RWTH BUSINESS SCHOOL

Mathematics & Statistics
M.Sc. Data Analytics and Decision Science

Prof. Dr. Thomas S. Lontzek

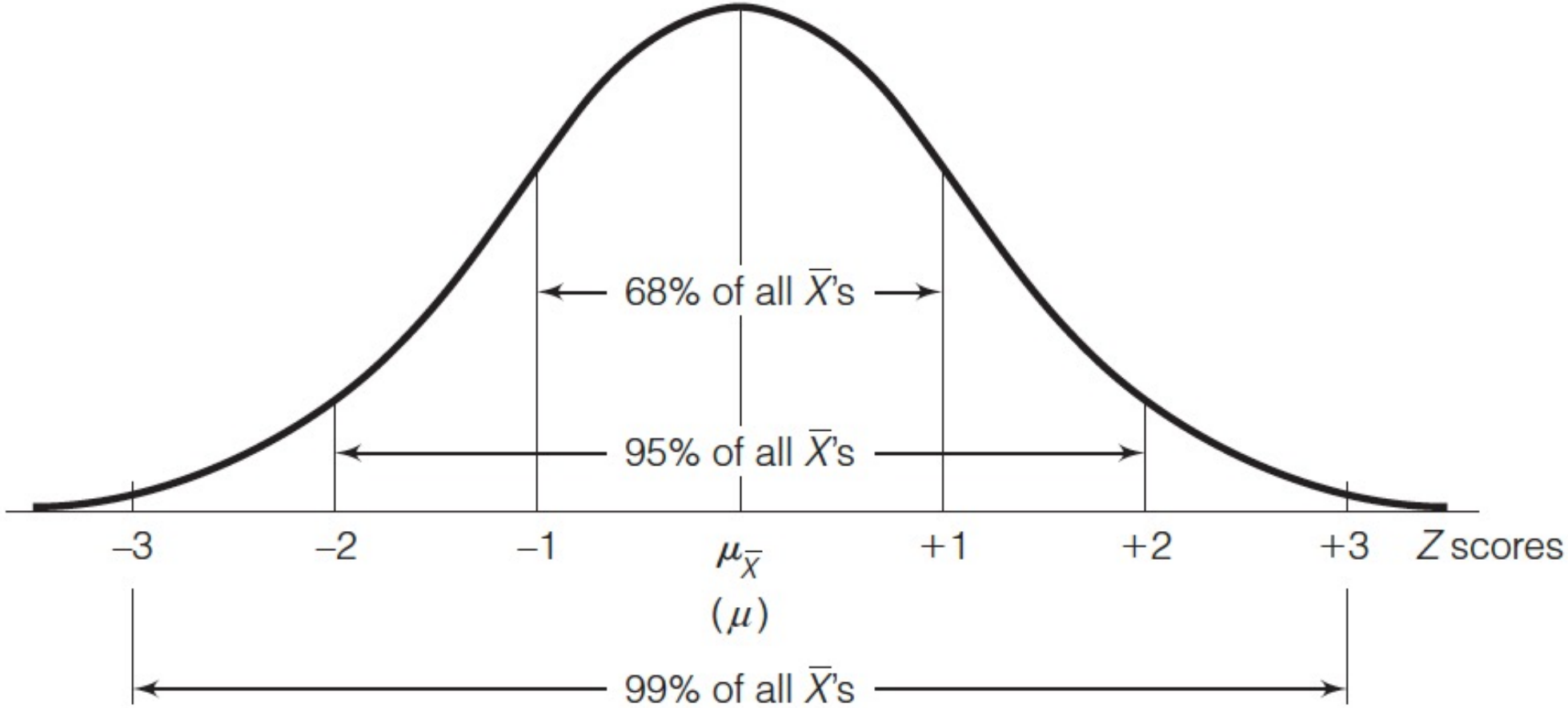
AACSB ACCREDITED

BUSINESS SCHOOL | RWTH AACHEN UNIVERSITY

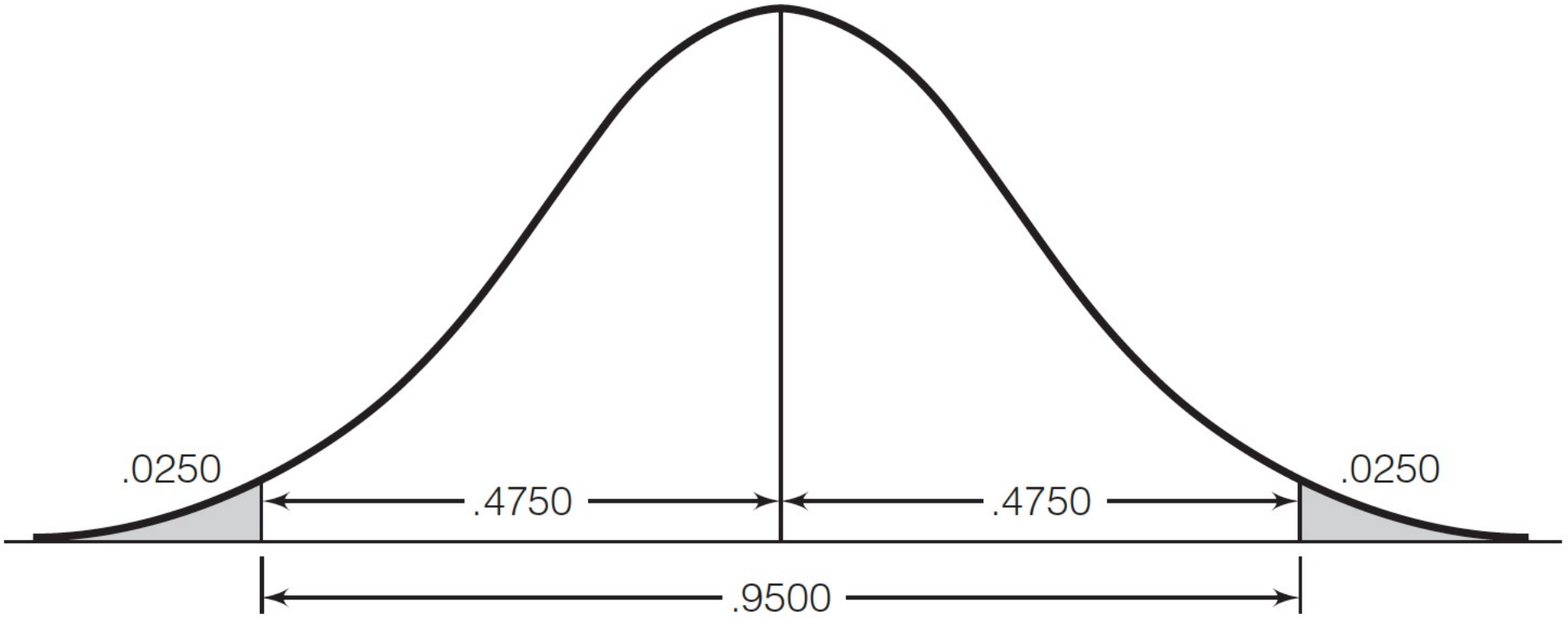
Central Limit Theorem

- One of the main goal in statistics is to quantifying how much confidence we can have in population estimates
- The CLT is a very simple concept
- E.g. uniform distribution
- If we show a histogram of the means of n samples from a population, the means will be normally distributed as n increases. (mostly safe if sample size >30)
- Practical implication: In experiment we often do not know from what distribution the data comes from. The CLT tells us not to worry: The sample means will be normally distributed

Areas Under The Sampling Distribution of Sample Means



The Sampling Distribution With Alpha of 5%



Z Scores of Various Levels of Alpha

Confidence Level	Alpha (α)	$\alpha/2$	Z Score
90%	0.100	0.0500	±1.65
95%	0.050	0.0250	±1.96
99%	0.010	0.0050	±2.58
99.9%	0.001	0.0005	±3.32
99.99%	0.0001	0.00005	±3.90

Standard Deviation and Standard Error

Standard deviation quantifies how much the data are spread out around the mean e.g. within a sample

The standard error (of the mean) measures how far the sample mean is from the population mean.

The standard error is the standard deviation of the sampling distribution

$$\text{standard deviation } \sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

$$\text{variance} = \sigma^2$$

$$\text{standard error } (\sigma_{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$$

where:

$\bar{x} =$ the sample's mean

$n =$ the sample size

Choosing Formulas for Confidence Intervals

If the Sample Statistic Is a	and	Use Formula
mean	the population standard deviation is known	$\text{c.i.} = \overline{X} \pm Z\left(\frac{\sigma}{\sqrt{N}}\right)$
mean	the population standard deviation is unknown	$\text{c.i.} = \overline{X} \pm Z\left(\frac{s}{\sqrt{N-1}}\right)$
proportion		$\text{c.i.} = P_s \pm Z\sqrt{\frac{P_u(1-P_u)}{N}}$

Estimating a Sample Mean (Population Standard Deviation Unknown)

