Point and Interval Estimates of Population parameters

Point Estimate and Interval Estimate

- A *point estimate* is a *single number* that is our "best guess" for the parameter.
- An *interval estimate* is an *interval of numbers* within which the parameter value is believed to fall.
- A *point estimate* does not tell us how close the estimate is likely to be to the parameter.
- An *interval estimate* is more useful. It incorporates a margin of error which helps us to gauge the accuracy of the point estimate.

Example. A General Social Survey asked to a sample of 1900 persons: "do you believe in hell?". 1387 responded yes. The point estimate for the proportion of adult Americans who would respond yes equals $1387/1900 = 0.73$ (point estimate). The interval estimate is [0.71,0.75], it predicts a *margin of error* of 0.02.

Properties of Point Estimators

For any particular parameter, there are several possible point estimates. For a normal distribution, for instance, the center is the mean and the median since that distribution is symmetric. So, two possible estimates of that center value are the sample mean and the sample median. What makes a particular estimate better than others? A *good estimator* of a parameter has two desirable properties:

Property 1: A good estimator has a sampling distribution that is centered at the parameter, i.e. the mean of the distribution equals the true value of the parameter. An estimator with this property is **unbiased**.

- The sample mean is an unbiased estimator of the population mean.
- The sample proportion is an unbiased estimator of the population proportion.

Example

Population (1 = speak a second language)

X	f_k
0	0.80
1	0.20

Sample Space for n=2

X_1, X_2	$p(X_1, X_2)$	$\hat{p}$
0,0	0.8×0.8	0
0,1	0.8×0.2	0.5
1,0	0.2×0.8	0.5
1,1	0.2×0.2	1

We note that $\hat{p}$ and X_1 are both unbiased, which is the best?

$\hat{p}$	p	$\hat{p} \times p(\hat{p})$
0	0.64	0
0.5	0.320	0.16
1	0.04	0.04
Tot.	1	$0.20 = \mathrm{E}(\hat{p})$

X_1	$p(X_1)$	$X_1 \times p(X_1)$
0	0.8	0
1	0.2	0.2
Tot.	1	$0.20 = \mathrm{E}(X_1)$

Standard Error

- Error of estimation = the difference between the statistic (estimate) and the true value of the parameter.
- Standard error (se) = the standard absolute error of estimation of the estimator (square root of squared errors mean).
- It is important to note that when the mean of the estimator sampling distribution is equal to the value of the parameter, then the standard error coincides with the standard deviation. This is so because the deviations from the mean coincide with the errors of estimation.

Property 2: A good estimator has a *small standard error* compared to other estimators.

- This means it tends to fall closer than other estimates to the parameter.
- The sample mean has a smaller standard error than the sample median when estimating the population mean of a normal distribution.

Example (cont'd) in pink: variances of p_hat and X1

$\hat{p}$	$p(\hat{p})$	$p(\hat{p}) \times (\hat{p} - 0.2)^2$
0	0.64	$0.64 \times 0.04 = 0.0256$
0.5	0.32	$0.32 \times 0.09 = 0.0288$
1	0.04	$0.04 \times 0.64 = 0.0256$
	1.00	0.0800

X_1	$p(X_1)$	$X_1 \times p(X_1)$	$p(X_1) \times (X_1 - 0.2)^2$
0	0.80	0	$0.8 \times 0.04 = 0.032$
1	0.20	0.20	$0.2 \times 0.64 = 0.128$
			0.160

$\hat{p}$ is better than X_1 because it has a smaller variance $\Rightarrow$ smaller standard deviation $\Rightarrow$ smaller standard error.

