confidence interval: $z_{\alpha/2} = 1.96$

For a 99% confidence interval: $z_{\alpha/2} = 2.58$

Confidence Intervals for Proportions

Let p = population proportion

$\hat{p}$ = sample proportion $= \dfrac{X}{n}$ $\qquad \hat{q} = 1 - \hat{p}$

X = number of sample units that possess the characteristics of interest and n = sample size.

when $np \geq 15$ and $nq \geq 15$,

$$(1-\alpha)100\% \text{ CI for } p = \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

$$= \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Example: Air Conditioned Households

In a recent survey of 150 households, 54 had central air conditioning. What is the point estimate of the proportion of households that have central air conditioning?

Since X = *54 and* n = *150,*

Point estimate of *p* = $\hat{p} = \frac{X}{n} = \frac{54}{150} = 0.36 = 36\%$

$$\hat{q} = 1 - \hat{p} = 1 - 0.36 = 0.64 = 64\%$$

Example: Religious Books

A survey of 1721 people found that 15.9% of individuals purchase religious books at a Christian bookstore.

Find the 95% confidence interval of the true proportion of people who purchase their religious books at a Christian bookstore.

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

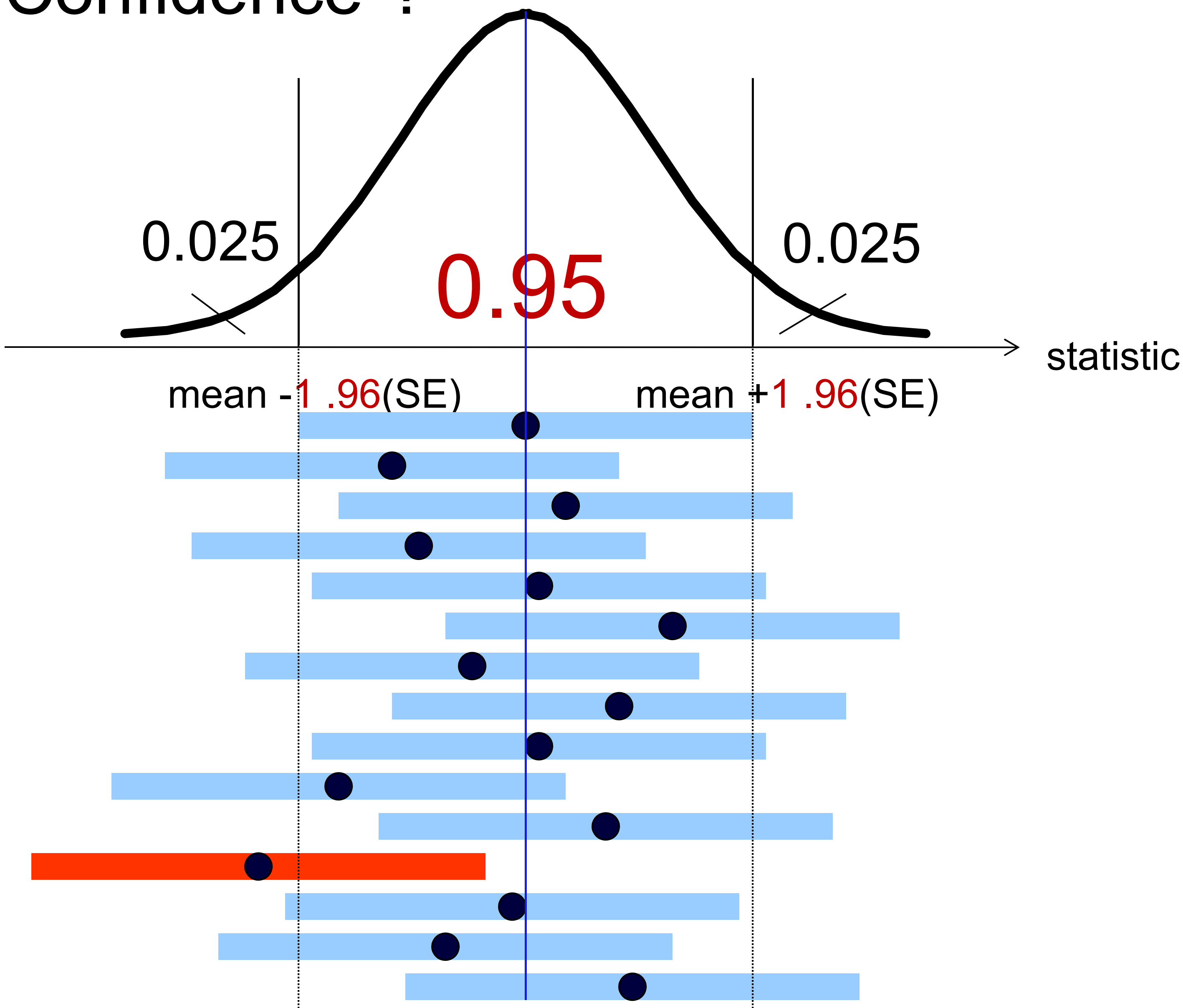
$$0.159 - 1.96\sqrt{\frac{(0.159)(0.841)}{1721}} < p < 0.159 + 1.96\sqrt{\frac{(0.159)(0.841)}{1721}}$$