A study of the leisure activities of Americans was conducted on a sample of 1000 households. The respondents identified television viewing as a major form of recreation. If the sample reported an average of 6.2 hours of television viewing a day, what is the estimate of the population mean? The information from the sample is

$$\overline{X} = 6.2$$
$$s = 0.7$$
$$N = 1000$$

If we set alpha at 0.05, the corresponding Z score will be ± 1.96, and the 95% confidence interval will be

$$\text{c.i.} = \overline{X} \pm Z\left(\frac{s}{\sqrt{N-1}}\right)$$

$$\text{c.i.} = 6.2 \pm 1.96\left(\frac{0.7}{\sqrt{1000-1}}\right)$$

$$\text{c.i.} = 6.2 \pm 1.96\left(\frac{0.7}{31.61}\right)$$

$$\text{c.i.} = 6.2 \pm 1.96(0.02)$$

$$\text{c.i.} = 6.2 \pm 0.04$$

Based on this result, we would estimate that the population spends an average of 6.2 ± 0.04 hours per day viewing television. The lower limit of our interval estimate $(6.2 - 0.04)$ is 6.16, and the upper limit $(6.2 + 0.04)$ is 6.24. Thus, another way to state the interval would be

$$6.16 \leq \mu \geq 6.24$$

The population mean is greater than or equal to 6.16 and less than or equal to 6.24. Because alpha was set at the 0.05 level, this estimate has a 5% chance of being wrong (that is, of not containing the population mean).

Estimating a Sample Mean (Known Population Standard Deviation)

Suppose you wanted to estimate the average IQ of a community and had randomly selected a sample of 200 residents, with a sample mean IQ of 105. Assume that the population standard deviation for IQ scores is about 15. If we are willing to run a 5% chance of being wrong and set alpha at 0.05, the corresponding Z score will be 1.96. These values can be used to construct a confidence interval: interval

$$\text{c.i.} = \overline{X} \pm Z\left(\frac{\sigma}{\sqrt{N}}\right)$$

$$\text{c.i.} = 105 \pm 1.96\left(\frac{15}{\sqrt{200}}\right)$$

$$\text{c.i.} = 105 \pm 1.96\left(\frac{15}{14.14}\right)$$

$$\text{c.i.} = 105 \pm (1.96)(1.06)$$

$$\text{c.i.} = 105 \pm 2.08$$

Our estimate is that the average IQ for the population in question is somewhere between 102.92 and 107.08. Since 95% of all possible sample means are within 1.96 Z ’s (or 2.08 IQ units in this case) of the mean of the sampling distribution, the odds are very high that our interval will contain the population mean.

Estimating a Population Proportion

If 45% of a random sample of 1000 Americans reports that walking is their major physical activity, what is the estimate of the population value? The sample information is

$$P_s = 0.45$$
$$N = 1000$$

Note that the percentage of "walkers" has been stated as a proportion. If we set alpha at 0.05, the corresponding Z score will be ± 1.96, and the interval estimate of the population proportion will be

$$\text{c.i.} = P_s \pm Z\sqrt{\frac{P_u(1 - P_u)}{N}}$$

$$\text{c.i.} = 0.45 \pm 1.96\sqrt{\frac{(0.5)(0.5)}{1000}}$$

Most conservative choice

$$\text{c.i.} = 0.45 \pm 1.96\sqrt{0.00025}$$
$$\text{c.i.} = 0.45 \pm (1.96)(0.016)$$
$$\text{c.i.} = 0.45 \pm 0.03$$

We estimate that the proportion of the population for which walking is the major form of physical exercise is between 0.42 and 0.48. That is, the lower limit of the interval estimate is $(0.45 - 0.03)$ or 0.42, and the upper limit is $(0.45 + 0.03)$ or 0.48. We may also express this result in percentages: Between 42% and 48% of the population walk as their major form of physical exercise. This interval has a 5% chance of not containing the population value.

Application: Polling of US Presidential Elections – "Too Close to Call"

Date/Poll	Margin of Error	Obama	McCain
11-3-2008, CSPAN	±3%	54%	43%
11-2-2008, Gallup	±2%	53%	42%
11-2-2008, Fox News	±3%	50%	43%
11-1-2008, NBC News	±3%	51%	43%
11-1-2008, CNN	±4%	53%	46%
11-1-2008, ABC News	±2%	54%	43%
11-1-2008, CBS News	±4%	54%	41%
	ACTUAL VOTE	55%	44%

Date of Poll	Bush	Kerry
November 1	48%	46%
October 25	49%	46%
October 18	50%	45%
ACTUAL VOTE	51%	48%

95% confidence intervals are accurate only to within 2–4 percentage points, depending on sample size.