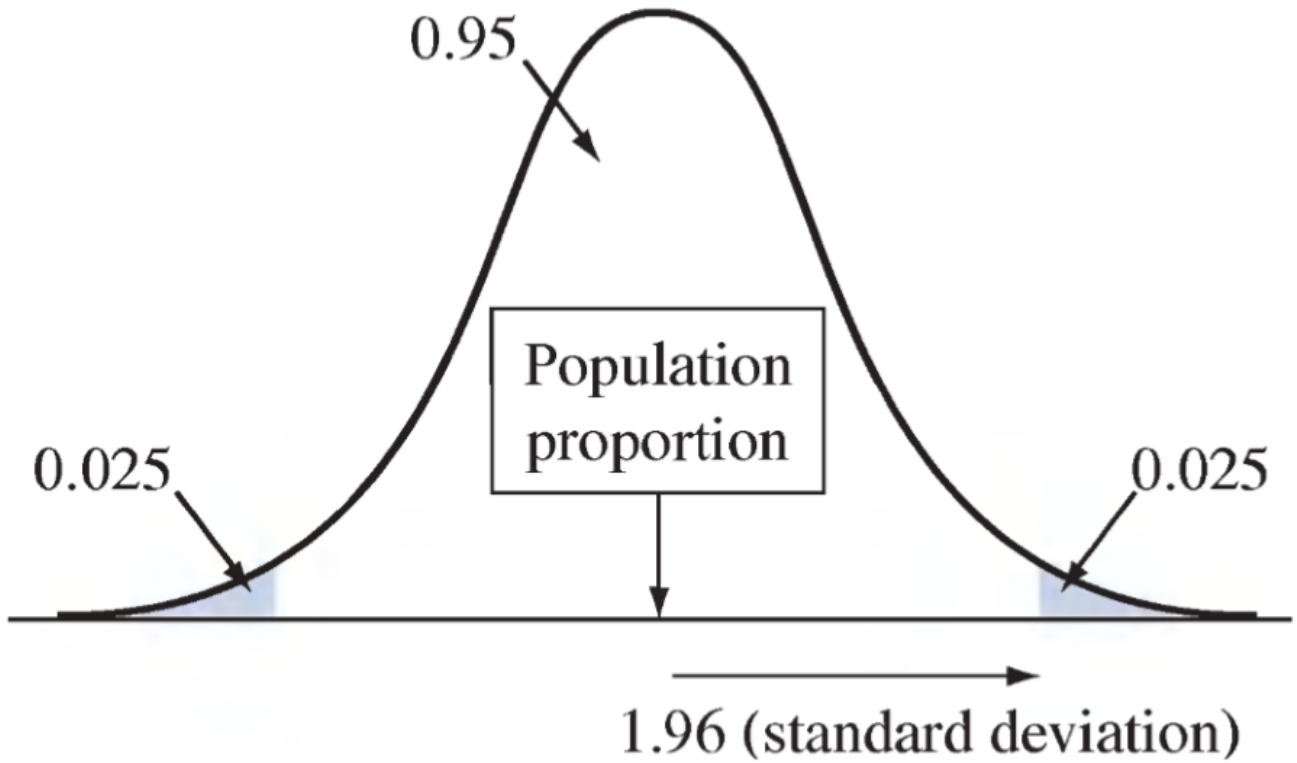
Confidence Intervals

- A *confidence interval* is an interval containing the most believable values for a parameter
- The probability that this method produces an interval that contains the parameter is called the *confidence level*
  - This is a number chosen to be close to 1, most commonly 0.95

Logic of Confidence Intervals

- To construct a confidence interval for a population proportion, start with the *sampling distribution of a sample proportion*
  - gives the possible values for the sample proportion and their probabilities
  - is approximately a normal distribution for large random samples by the CLT $np >= 15$, $n(1-p) >= 15$
  - has mean equal to the population proportion

- Fact: Approximately 95% of a normal distribution falls within 1.96 standard deviations of the mean
  - With probability 0.95, the sample proportion falls within about 1.96 standard errors of the population proportion
  - The distance of 1.96 standard errors is the *margin of error* in calculating a 95% confidence interval for the population proportion



Example (cont'd)

Population (1 = speak a second language)

