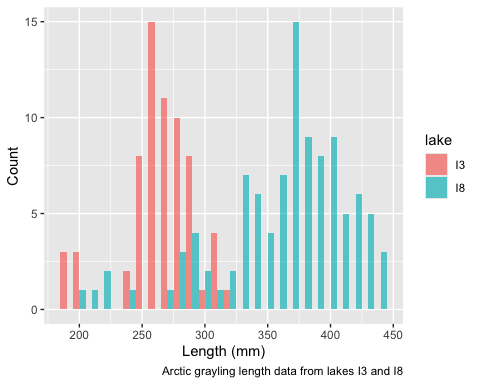
Lecture 04: Probability and Inference

Bill Perry

# **Lecture 4: Probability and Statistical Inference**

* Review of probability distributions
* Standard normal distribution and Z-scores
* Standard error and confidence intervals
* Statistical inference fundamentals
* Hypothesis testing principles

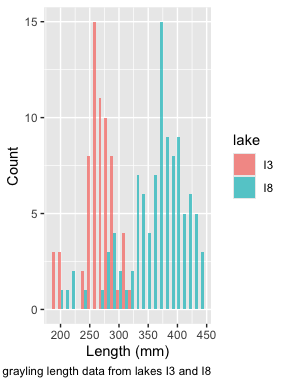


|  |
| --- |
| Practice Exercise 1: Exploring the Grayling Dataset |
| Let’s explore the Arctic grayling data from lakes I3 and I8. Use the grayling\_df data frame to create basic summary statistics.  # Write your code here to explore the basic structure of the data # also note plottig a box plot is really useful str(grayling\_df)  spc\_tbl\_ [168 × 5] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  $ site : num [1:168] 113 113 113 113 113 113 113 113 113 113 ...  $ lake : chr [1:168] "I3" "I3" "I3" "I3" ...  $ species : chr [1:168] "arctic grayling" "arctic grayling" "arctic grayling" "arctic grayling" ...  $ length\_mm: num [1:168] 266 290 262 275 240 265 265 253 246 203 ...  $ mass\_g : num [1:168] 135 185 145 160 105 145 150 130 130 71 ...  - attr(\*, "spec")=  .. cols(  .. site = col\_double(),  .. lake = col\_character(),  .. species = col\_character(),  .. length\_mm = col\_double(),  .. mass\_g = col\_double()  .. )  - attr(\*, "problems")=<externalptr>  summary(grayling\_df)  site lake species length\_mm   Min. :113 Length:168 Length:168 Min. :191.0   1st Qu.:113 Class :character Class :character 1st Qu.:270.8   Median :118 Mode :character Mode :character Median :324.5   Mean :116 Mean :324.5   3rd Qu.:118 3rd Qu.:377.0   Max. :118 Max. :440.0     mass\_g   Min. : 53.0   1st Qu.:151.2   Median :340.0   Mean :351.2   3rd Qu.:519.5   Max. :889.0   NA's :2 |

# **Lecture 4:** Probability Distributions

## Probability Distribution Functions

* A **probability distribution** describes the probability of different outcomes in an experiment
* We’ve seen histograms of observed data
* Theoretical distributions help us model and understand real-world data
* We will focus on a standard normal distribution and a t distribution

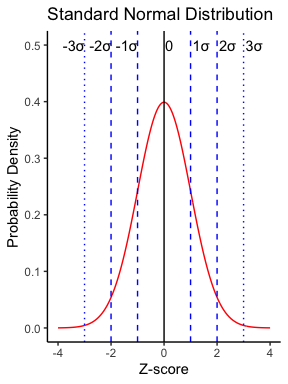


# **Lecture 4:** The Standard Normal Distribution

The standard normal distribution is crucial for understanding statistical inference:

* Has mean (μ) = 0 and standard deviation (σ) = 1
* Symmetrical bell-shaped curve
* Area under the curve = 1 (total probability)
* Approximately:
  + 68% of data within ±1σ of the mean
  + **95% of data within ±2σ of the mean - really 1.96σ**
  + 99.7% of data within ±3σ of the mean

Z-scores allow us to convert any normal distribution to the standard normal distribution.



|  |
| --- |
| Practice Exercise 2: Calculating Z-scores |
| Let’s practice converting raw values to Z-scores using the Arctic grayling data.  # Calculate the mean and standard deviation of fish lengths mean\_length <- mean(grayling\_df$length\_mm, na.rm = TRUE) sd\_length <- sd(grayling\_df$length\_mm, na.rm = TRUE)  # Calculate Z-scores for fish lengths grayling\_df <- grayling\_df %>%  mutate(z\_score = (length\_mm - mean\_length) / sd\_length)  # View the first few rows with Z-scores head(grayling\_df)  # A tibble: 6 × 6  site lake species length\_mm mass\_g z\_score  <dbl> <chr> <chr> <dbl> <dbl> <dbl> 1 113 I3 arctic grayling 266 135 -0.900 2 113 I3 arctic grayling 290 185 -0.531 3 113 I3 arctic grayling 262 145 -0.961 4 113 I3 arctic grayling 275 160 -0.761 5 113 I3 arctic grayling 240 105 -1.30  6 113 I3 arctic grayling 265 145 -0.915 |

# What proportion of fish are within 1 standard deviation of the mean?  
within\_1sd <- sum(abs(grayling\_df$z\_score) <= 1, na.rm = TRUE) / sum(!is.na(grayling\_df$z\_score))  
cat("Proportion within 1 SD:", round(within\_1sd \* 100, 1), "%\n")

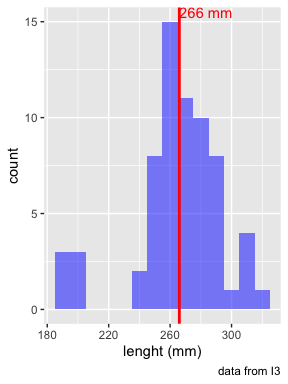
Proportion within 1 SD: 64.3 %

# Lecture 4: Standard normal distribution

You want to know things about this population like

* probability of afish having a certain length (e.g., > 300 mm)
* Can solve this by integrating under curve
* But it is tedious to do every time
* Instead
  + we can use the *standard normal distribution* (SND)