Hypothesis Testing - Five Steps

Step 1. Making assumptions and meeting test requirements

Step 2. Stating the null hypothesis

Step 3. Selecting the sampling distribution and establishing the critical region

Step 4. Computing the test statistic

Step 5. Making a decision and interpreting the results of the test

Hypothesis Testing – Finding Critical Z-Scores for One- and Two- Tailed Tests

One- vs. Two-Tailed Tests, $\alpha = 0.05$

If the Research Hypothesis Uses	The Test Is	And Concern Is with	Z(critical) =
$\neq$	Two-tailed	Both tails	±1.96
$>$	One-tailed	Upper tail	+1.65
$<$	One-tailed	Lower tail	−1.65

		One-Tailed Value	
Alpha	Two-Tailed Value	Upper Tail	Lower Tail
0.10	±1.65	+1.29	−1.29
0.05	±1.96	+1.65	−1.65
0.01	±2.58	+2.33	−2.33
0.001	±3.32	+3.10	−3.10
0.0001	±3.90	−3.70	+3.70

Hypothesis Testing – A One Tailed Test

A sociologist has noted that sociology majors seem **more sophisticated**, charming, and cosmopolitan than the rest of the student body. A "Sophistication Scale" test has been administered to the entire student body and to a random sample of 100 sociology majors, and these results have been obtained:

Student Body	Sociology Majors
$\mu = 17.3$	$\overline{X} = 19.2$
$\sigma = 7.4$	$N = 100$

Hypothesis Testing – A One Tailed Test

Step 1. Making Assumptions and Meeting Test Requirements. Since we are using a mean to summarize the sample outcome, we must assume that the Sophistication Scale generates interval-ratio–level data. With a sample size of 100, the Central Limit Theorem applies, and we can assume that the sampling distribution is normal in shape.

Step 2. Stating the Null Hypothesis (H_0). The null hypothesis states that there is no difference between sociology majors and the general student body. The research hypothesis (H_1) will also be stated at this point. The researcher has predicted a direction for the difference ("Sociology majors are *more* sophisticated"), so a one-tailed test is justified. The one-tailed research hypothesis asserts that sociology majors have a higher (>) score on the Sophistication Scale. The two hypotheses may be stated as

$$H_0: \mu = 17.3$$
$$(H_1: \mu > 17.3)$$

Hypothesis Testing – A One Tailed Test

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

If alpha is set at 0.05, the critical region will begin at the Z score +1.65. That is, the researcher has predicted that sociology majors are *more* sophisticated and that this sample comes from a population that has a mean *greater than* 17.3, so he or she will be concerned only with sample outcomes in the upper tail of the sampling distribution. If sociology majors are *the same as* other students in terms of sophistication (if the H_0 is true) or if they are *less* sophisticated (and come from a population with a mean less than 17.3), the theory is not supported. These decisions may be summarized as

$$\begin{aligned} \text{Sampling distribution} &= Z \text{ distribution} \\ \alpha &= 0.05 \\ Z(\text{critical}) &= +1.65 \end{aligned}$$

Hypothesis Testing – A One Tailed Test

Step 4. Computing the Test Statistic.