X	f_k
0	0.80
1	0.20

If the sample size is 400 then the sampling distribution is well approximated by a normal with mean $p = 0.2$ and s.d. $\sqrt{(p(1-p)/n)} = \sqrt{(0.2 \times 0.8/400)} = 0.02$. It follows that

sample proportion +- 1.96(standard deviation)

$$P \quad (0.2 - 1.96 \times 0.02 < \hat{p} < 0.2 + 1.96 \times 0.02) = 0.95$$

$$P \quad (-1.96 \times 0.02 < \hat{p} - 0.2 < 1.96 \times 0.02) = 0.95$$

$$P \quad (-0.0392 < \hat{p} - 0.2 < 0.0392) = 0.95$$

The margin of error is about 4%

Margin of Error

- The *margin of error* measures how accurate the point estimate is likely to be in estimating a parameter.
- It is a multiple of the standard error of the sampling distribution of the estimate when the sampling distribution is a normal distribution.
- The distance of 1.96 standard errors is the margin of error for a 95% confidence interval for a parameter from a normal distribution.

Summary

A confidence interval is constructed by taking a point estimate and adding and subtracting a margin of error. The margin of error is based on the standard deviation of the sampling distribution of that estimator. When the sampling distribution is approximately normal, a 95% confidence interval has a margin of error equal to 1.96 standard deviations.

Constructing a Confidence Interval to Estimate a Population Proportion

Finding the 95% Confidence Interval for a Population Proportion

- We symbolize a population proportion by p
- The point estimate of the population proportion is the *sample proportion*
- We symbolize the sample proportion by $\hat{p}$
- A 95% confidence interval uses a margin of error = 1.96(standard errors)
- CI = [point estimate $\pm$ margin of error] =

$$\hat{p} \pm 1.96(\text{standard errors})$$

for a 95% confidence interval

- The exact standard error of a sample proportion equals:

$$se = \sqrt{\frac{p(1-p)}{n}}$$

- This formula depends on the unknown population proportion, p
- In practice, we do not know p, and we need to estimate the standard error as