# A tibble: 1 × 1  
 mean\_length  
 <dbl>  
1 266.

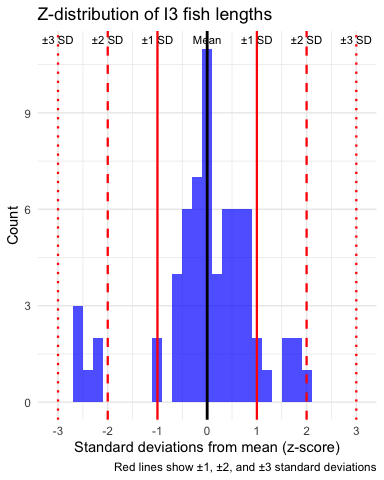


# Lecture 4: Standard normal distribution

Standard Normal Distribution

* “benchmark” normal distribution with µ = 0, σ = 1
* The Standard Normal Distribution is defined so that:
  + ~68% of the curve area within +/- 1 σ of the mean,
  + ~95% within +/- 2 σ of the mean,
  + ~99.7% within +/- 3 σ of the mean

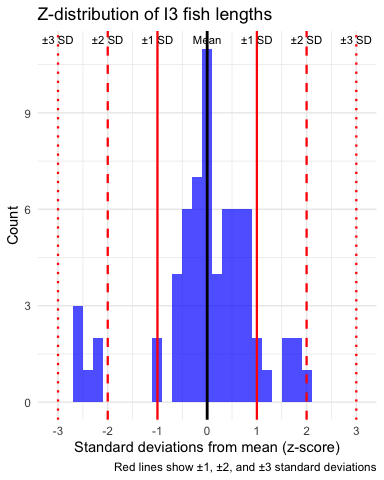
\*remember σ = standard deviation



# Lecture 4: Standard normal distribution

Areas under curve of Standard Normal Distribution

* Have been calculated for a range of sample sizes
* Can be looked up in z-table
* No need to integrate
* Any normally distributed data can be standardized
  + transformed into the standard normal distribution
  + a value can ber looked up in a table



# Lecture 4: Standard normal distribution

Done by converting original data points to z-scores

* Z-scores calculated as:

* z = z-score for observation
* xi = original observation
* µ = mean of data distribution
* σ = SD of data distribution

So lets do this for a fish that is 300mm long and guess the probability of catching something larger

z = (300 - 265.61)/28.3 = 1.215194

i3\_stats <- gray\_i3\_df %>%  
 summarize(  
 mean\_length = round(mean(length\_mm, na.rm = TRUE), 2),  
 sd\_length = sd(length\_mm, na.rm = TRUE),  
 n = sum(!is.na(length\_mm)),  
 se\_length = round(sd\_length / sqrt(sum(!is.na(length\_mm))), 2),  
 .groups = "drop"  
 )  
  
# Display the results  
i3\_stats

# A tibble: 1 × 4  
 mean\_length sd\_length n se\_length  
 <dbl> <dbl> <int> <dbl>  
1 266. 28.3 66 3.48

# Lecture 4: Standard normal distribution

Done by converting original data points to z-scores

* Z-scores calculated as:

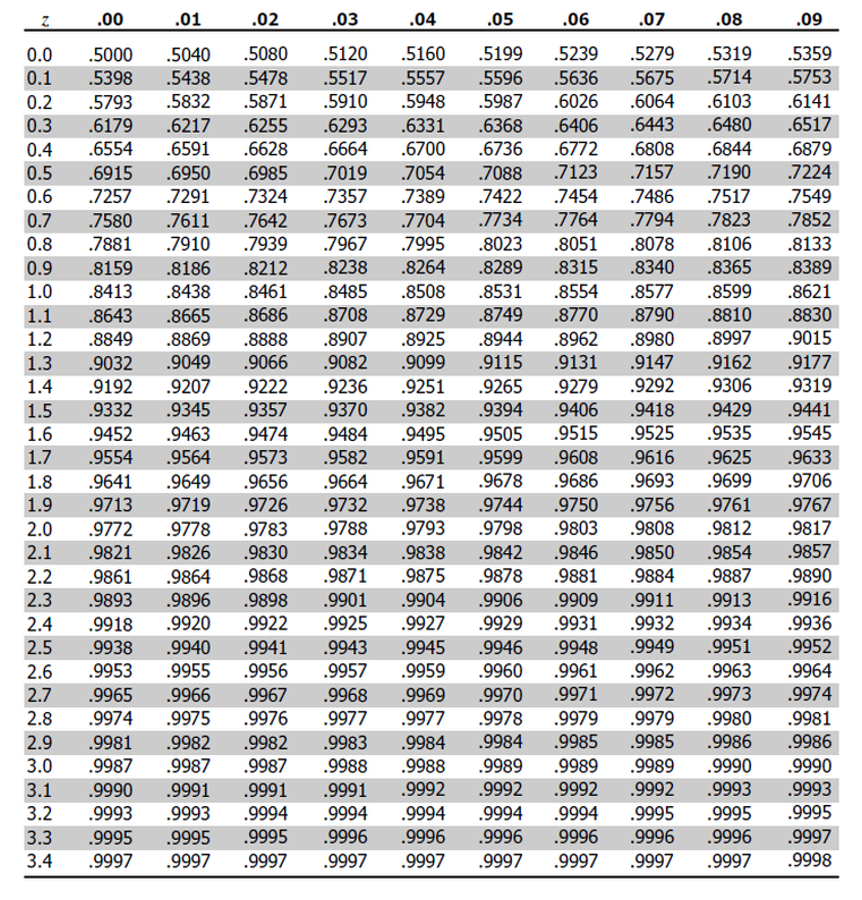
* z = z-score for observation
* xi = original observation
* µ = mean of data distribution
* σ = SD of data distribution

So lets do this for a fish that is 320mm long and guess the probability of catching something larger

z = (320 - 265.61)/28.3 = 1.92

or .9726 in table or 97.3% is the area left of the curve and

100 - 97.3 = 2.7% or 2.7% of fish are expected to be longer



# **Lecture 4:** Sampling a population - Std Error

The **standard error of the mean (SEM)** tells us how precise our sample mean is as an estimate of the population mean.