$$0.142 < p < 0.176$$

95% CI for *p* = (0.142, 0.176)

with **95% confidence,** the true percentage is between 14.2% and 17.6%.

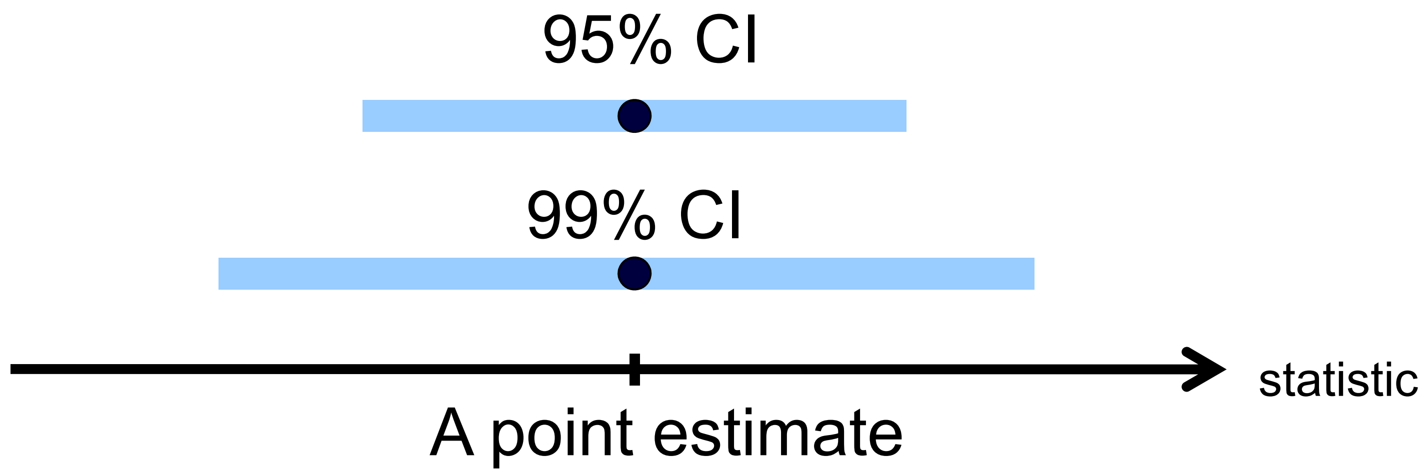
What Does It Mean to Say that We Have "95% Confidence"?



What Does It Mean to Say that We Have "95% Confidence"?

- If we used the 95% confidence interval method to estimate many population proportions, then *in the long run about 95% of those intervals would give correct results, containing the population proportion*

Different Confidence levels



- In using confidence intervals, we must compromise between the desired margin of error and the desired confidence of a correct inference
  - As the desired confidence level increases, the margin of error gets larger
- Is it possible to have a high confidence level and a small margin of error?

Example: Male Nurses

A sample of 500 nursing applications included 60 from men. Find the 90% confidence interval of the true proportion of men who applied to the nursing program.