$$Z(\text{obtained}) = \frac{\overline{X} - \mu}{\sigma/\sqrt{N}}$$

$$Z(\text{obtained}) = \frac{19.2 - 17.3}{7.4/\sqrt{100}}$$

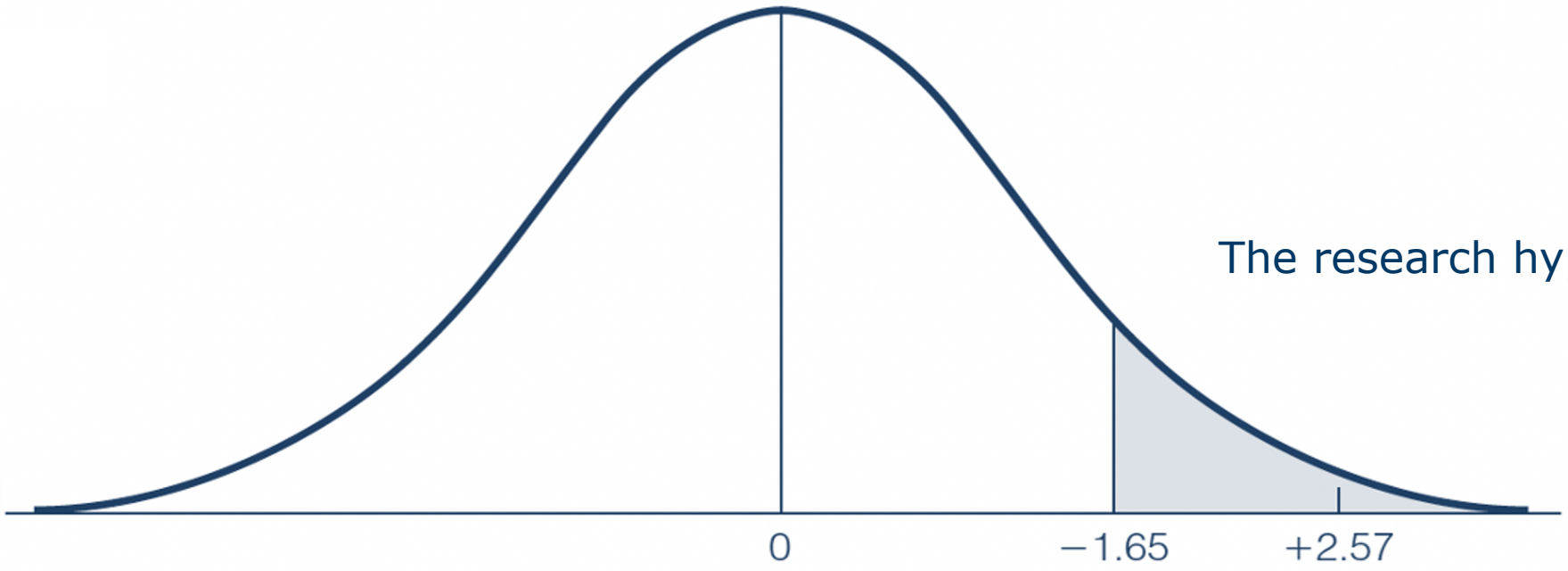
$$Z(\text{obtained}) = +2.57$$

Hypothesis Testing – A One Tailed Test

Step 5. Making a Decision and Interpreting Test Results. In this step, we will compare the Z(obtained) with the Z(critical):

$$Z(\text{critical}) = +1.65$$
$$Z(\text{obtained}) = +2.57$$

Z(obtained) Versus Z(critical) ($\alpha = 0.05$, one tailed test)



The research hypothesis is supported

Decision-Making in the Five Step Model

The H_0 Is Actually:	Decision	
	Reject	Fail to Reject
True	Type I, or α, error	OK
False	OK	Type II, or β, error