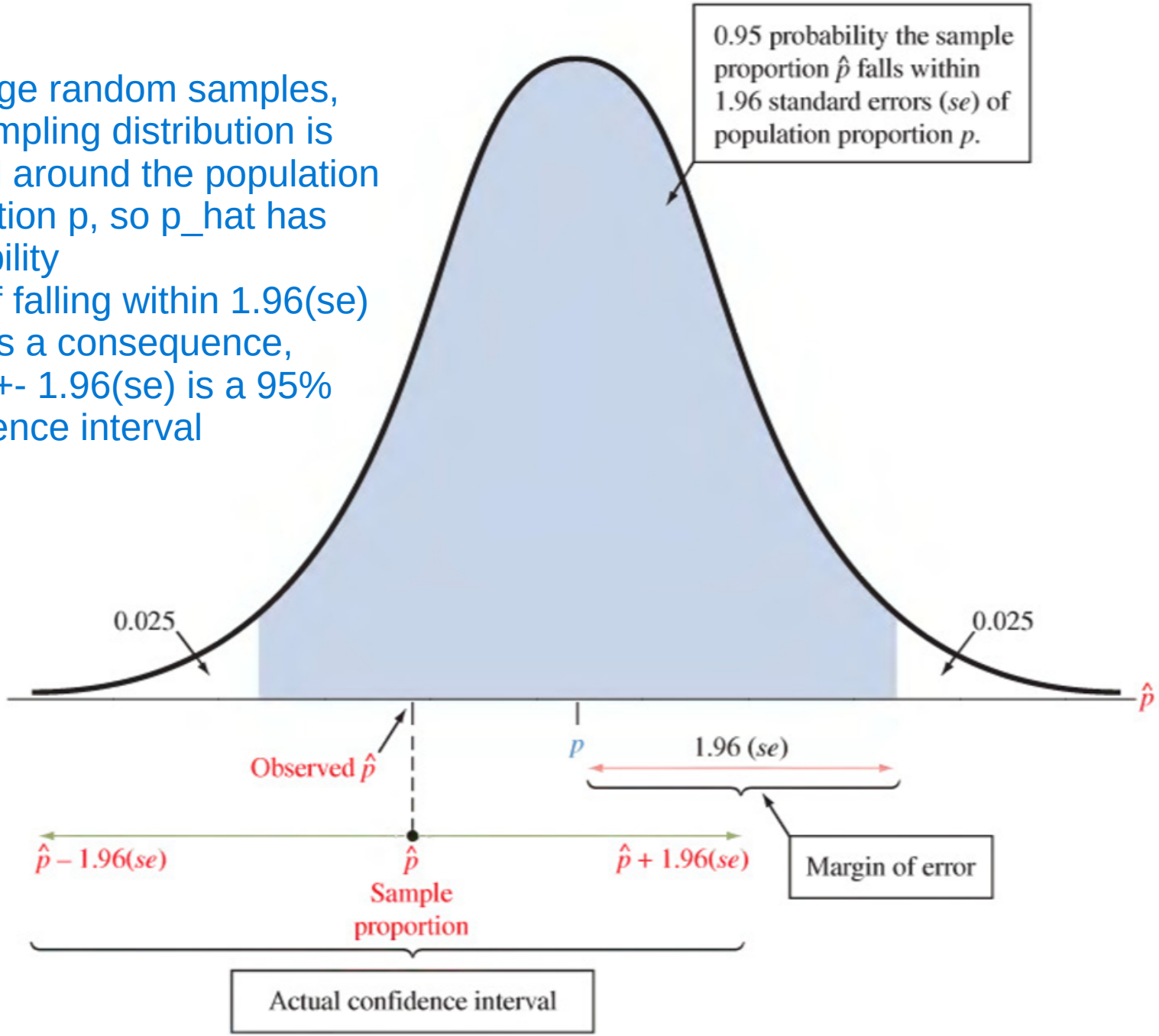
$$se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- A 95% confidence interval for a population proportion p is:

$$\hat{p} \pm 1.96(se), \text{ with } se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

A standard error is an estimated standard deviation of a sampling distribution

For large random samples, the sampling distribution is normal around the population proportion p, so p_hat has probability 0.95 of falling within 1.96(se) of p. As a consequence, p_hat +- 1.96(se) is a 95% confidence interval for p



Example. A survey asked to a sample of 1900 persons: "do you believe in hell?"; 1387 responded yes. The point estimate for the proportion of adult Americans who would respond yes equals 0.73. The margin of error (with a probability of 95%) is $1.96 \times se = 1.96 \times 0.444/\sqrt{(1900)} = 0.02$. The interval estimate is [0.71,0.75].

Sample Size Needed for Large-Sample Confidence Interval for a Proportion The sampling distribution of the sample proportion is well approximated by a normal if both np and n(1-p) are greater than or equal to 15. In this case we do not know the true value of p but only its point estimate. It follows that the rule becomes

For the 95% confidence interval for a proportion p to be valid, you should have at least 15 successes and 15 failures:

$$n\hat{p} \geq 15 \text{ and } n(1-\hat{p}) \geq 15$$

How Can We Use Confidence Levels Other than 95%?

- "*95% confidence*" means that there is a 95% chance that a sample proportion value occurs such that the confidence interval contains the unknown value of the population proportion, p
- *With probability* 0.05, the method produces a confidence interval that misses p
- In practice, the confidence level 0.95 is the most common choice
- But, some applications require greater (or less) confidence
- To increase the chance of a correct inference, we use a larger confidence level, such as 0.99.
- With probability 0.99, the sample proportion falls within about 2.5758 standard errors of the population proportion. It follows that the 99% CI will be

$$\hat{p} \pm 2.5758(se), \text{ with } se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- 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