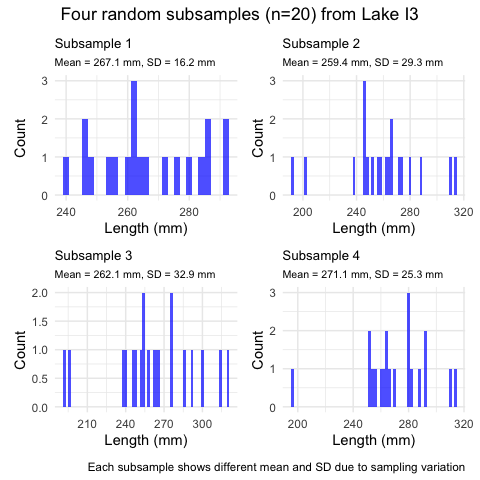
Standard Error Formula:

Where:

* is the sample standard deviation
* is the sample size

**Key properties:**

* SEM decreases as sample size increases
* SEM is used to construct confidence intervals
* SEM measures the precision of the sample mean



|  |
| --- |
| Practice Exercise 5: Sampling Distributions |
| Let’s explore how sample size affects our estimates by taking samples of different sizes:  # Set seed for reproducibility set.seed(456)  # Create samples of different sizes small\_sample <- grayling\_df %>% sample\_n(5) medium\_sample <- grayling\_df %>% sample\_n(30) large\_sample <- grayling\_df %>% sample\_n(125)  # Calculate mean and standard error for each sample small\_mean <- mean(small\_sample$length\_mm, na.rm = TRUE) small\_se <- sd(small\_sample$length\_mm, na.rm = TRUE) / sqrt(10)  medium\_mean <- mean(medium\_sample$length\_mm, na.rm = TRUE) medium\_se <- sd(medium\_sample$length\_mm, na.rm = TRUE) / sqrt(30)  large\_mean <- mean(large\_sample$length\_mm, na.rm = TRUE) large\_se <- sd(large\_sample$length\_mm, na.rm = TRUE) / sqrt(100)  # Create a data frame with the results results <- data.frame(  Sample\_Size = c(10, 30, 100),  Mean = c(small\_mean, medium\_mean, large\_mean),  SE = c(small\_se, medium\_se, large\_se) )  # Display the results results  Sample\_Size Mean SE 1 10 302.000 26.607330 2 30 319.200 12.082989 3 100 323.328 6.478149  What do you observe about the standard error as sample size increases? Why does this happen? |

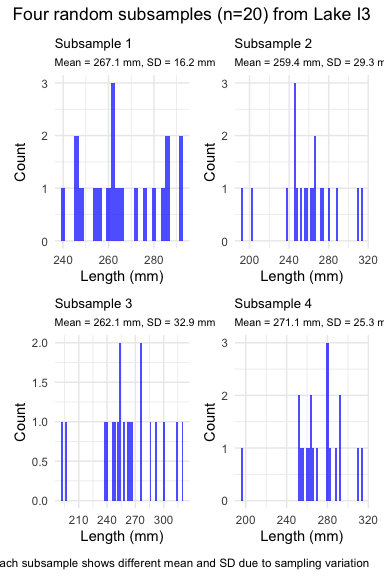
# **Lecture 4:** Estimating µ - population mean

## Every sample gives slightly different estimate of µ

* Can take many samples and calculate means
* Plot the frequency distribution of means
* Get the “sampling distribution of means”

## 3 important properties:

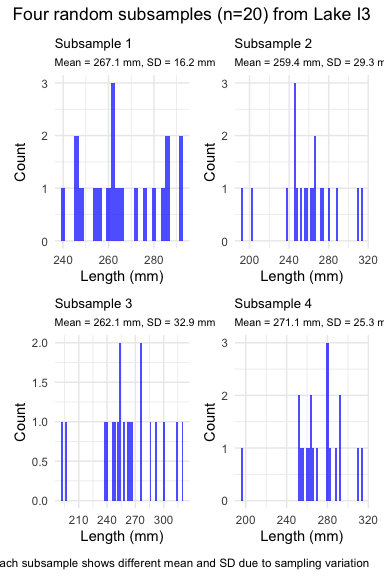
* Sampling distribution of means (SDM) from normal population will be normal
* Large Sampling distribution of means from any population will be normal (Central Limit Theorem)
* The mean of Sampling distribution of means will equal µ or the mean



# **Lecture 4:** Estimating µ - population mean

## Given above

* can estimate the standard deviation of sample means
* “Standard error of sample mean”
* How good is your estimate of population mean? (based on the sample collected)
* quantifies how much the sample means are expected to vary from samples
* gives an estimate of the error associated with using to estimate …

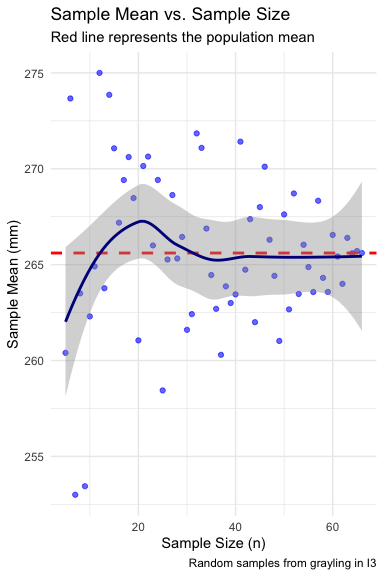


# **Lecture 4:** Estimating µ - population mean

Notice: - depends on - sample s (standard deviation) - sample n - ()

How and why? - Decreases with sample n - number - increases with sample s - standard deviation

* Large sample, low s = greater confidence in estimate of



# **Lecture 4:** Standard Error of the Mean

The **standard error of the mean (SEM)** tells us how precise our sample mean is as an estimate of the population mean.