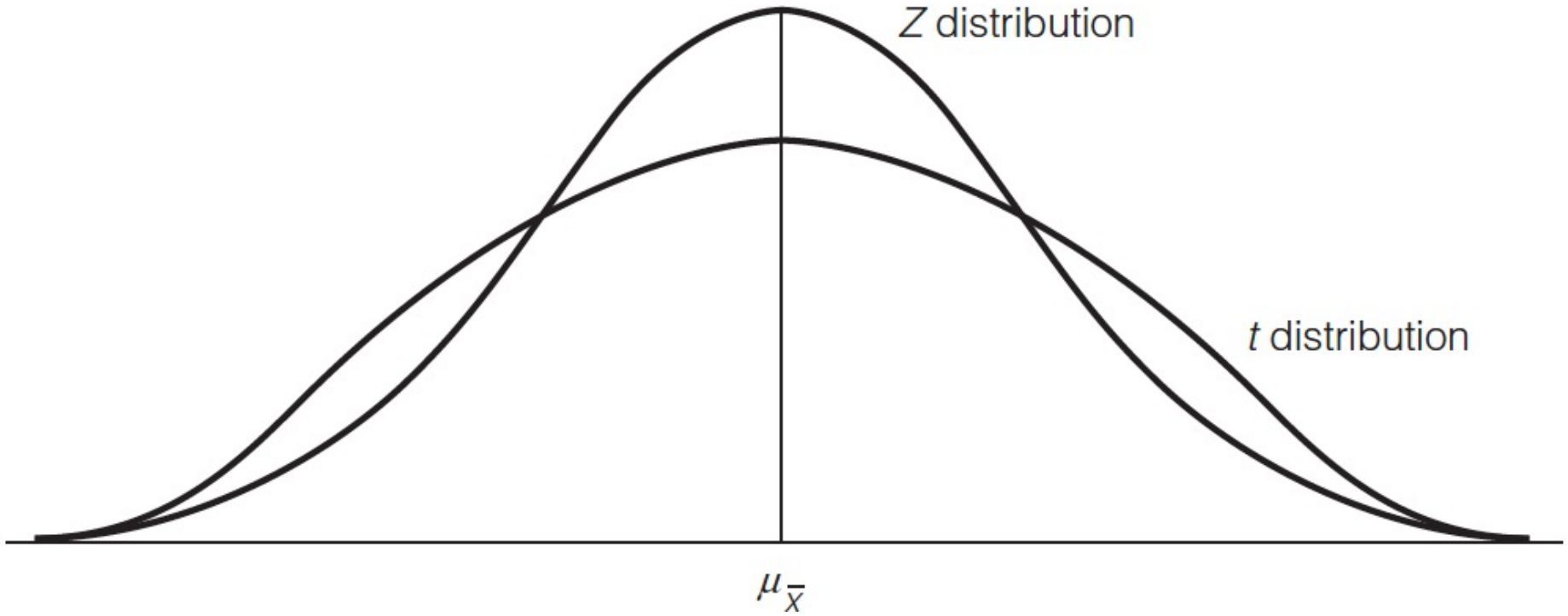
The Student's t Distribution

- For large samples (that is, samples with 100 or more cases), the sample standard deviation yields an adequate estimate of of the population standard deviation.

- For smaller samples, however, when the population standard deviation is unknown, an alternative distribution called the Student's t distribution must be used to find areas under the sampling distribution and establish the critical region.

- The shape of the t distribution varies as a function of sample size: For small samples, the t distribution is much flatter than the Z distribution, but, as sample size increases, the t distribution comes to resemble the Z distribution more and more until the two are essentially identical for sample sizes greater than 120.

The Student's t Distribution



The Student's t Distribution: Example of Sample Means

A researcher wonders if commuter students are different from the general student body in terms of academic achievement. She has gathered a random sample of 30 commuter students and has learned from the registrar that the mean grade-point average for all students is 2.50 ($\mu = 2.50$), but the standard deviation of the population (σ) has never been computed. Sample data are reported here. Is the sample from a population that has a mean of 2.50?

Student Body	Commuter Students
$\mu = 2.50\ (= \mu_{\overline{X}})$	$\overline{X} = 2.78$
$\sigma = ?$	$s = 1.23$
	$N = 30$

The Student's t Distribution: Example

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Random sampling
Level of measurement is interval-ratio
Sampling distribution is normal

Step 2. Stating the Null Hypothesis.

$$H_0: \mu = 2.50$$
$$(H_1: \mu \neq 2.50)$$

You can see from the research hypothesis that the researcher has not predicted a direction for the difference. This will be a two-tailed test.

The Student's t Distribution: Example

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region. Since σ is unknown and the sample size is small, the t distribution will be used to find the critical region. Alpha will be set at 0.01.

$$\begin{aligned} \text{Sampling distribution} &= t \text{ distribution} \\ \alpha &= 0.01, \text{ two-tailed test} \\ \text{df} &= (N - 1) = 29 \\ t(\text{critical}) &= \pm 2.756 \end{aligned}$$

Step 4. Computing the Test Statistic.