Example. Exit poll: Out of 1400 voters, 660 voted for the Democratic candidate. The CIs are

- 95%

$$0.471 \pm 1.96 \times 0.499/\sqrt{(1400)}$$
$$0.471 \pm 0.026$$
$$0.445 < p < 0.497$$

- 99%

$$0.471 \pm 2.5758 \times 0.499/\sqrt{(1400)}$$
$$0.471 \pm 0.034$$
$$0.437 < p < 0.505$$

Example. A recent GSS asked "If the wife in a family wants children, but the husband decides that he does not want any children, is it all right for the husband to refuse to have children?" Of 598 respondents, 366 said yes. The CIs are

- 95%

$$0.6121 \pm 1.96 \times 0.487/\sqrt{(598)}$$
$$0.612 \pm 0.039$$
$$0.573 < p < 0.651$$

- 99%

$$0.6121 \pm 2.5758 \times 0.487/\sqrt{(598)}$$
$$0.612 \pm 0.051$$
$$0.561 < p < 0.663$$

Summary

p_hat = x / n

The CI for the proportion is

$$\hat{p} \pm z(se), \text{ with } \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z depends on the confidence level:

- confidence level 95% $\Rightarrow z = 1.96$
- confidence level 99% $\Rightarrow z = 2.5758$

The sample size n should be large enough so that the number of success, $n\hat{p}$, and the number of failures, $n(1-\hat{p})$, are both at least 15.

Effect of the Sample Size

- The *margin of error* for a confidence interval:
  - Increases as the confidence level increases
  - Decreases as the sample size increases

Interpretation of the Confidence Level

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

Example (referendum). During a referendum, in a sample of $n = 100$ exit polls, we found a proportion of 0.55 YES. What can we conclude? First of all, we calculate confidence intervals for the proportion

- 95%: $0.55 \pm 1.96 \times 0.497/\sqrt{(100)} = 0.55 \pm 0.097$
  $\Rightarrow 0.453 < p < 0.647$
- 99%: $0.55 \pm 2.5758 \times 0.497/\sqrt{(100)} = 0.55 \pm 0.128$
  $\Rightarrow 0.422 < p < 0.678$

We observe that for p are plausible values of less than 0.5, so we can not yet conclude anything about the outcome of the referendum.

Now suppose to have $n = 1000$ exit polls that confirm an estimate of 0.55 for the YES. What can we conclude? We observe that the confidence intervals now are

- 95%: $0.55 \pm 1.96 \times 0.497/\sqrt{(1000)} = 0.55 \pm 0.031$
  $\Rightarrow 0.519 < p < 0.581$
- 99%: $0.55 \pm 2.5758 \times 0.497/\sqrt{(1000)} = 0.55 \pm 0.040$
  $\Rightarrow 0.510 < p < 0.590$

and we can therefore conclude in favor of the victory of the YES.

Confidence Interval to Estimate a Population Mean

- CI: Point estimate $\pm$ margin of error
- The sample mean is the point estimate of the population mean
- The exact standard error of the sample mean is $\sigma/\sqrt{n}$
- In practice, we do not know the population standard deviation σ, we estimate it by the sample standard deviation, s
- For large n from any population and also for every n from an underlying population that is normal, the confidence interval for the population mean is

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

in theory: sample mean +- the z-score of the standard error

Example. We want to make inference about the mean of the weight distribution of 10 year-old children. To this purpose, we draw a sample of 144 children and compute the sum of the weights, that is 5760 kg, and the sum of the squared deviations from the mean, that is 3575. How do we estimate the population mean?