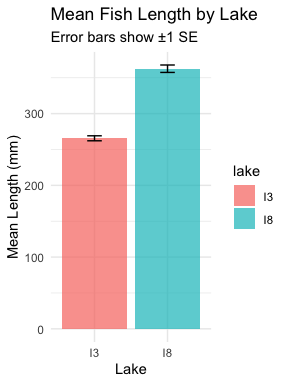
Standard Error Formula:

Where:

* is the sample standard deviation
* is the sample size

**Key properties:**

* SEM decreases as sample size increases
* SEM is used to construct confidence intervals
* SEM measures the precision of the sample mean



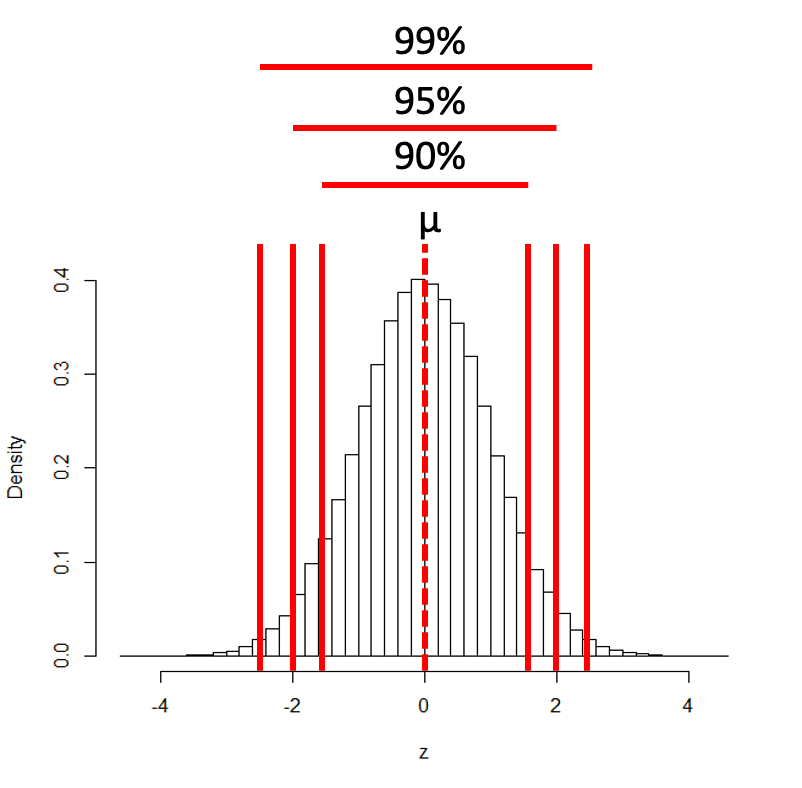
# **Lecture 4:** Confidence Intervals

A **confidence interval** is a range of values that is likely to contain the true population parameter.

95% Confidence Interval Formula:

Where:

* ȳ is the sample mean
* 𝑛 is the sample size
* σ is the population standard deviation
* z is the z-value corresponding the probability of the CI



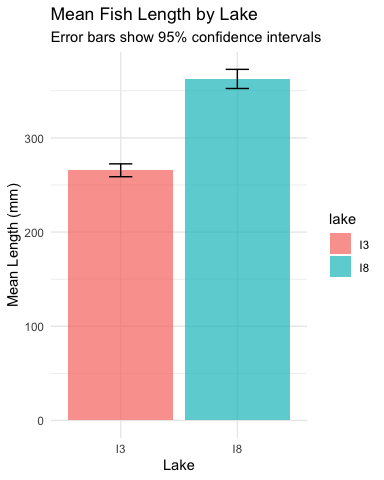
# **Lecture 4:** Confidence Intervals

A **confidence interval** is a range of values that is likely to contain the true population parameter.

**Interpretation:** If we were to take many samples and calculate the 95% CI for each, about 95% of these intervals would contain the true population mean.

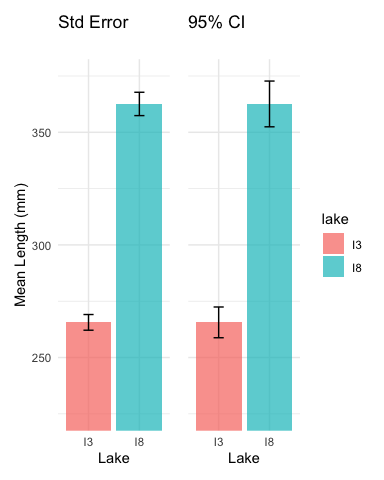
**Common misinterpretation:** “There is a 95% probability that the true mean is in this interval.”

* Interpret 95% CI to mean:
  + Range of values that contains µ (population mean) with 95% probability
* More correctly:
  + If we took 100 samples from population
  + calculate a CI from each
  + 95 of the 100 CIs will contain the true population mean - µ



# **Lecture 4:** Compare the SE and CI plots

Lets compare what the two plots look like near each other



|  |
| --- |
| Practice Exercise 3: Calculating Standard Error and Confidence Intervals |
| Calculate the standard error and 95% confidence interval for the mean length of Arctic grayling in each lake.  # Calculate the standard error and confidence intervals by lake ci\_results <- grayling\_df %>%  group\_by(lake) %>%  summarize(  mean\_length = round(mean(length\_mm, na.rm = TRUE), 2),  sd\_length = sd(length\_mm, na.rm = TRUE),  n = sum(!is.na(length\_mm)),  se\_length = round(sd\_length / sqrt(n), 2),  ci = round(1.96 \* se\_length, 2),  ci\_lower = round(mean\_length - 1.96 \* se\_length, 2),  ci\_upper = round(mean\_length + 1.96 \* se\_length, 2),  .groups = "drop"  )  # Display the results ci\_results  # A tibble: 2 × 8  lake mean\_length sd\_length n se\_length ci ci\_lower ci\_upper  <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <dbl> 1 I3 266. 28.3 66 3.48 6.82 259. 272. 2 I8 363. 52.3 102 5.18 10.2 352. 373.  What do these confidence intervals tell us about the difference between lakes? |

# **Lecture 4:** Confidence intervals

In the more typical case DON’T know the population σ

* estimate it from the samples when don’t know the population σ
* and when sample size is <~30)
* can’t use the standard normal (z) distribution

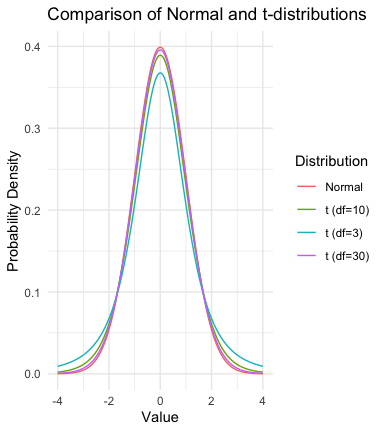
*Instead, we use Student’s t distribution*



# **Lecture 4:** Understanding t-distribution

When sample sizes are small, the **t-distribution** is more appropriate than the normal distribution.