$$t(\text{obtained}) = \frac{\overline{X} - \mu}{s/\sqrt{N - 1}}$$

$$t(\text{obtained}) = \frac{2.78 - 2.50}{1.23/\sqrt{29}}$$

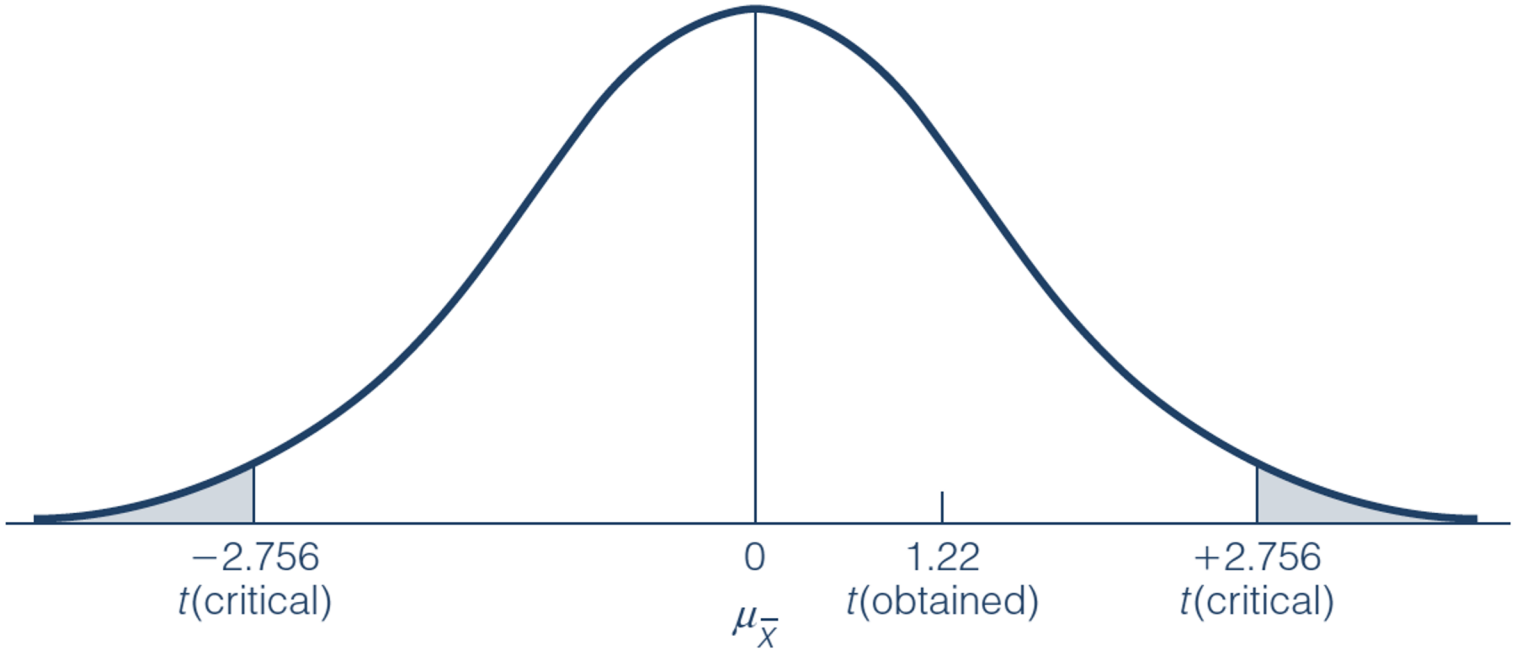
$$t(\text{obtained}) = \frac{0.28}{0.23}$$

$$t(\text{obtained}) = +1.22$$

The Student's t Distribution: Example

Step 5. Making a Decision and Interpreting Test Results. The test statistic does not fall into the critical region. Therefore, the researcher fails to reject the H_0. The difference between the sample mean (2.78) and the population mean (2.50) is not statistically significant. The difference is no greater than what would be expected if only random chance were operating.

Sampling Distribution Showing *t*(obtained) Versus *t*(critical)
($\alpha = 0.05$, two-tailed test, df = 29)



BUSINESS SCHOOL | RWTH AACHEN UNIVERSITY

Hypothesis Testing: Example of Sample Proportions

A random sample of 122 households in a low-income neighbourhood revealed that 53 (or a proportion of 0.43) of the households were headed by females. In the city as a whole, the proportion of female-headed households is 0.39. Are households in the lower-income neighbourhood significantly different from the city as a whole in terms of this characteristic?

Hypothesis Testing: Example of Sample Proportions

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Random sampling
Level of measurement is nominal
Sampling distribution is normal in shape

Step 2. Stating the Null Hypothesis. The research question, as stated earlier, asks only if the sample proportion is *different from* the population proportion. Because we have not predicted a direction for the difference, a two-tailed test will be used.

$$H_0: P_u = 0.39$$
$$(H_1: P_u \neq 0.39)$$

Hypothesis Testing: Example of Sample Proportions

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

$$\begin{aligned} \text{Sampling distribution} &= t \text{ distribution} \\ \alpha &= 0.10, \text{ two-tailed test} \\ Z(\text{critical}) &= \pm 1.65 \end{aligned}$$

Step 4. Computing the Test Statistic.