- Mean point estimate: $\bar{x} = 5760/144 = 40$.
- Variance point estimate: $s^2 = 3575/143 = 25$.
- Interval estimate 95%: $40 \pm 1.96 \times 5/\sqrt{(144)} = 40 \pm 0.817$ $\Rightarrow 39.183 < \mu < 40.817$
- Interval estimate 99%: $40 \pm 2.5758 \times 5/\sqrt{(144)} = 40 \pm 1.073$ $\Rightarrow 38.927 < \mu < 41.073$

Confidence Interval for a Population Mean when n is small and the population is normal

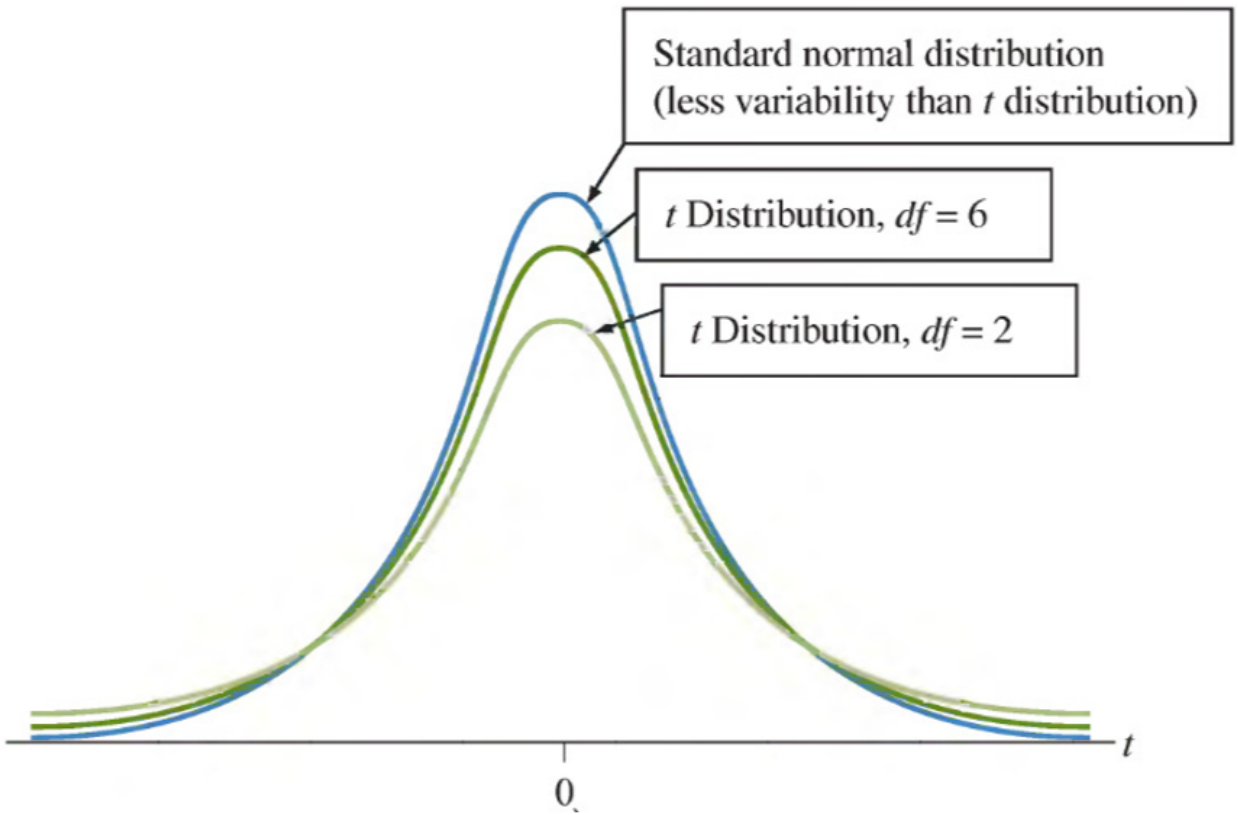
- Substituting the sample standard deviation s for to get s.e. $= s/\sqrt{n}$ introduces extra error
- To account for this increased error, we replace the z-score by a slightly larger score, the t-score

$$\bar{x} \pm t_{\mathrm{df},0.025}\left(\frac{s}{\sqrt{n}}\right);\ \mathrm{df} = n-1$$