For 90% CL → $\alpha/2 = 0.05$ → $z_{.05} = 1.65$

$$\hat{p} = \frac{X}{n} = \frac{60}{300} = 0.12 \quad , \quad \hat{q} = 0.88$$

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.12 - 1.65\sqrt{\frac{(0.12)(0.88)}{500}} < p < 0.12 + 1.65\sqrt{\frac{(0.12)(0.88)}{500}}$$

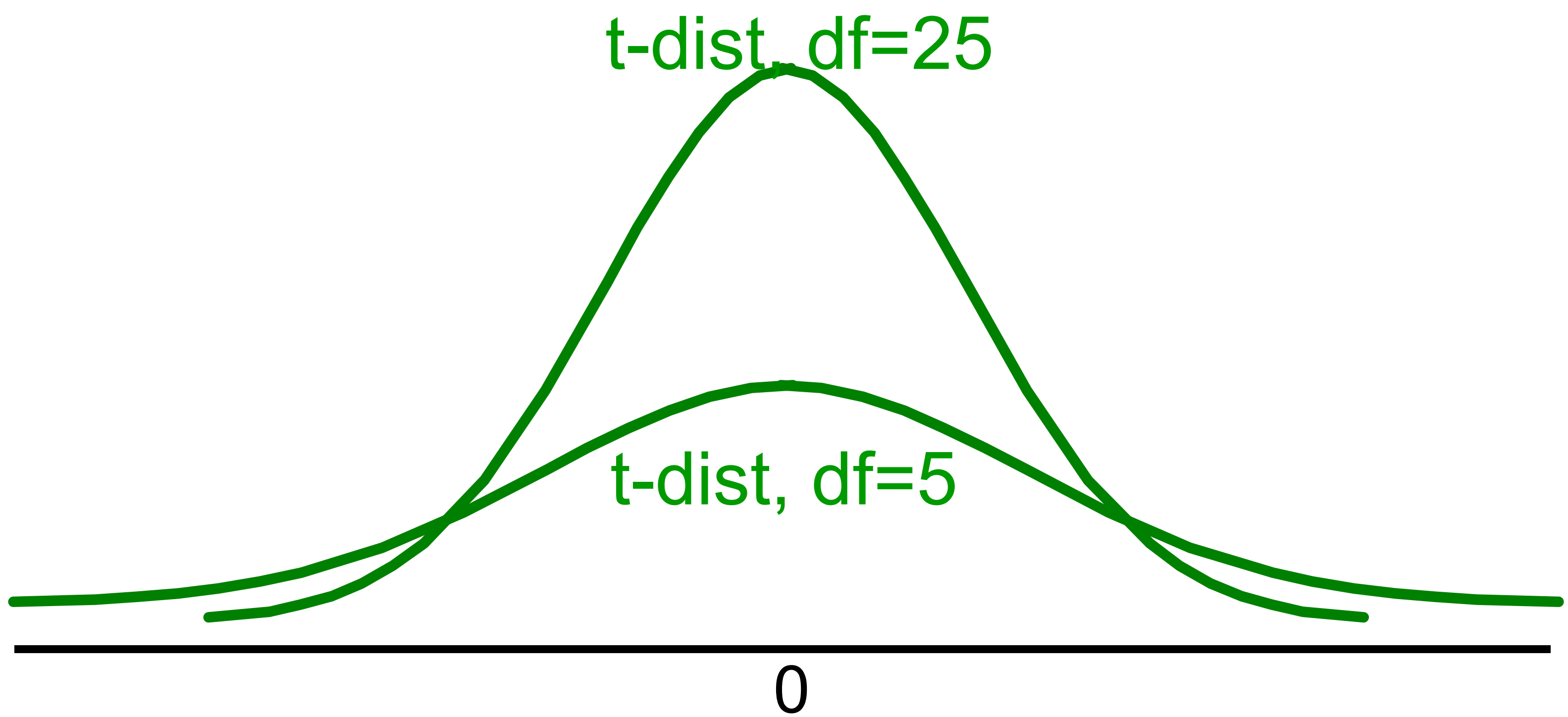
$$0.12 - 0.024 < p < 0.12 + 0.024$$

$$.096 < p < 0.144$$

We can be 90% confident that the percentage of male applicants is between 9.6% and 14.4%.