$$Z(\text{obtained}) = \frac{P_s - P_u}{\sqrt{P_u(1 - P_u)/N}}$$

$$Z(\text{obtained}) = \frac{0.43 - 0.39}{\sqrt{(0.39)(0.61)/122}}$$

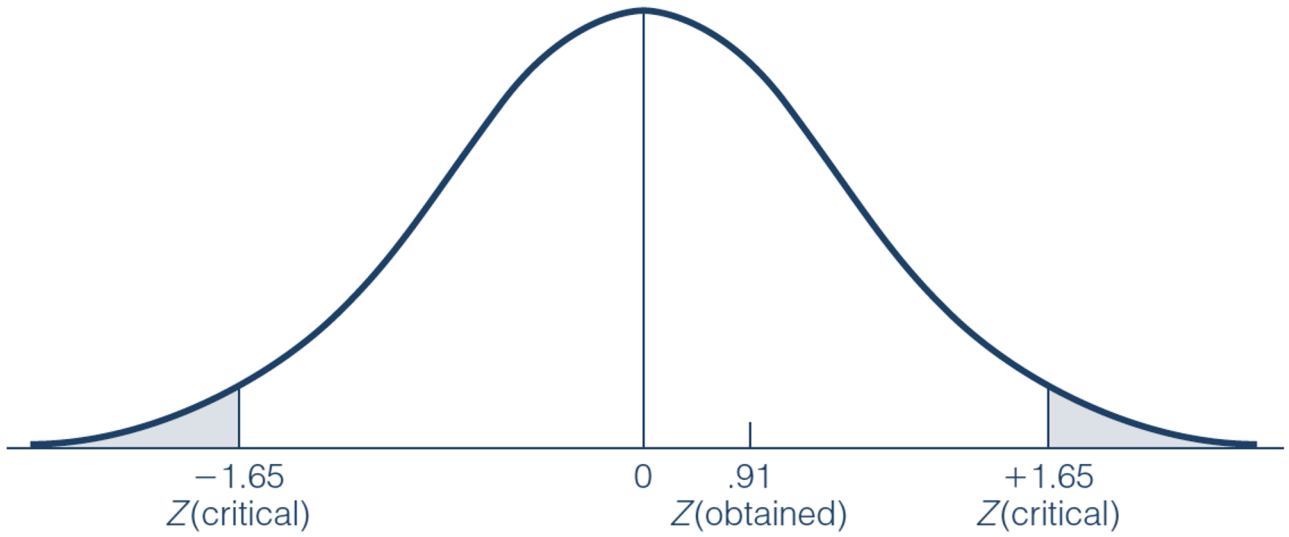
$$Z(\text{obtained}) = \frac{0.04}{0.044}$$

$$Z(\text{obtained}) = +0.91$$

Hypothesis Testing: Example of Sample Proportions

Step 5. Making a Decision and Interpreting Test Results. The test statistic, Z(obtained), does not fall into the critical region. Therefore, we fail to reject the H_0. There is no statistically significant difference between the low-income community and the city as a whole in terms of the proportion of households headed by females.

Sampling Distribution Showing *Z*(obtained) Versus *Z*(critical)
(α = 0.10, two-tailed test)



Hypothesis Testing: Significance of Sample Means

A scale measuring satisfaction with family life has been administered to a sample of married respondents. On this scale, higher scores indicate greater satisfaction. The sample has been divided into respondents with no children and respondents with at least one child, and means and standard deviations have been computed for both groups. Is there a significant difference in satisfaction with family life between these two groups?

The sample information is:

Sample 1 (No Children)	Sample 2 (At Least One Child)
$\overline{X}_1 = 11.3$	$\overline{X}_2 = 10.8$
$s_1 = 0.6$	$s_2 = 0.5$
$N_1 = 78$	$N_2 = 93$

We can see from the sample results that respondents with no children are more satisfied. Is this difference significant?

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Independent random samples
Level of measurement is interval-ratio
Sampling distribution is normal

Step 2. Stating the Null Hypothesis.

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

Sampling distribution = Z distribution
Alpha = 0.05, two-tailed
Z(critical) = ± 1.96

Hypothesis Testing: Significance of Sample Means

Step 4. Computing the Test Statistic.

$$\sigma_{\overline{X}-\overline{X}} = \sqrt{\frac{s_1^2}{N_1 - 1} + \frac{s_2^2}{N_2 - 1}}$$

$$\sigma_{\overline{X}-\overline{X}} = \sqrt{\frac{(0.6)^2}{78 - 1} + \frac{(0.5)^2}{93 - 1}}$$

$$\sigma_{\overline{X}-\overline{X}} = \sqrt{0.007}$$

$$\sigma_{\overline{X}-\overline{X}} = 0.08$$

$$Z(\text{obtained}) = \frac{(\overline{X}_1 - \overline{X}_2)}{\sigma_{\overline{X}-\overline{X}}}$$

$$Z(\text{obtained}) = \frac{11.3 - 10.8}{0.08}$$

$$Z(\text{obtained}) = \frac{0.50}{0.08}$$

$$Z(\text{obtained}) = 6.25$$

Step 5. Making a Decision and Interpreting the Results of the Test. Comparing the test statistic with the critical region,

$$Z(\text{obtained}) = 6.25$$

$$Z(\text{critical}) = \pm 1.96$$

We reject the null hypothesis. Parents and childless couples are significantly different in their satisfaction with family life. Given the direction of the difference, we can also note that childless couples are significantly happier.

RWTH BUSINESS SCHOOL

Mathematics & Statistics
M.Sc. Data Analytics and Decision Science

Prof. Dr. Thomas S. Lontzek

AACSB ACCREDITED

BUSINESS SCHOOL | RWTH AACHEN UNIVERSITY