* Similar to normal distribution but with heavier tails
* Shape depends on **degrees of freedom** (df = n-1)
* With large df (>30), approaches the normal distribution
* Used for:
  + Small sample sizes
  + When population standard deviation is unknown
  + Calculating confidence intervals
  + Conducting t-tests



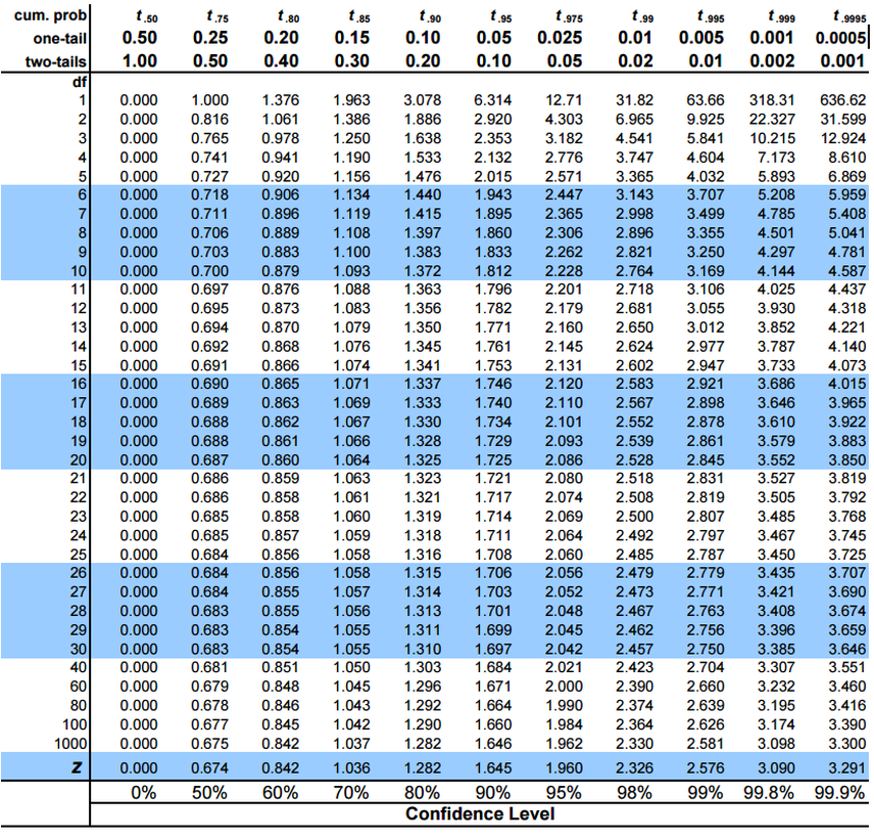
|  |
| --- |
| Practice Exercise 4: Using the t-distribution |
| Let’s compare confidence intervals using the normal approximation (z) versus the t-distribution for our fish data.  # Calculate CI using both z and t distributions for a smaller subset small\_sample <- grayling\_df %>%   filter(lake == "I3") %>%   slice\_sample(n = 10)  # Calculate statistics sample\_mean <- mean(small\_sample$length\_mm) sample\_sd <- sd(small\_sample$length\_mm) sample\_n <- nrow(small\_sample) sample\_se <- sample\_sd / sqrt(sample\_n)  # Calculate confidence intervals z\_ci\_lower <- sample\_mean - 1.96 \* sample\_se z\_ci\_upper <- sample\_mean + 1.96 \* sample\_se  # For t-distribution, get critical value for 95% CI with df = n-1 t\_crit <- qt(0.975, df = sample\_n - 1) t\_ci\_lower <- sample\_mean - t\_crit \* sample\_se t\_ci\_upper <- sample\_mean + t\_crit \* sample\_se  # Display results cat("Mean:", round(sample\_mean, 1), "mm\n")  Mean: 255.3 mm  cat("Standard deviation:", round(sample\_sd, 2), "mm\n")  Standard deviation: 26.26 mm  cat("Standard error:", round(sample\_se, 2), "mm\n")  Standard error: 8.31 mm  cat("95% CI using z:", round(z\_ci\_lower, 1), "to", round(z\_ci\_upper, 1), "mm\n")  95% CI using z: 239 to 271.6 mm  cat("95% CI using t:", round(t\_ci\_lower, 1), "to", round(t\_ci\_upper, 1), "mm\n")  95% CI using t: 236.5 to 274.1 mm  cat("t critical value:", round(t\_crit, 3), "vs z critical value: 1.96\n")  t critical value: 2.262 vs z critical value: 1.96 |

# Student’s t-distribution

To calculate CI for sample from “unknown” population:

Where:

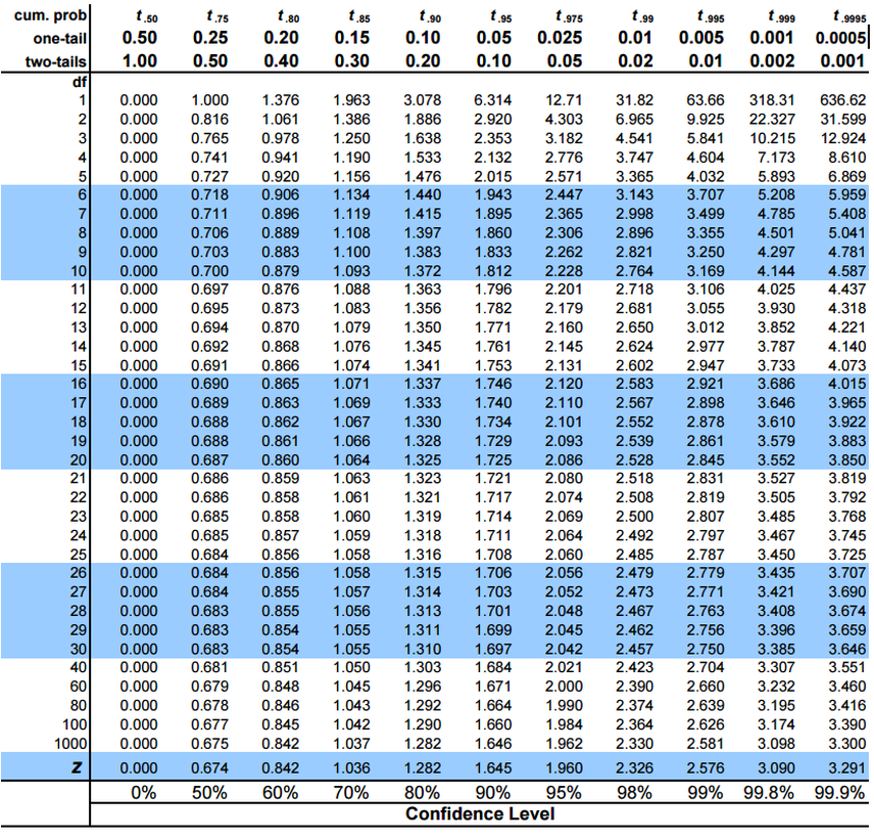
* ȳ is sample mean
* 𝑛 is sample size
* s is sample standard deviation
* t t-value corresponding the probability of the CI
* t in t-table for different degrees of freedom (n-1)



# **Lecture 5:** Student’s t-distribution

Here is a t-table

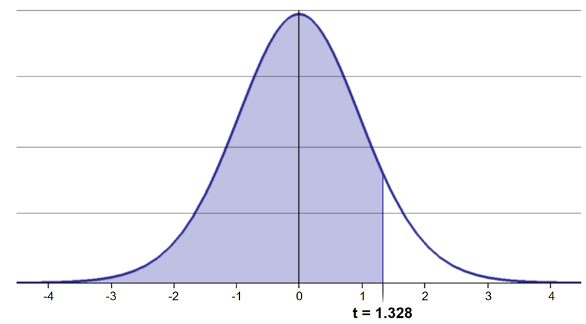
* Values of t that correspond to probabilities
* Probabilities listed along top
* Sample dfs are listed in the left-most column
* Probabilities are given for one-tailed and two-tailed “questions”

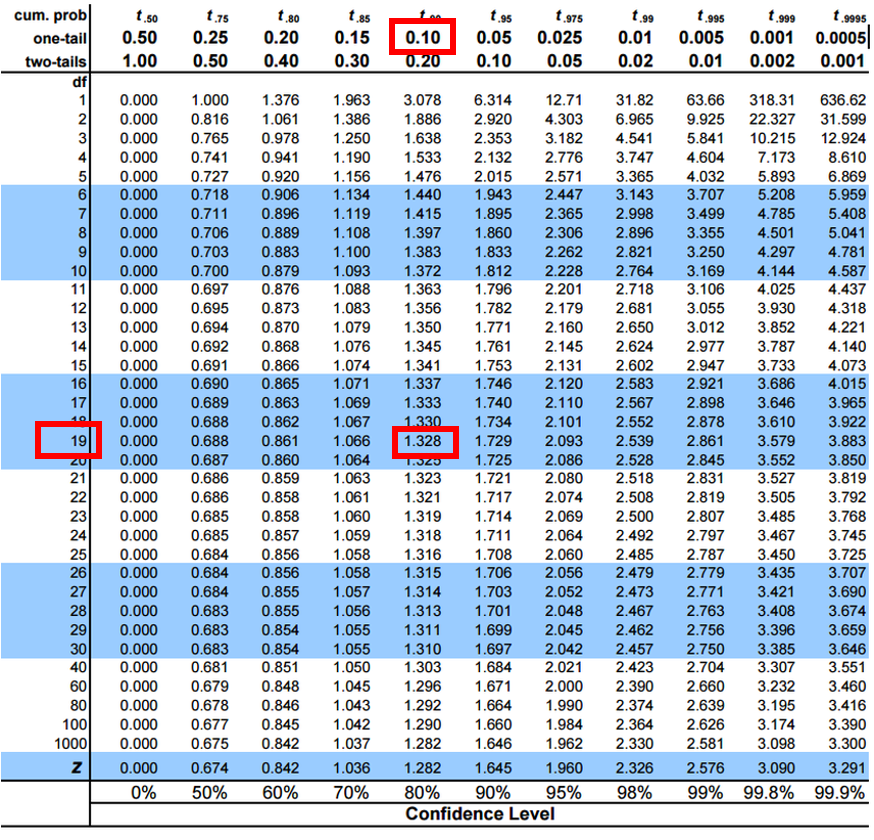


# **Lecture 5:** Student’s t-distribution

One-tailed questions: area of distribution left or (right) of a certain value

* n=20 (df=19) - 90% of the observations found left
* t= 1.328 (10% are outside)