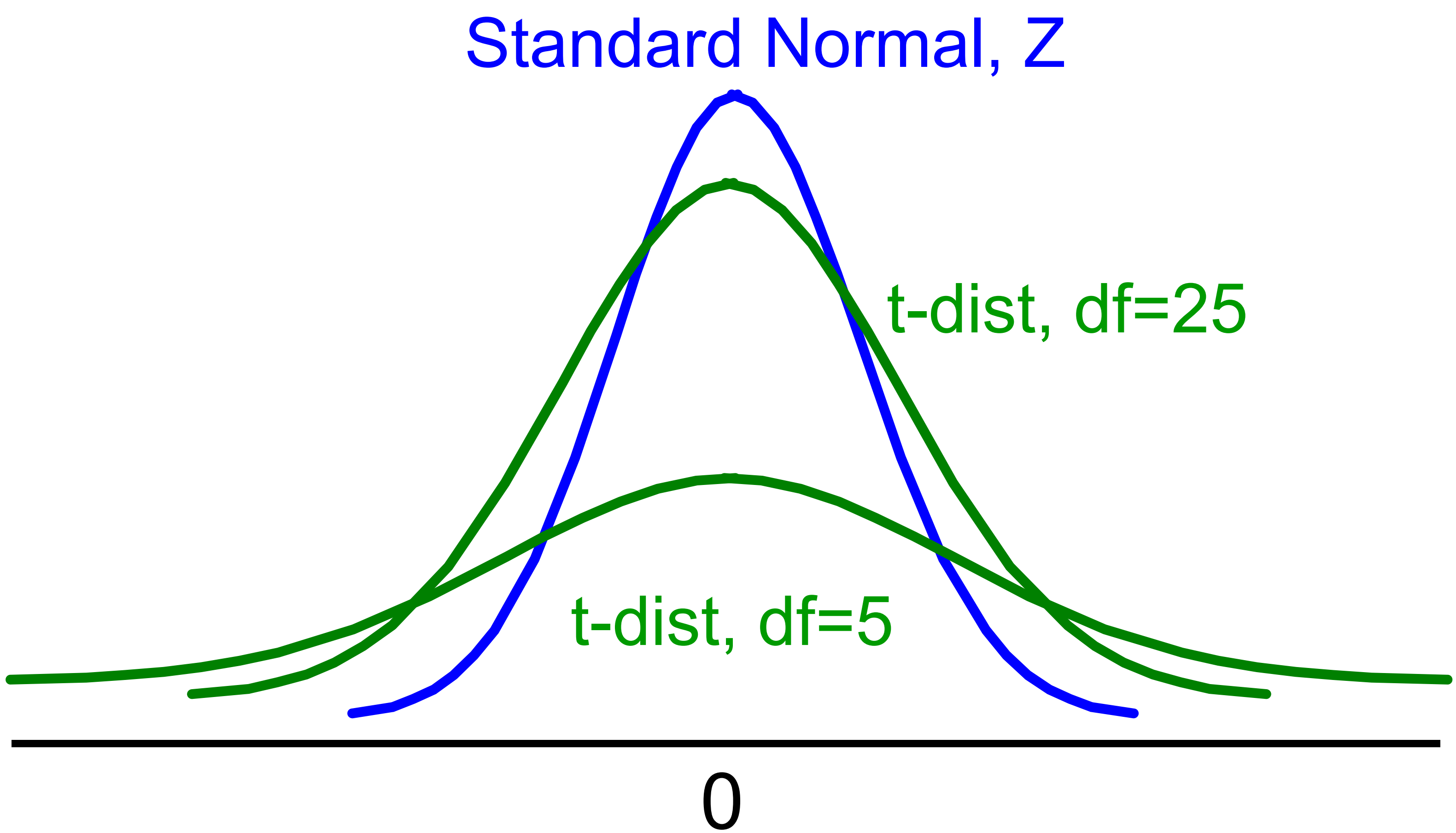
Characteristics of the *t* Distribution

1. It is bell-shaped (unimodal, symmetric)
2. The mean is 0.
3. The variance is greater than 1.
4. The *t* distribution is actually a family of curves based on the concept of degrees of freedom (*d.f.* = *n* – 1).