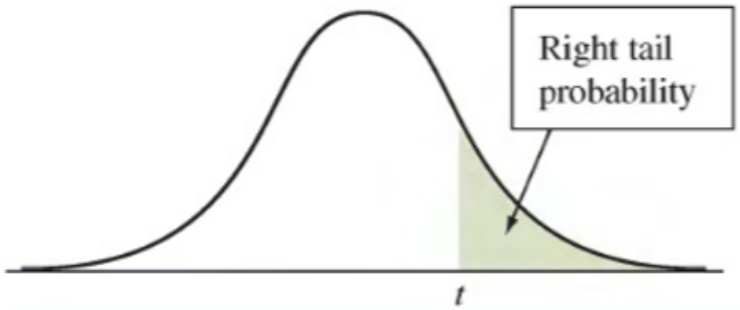
- To use this method, you need:
  - Data obtained by randomization (random sample)
  - An approximately normal population distribution

Properties of the t Distribution

- The t-distribution is bell shaped and symmetric about 0 (=mean)
- The probabilities depend on the degrees of freedom, $df = n - 1$
- The t-distribution has thicker tails than the standard normal distribution, i.e., it is more spread out



How Do We Find a t Confidence Interval for Other Confidence Levels?



For example, t_ .025 has probability 0.025 in the right tail, a two-tail probability of 0.05, and is used in 95% confidence intervals

REM: left_tail = - right_tail

Table B *t* **Distribution Critical Values**

	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208

Example. A study of 7 American adults from an SRS yields an average height of 67.2 inches and a standard deviation of 3.9 inches. Assuming the heights are normally distributed, 95% and 99% confidence intervals for the average height of all American adults (μ) are:

- 95%: $67.2 \pm 2.447 \times 3.9/\sqrt{(7)} = 67.2 \pm 3.607$
  $63.593 < \mu < 70.807$

  We are 95% confident that the average height of all American adults is between 63.6 and 70.8 inches.

- 99%: $67.2 \pm 3.707 \times 3.9/\sqrt{(7)} = 67.2 \pm 5.465$
  $61.735 < \mu < 72.665$

  We are 99% confident that the average height of all American adults is between 61.7 and 72.7 inches.

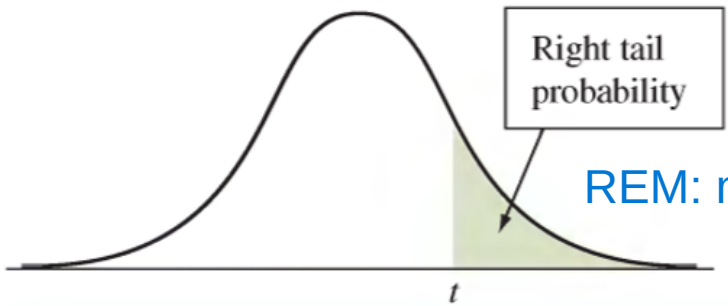
If the Population is Not Normal, is the Method "Robust"?

A statistical method is said to be robust with respect to a particular assumption if it performs adequately even when that assumption is modestly violated

- 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
- We say the t-distribution is a *robust method* in terms of the normality assumption
- How problematic is it if we use the t-confidence interval even if the population distribution is not normal?
- For large random samples, it is not problematic because of the Central Limit Theorem, ensuring a normal distribution of the sample mean
- What if n is small?
- Confidence intervals using t-scores usually work quite well except for when extreme outliers are present. The method is robust.

The Std Normal Distribution is the t Distribution with df $= \infty$

You can think of the standard normal distribution as a t distribution with df = infinity



REM: margin of error -> standard error -> sample size

Table B *t* **Distribution Critical Values**

			•••			
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.299	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
80	1.292	1.664	1.990	2.374	2.639	3.195
100	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.091

see section 8.4 of book for extras and page 392 for summary