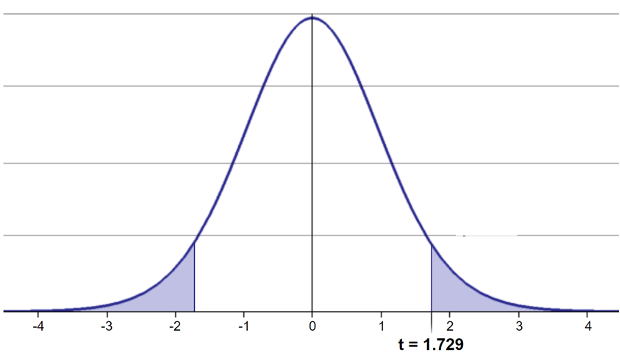


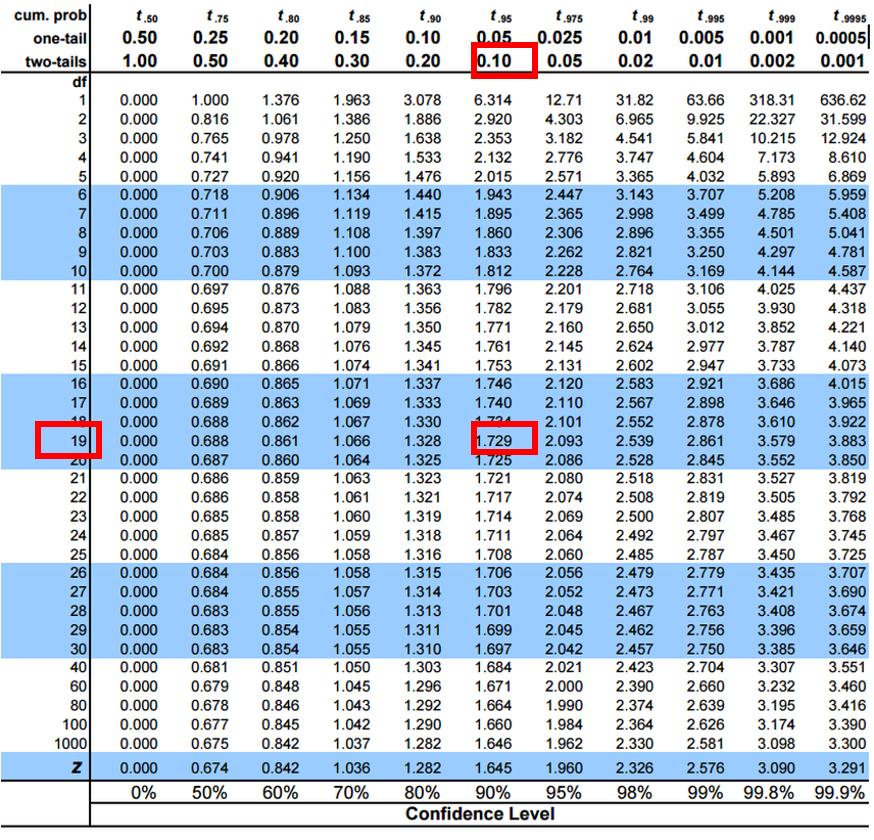


# **Lecture 5:** Student’s t-distribution

Two-tailed questions refer to area between certain values

* n= 20 (df=19), 90% of the observations are between
* t=-1.729 and t=1.729 (10% are outside)



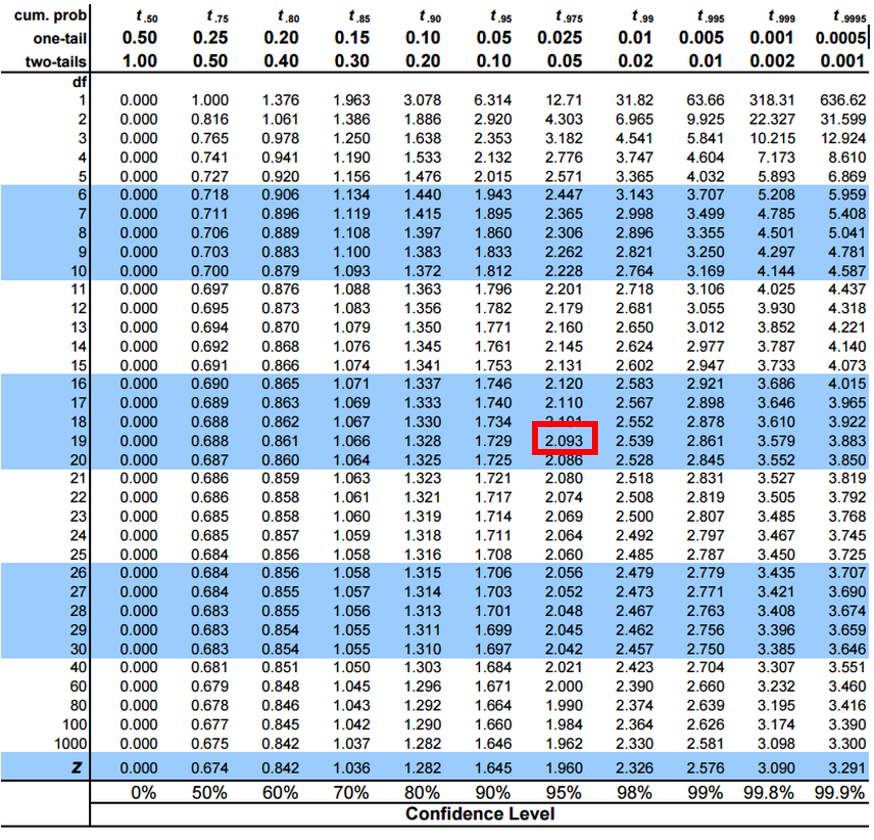


# **Lecture 5:** Student’s t-distribution

Let’s calculate CIs again:

Use two-sided test

* 95% CI Sample A: = 272.8 ± 2.262 \* (37.81/(9^0.5)) = 1.650788
* The 95% CI is between 244.3 and 301.3
* “The 95% CI for the population mean from sample A is 272.8 ± 28.5”



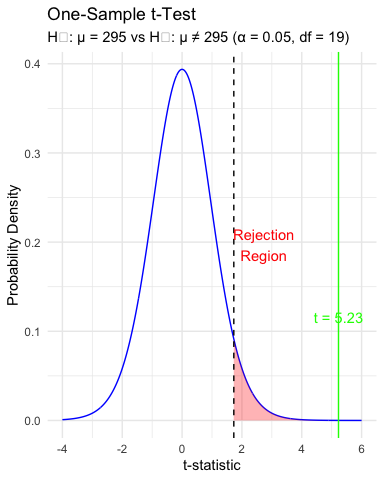
# **Lecture 4:** Intro to Hypothesis Testing

Hypothesis testing is a systematic way to evaluate research questions using data.