Characteristics of the *t* Distribution

5. As the sample size increases, the *t* distribution approaches the standard normal distribution.



Confidence Intervals for Means

$$(1-\alpha)100\%\ \text{CI} = \text{point estimate} \pm\ \text{margin of error}$$

$$(1-\alpha)100\%\ \text{CI for } \mu = \overline{X} \pm Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

σ is unknown!

- For large *n* (*n* ≥30)
- For small *n* from a normal population

Confidence Interval for the Mean

$$(1-\alpha)100\% \text{ CI for } \mu = \overline{X} \pm Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$(1-\alpha)100\% \text{ CI for } \mu = \overline{X} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$\overline{X} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \overline{X} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

The degrees of freedom are *n - 1.*

- An approximately normal population

Example: Using The *t* distribution

Find the $t_{\alpha/2}$ value for a 95% confidence interval when the sample size is 22.

Degrees of freedom are d.f. = 21.

$\alpha =$	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
$\nu = 1$	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
⋮				⋮			
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725

Example: Sleeping Time

Ten randomly selected people were asked how long they slept at night. The mean time was 7.1 hours, and the standard deviation was 0.78 hour. Find the 95% confidence interval of the mean time. Assume the variable is normally distributed.

Since σ is unknown and *s* must replace it, the *t* distribution must be used for the confidence interval. Hence, with 9 degrees of freedom, $t_{\alpha/2} = 2.262$.

$$\bar{X} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{X} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$7.1 - 2.262\left(\frac{0.78}{\sqrt{10}}\right) < \mu < 7.1 + 2.262\left(\frac{0.78}{\sqrt{10}}\right)$$

Example: Sleeping Time

$$7.1 - 2.262\left(\frac{0.78}{\sqrt{10}}\right) < \mu < 7.1 + 2.262\left(\frac{0.78}{\sqrt{10}}\right)$$

$$7.1 - 0.56 < \mu < 7.1 + 0.56$$

$$6.5 < \mu < 7.7$$

One can be 95% confident that the population mean is between 6.5 and 7.7 hours.

Example: Home Fires by Candles

The data represent a sample of the number of home fires started by candles for the past several years. Find the 99% confidence interval for the mean number of home fires started by candles each year.

5460 5900 6090 6310 7160 8440 9930

Step 1: Find the mean and standard deviation.

The mean is $\bar{X}$ = 7041.4 and the standard deviation s = 1610.3.

Step 2: Find $t_{\alpha/2}$. The confidence level is 99%, and the degrees of freedom d.f. = 6

$t_{.005}$ = 3.707.

Example: Home Fires by Candles

Step 3: Substitute in the formula.

$$\overline{X} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \overline{X} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$7041.4 - 3.707\left(\frac{1610.3}{\sqrt{7}}\right) < \mu < 7041.4 - 3.707\left(\frac{1610.3}{\sqrt{7}}\right)$$

$$7041.4 - 2256.2 < \mu < 7041.4 + 2256.2$$

$$4785.2 < \mu < 9297.6$$

One can be 99% confident that the population mean number of home fires started by candles each year is between 4785.2 and 9297.6, based on a sample of home fires occurring over a period of 7 years.

What If the Population is Not Normal?

- A basic assumption of the confidence interval using the *t*-distribution is that the population distribution is normal.
- Many variables have distributions that are far from normal.
- How problematic is it if we use the *t*- confidence interval even if the population distribution is not normal?
- For large random samples, it's not problematic.

  The Central Limit Theorem applies: for large n, the sampling distribution is bell-shaped even when the population is not.

What If the Population is Not Normal?

- What about a confidence interval using the *t*-distribution when n is small?
  - Even if the population distribution is not normal, confidence intervals using *t*-scores usually work quite well
- We say the *t*-distribution is a *robust method* in terms of the normality assumption

Cases Where the *t*- Confidence Interval Does Not Work

- With binary data
- With data that contain extreme outliers

The *t*-Distribution with df = ∞

$\alpha =$	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
$\nu = 1$	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
⋮							
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

The 2002 GSS asked: “What do you think is the ideal number of children in a family?”

The 497 females who responded had a median of 2, mean of 3.0, and standard deviation of 1.8.

1. What is the population mean?

2. What is the point estimate of the population mean?

   a. 497
   b. 2
   c. 3.0
   d. 1.8
   e. Unknown

3. What is the 95% confidence interval of the population mean?

Sample Size Determination

How large a sample is necessary to make an accurate interval estimate?

It depends on

1. The population variance.
2. The confidence level.
3. The maximum error (how close the estimator is to the parameter).

$$(1-\alpha)100\%\ \text{CI for } p = \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$ **Error**

$$(1-\alpha)100\%\ \text{CI for } \mu = \overline{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$ **Error**

Sample Size Determination for Proportion Estimation

The margin of error is $E = Z_{\alpha/2}\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$

Thus, the formula for *minimum* sample size needed for interval estimate of a population proportion is

$$n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2$$

If there is no information on the proportion, use **0.5** as the estimate.

$$n = \left(\frac{1}{4}\right)\left(\frac{Z_{\alpha/2}}{E}\right)^2$$

If necessary, round up to the next whole number.

Example: Home Computers

A researcher wishes to estimate, with 95% confidence, the proportion of people who own a home computer. A previous study shows that 40% of those interviewed had a computer at home. The researcher wishes to be accurate within 2% of the true proportion. Find the minimum sample size necessary.

$$n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2 = (0.40)(0.60)\left(\frac{1.96}{0.02}\right)^2 = 2304.96$$

The researcher should interview a sample of at least 2305 people.

Example: Car Phone Ownership

The same researcher wishes to estimate the proportion of executives who own a car phone. She wants to be 90% confident and be accurate within 5% of the true proportion. Find the minimum sample size necessary.

Since there is no prior knowledge of $\hat{p}$, statisticians assign the values $\hat{p}$ = 0.5 and $\hat{q}$ = 0.5. The sample size obtained by using these values will be large enough to ensure the specified degree of confidence.

$$n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2 = (0.50)(0.50)\left(\frac{1.65}{0.05}\right)^2 = 272.25$$

The researcher should ask at least 273 executives.

Example: Exit Poll

A television network plans to predict the outcome of an election between two candidates – A and B. They will do this with an exit poll that randomly samples votes on election day.

The final poll a week before election day estimated Mr. A to be well ahead, 58% to 42%. So the outcome is not expected to be close.

The researchers decide to use a sample size for which the margin of error is 0.04

What is the sample size, *n* for which a 95% confidence interval for the population proportion has margin of error equal to 0.04?

Example: Exit Poll

$$n = \hat{p}(1-\hat{p})\left(\frac{1.96}{m}\right)^2$$

$$= (0.58)(0.42)\left(\frac{1.96}{0.04}\right)^2$$

$$= 584.9$$

A random sample of size *n = 585* should give a margin of error of about 0.04 for a 95% confidence interval for the population proportion.

Sample Size determination for Mean Estimation

The margin of error is $E = Z_{\alpha/2}\left(\dfrac{\sigma}{\sqrt{n}}\right)$

Thus, the formula for *minimum* sample size needed for interval estimate of a population mean is

$$n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2$$

If there is no information on the *σ*, it can be estimated using *s* or the range rule of thumb.

If necessary, round up to the next whole number.

Example: Depth of a River

A scientist wishes to estimate the average depth of a river. He wants to be 99% confident that the estimate is accurate within 2 feet. From a previous study, the standard deviation of the depths measured was 4.38 feet.

$$99\% \rightarrow z = 2.58, E = 2, \sigma = 4.38$$

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left(\frac{2.58 \cdot 4.38}{2}\right)^2 = 31.92 = 32$$

Therefore, to be 99% confident that the estimate is within 2 feet of the true mean depth, the scientist needs at least a sample of 32 measurements.

Example: Mean Education

A social scientist plans a study of adult South Africans to investigate educational attainment in the black community.

How large a sample size is needed so that a 95% confidence interval for the mean number of years of education has margin of error equal to 1 year? (You may assume the distribution is bell-shaped.)

No prior information about the standard deviation of educational attainment is available. We might guess that the sample education values fall within a range of about 18 years.

s = range / 6 = 18 / 6 = 3

So 3 is a crude estimate of s

Example: Mean Education

The desired margin of error is E = 1 year

The required sample size is:

$$n = \left(\frac{Z_{\alpha/2} s}{E} \right)^2 = \left(\frac{1.96 \bullet 3}{1} \right)^2 = 34.6$$