**Key components:**

1. **Null hypothesis (H₀)**: Typically assumes “no effect” or “no difference”
2. **Alternative hypothesis (Hₐ)**: The claim we’re trying to support
3. **Statistical test**: Method for evaluating evidence against H₀
4. **P-value**: Probability of observing our results (or more extreme) if H₀ is true
5. **Significance level (α)**: Threshold for rejecting H₀, typically 0.05

**Decision rule**: Reject H₀ if p-value < α



# **Lecture 4:** Intro to Hypothesis Testing

Hypothesis testing is a systematic way to evaluate research questions using data.

**Key components:**

1. **Null hypothesis (H₀)**: Typically assumes “no effect” or “no difference”
2. **Alternative hypothesis (Hₐ)**: The claim we’re trying to support
3. **Statistical test**: Method for evaluating evidence against H₀
4. **P-value**: Probability of observing our results (or more extreme) if H₀ is true
5. **Significance level (α)**: Threshold for rejecting H₀, typically 0.05

**Decision rule**: Reject H₀ if p-value < α



|  |
| --- |
| Practice Exercise 5: Lets practice a One-Sample t-Test |
| Let’s perform a one-sample t-test to determine if the mean fish length in Toolik Lake differs from 50 mm:  # get only lake I# i3\_df <- grayling\_df %>% filter(lake=="I3")  # what is the mean i3\_mean <- mean(i3\_df$length\_mm, na.rm=TRUE) cat("Mean:", round(i3\_mean, 1), "mm\n")  Mean: 265.6 mm  # Perform a one-sample t-test t\_test\_result <- t.test(i3\_df$length\_mm, mu = 260)  # View the test results t\_test\_result  One Sample t-test  data: i3\_df$length\_mm t = 1.6091, df = 65, p-value = 0.1124 alternative hypothesis: true mean is not equal to 260 95 percent confidence interval:  258.6481 272.5640 sample estimates: mean of x   265.6061  Interpret this test result by answering these questions:   1. What was the null hypothesis? 2. What was the alternative hypothesis? 3. What does the p-value tell us? 4. Should we reject or fail to reject the null hypothesis at α = 0.05? 5. What is the practical interpretation of this result for fish biologists? |

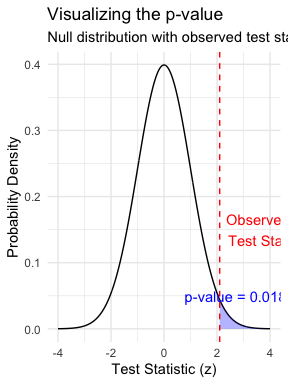
|  |
| --- |
| Practice Exercise 6: Formulating Hypotheses |
| For the following research questions about Arctic grayling, write the null and alternative hypotheses:   1. Are fish in Lake I8 longer than fish in Lake I3? 2. Is the mean length of Arctic grayling in these lakes different from 300 mm? 3. Is there a relationship between fish length and mass?   # Let's test one of these hypotheses: Are fish in Lake I8 longer than fish in Lake I3?  # Perform an independent t-test t\_test\_result <- t.test(length\_mm ~ lake, data = grayling\_df,   alternative = "less") # H₀: μ\_I3 ≥ μ\_I8, H₁: μ\_I3 < μ\_I8  # Display the results t\_test\_result  Welch Two Sample t-test  data: length\_mm by lake t = -15.532, df = 161.63, p-value < 2.2e-16 alternative hypothesis: true difference in means between group I3 and group I8 is less than 0 95 percent confidence interval:  -Inf -86.66138 sample estimates: mean in group I3 mean in group I8   265.6061 362.5980  Based on this t-test, what can we conclude about the difference in fish length between the two lakes? |

# **Lecture 4:** Understanding P-values

A **p-value** is the probability of observing the sample result (or something more extreme) if the null hypothesis is true.

**Common interpretations:** - p < 0.05: Strong evidence against H₀ - 0.05 ≤ p < 0.10: Moderate evidence against H₀ - p ≥ 0.10: Insufficient evidence against H₀

**Common misinterpretations:** - p-value is NOT the probability that H₀ is true - p-value is NOT the probability that results occurred by chance - Statistical significance ≠ practical significance



# **Lecture 4:** Type I and Type II Errors

When making decisions based on hypothesis tests, two types of errors can occur:

**Type I Error (False Positive)** - Rejecting H₀ when it’s actually true - Probability = α (significance level) - “Finding an effect that isn’t real”

**Type II Error (False Negative)** - Failing to reject H₀ when it’s actually false - Probability = β - “Missing an effect that is real”

**Statistical Power = 1 - β** - Probability of correctly rejecting a false H₀ - Increases with: - Larger sample size - Larger effect size - Lower variability - Higher α level



|  |
| --- |
| Practice Exercise 6: Interpreting P-values and Errors |
| Given the following scenarios, identify whether a Type I or Type II error might have occurred:   1. A researcher concludes that a new fishing regulation increased grayling size, when in fact it had no effect. 2. A study fails to detect a real decline in grayling population due to warming water, concluding there was no effect. 3. Let’s calculate the power of our t-test to detect a 30 mm difference in length between lakes:   # Calculate power for detecting a 30 mm difference # First determine parameters lake\_I3 <- grayling\_df %>% filter(lake == "I3") lake\_I8 <- grayling\_df %>% filter(lake == "I8")   n1 <- nrow(lake\_I3) n2 <- nrow(lake\_I8) sd\_pooled <- sqrt((var(lake\_I3$length\_mm) \* (n1-1) +   var(lake\_I8$length\_mm) \* (n2-1)) /   (n1 + n2 - 2))  # Calculate power effect\_size <- 30 / sd\_pooled # Cohen's d df <- n1 + n2 - 2 alpha <- 0.05 power <- power.t.test(n = min(n1, n2),   delta = effect\_size,  sd = 1, # Using standardized effect size  sig.level = alpha,  type = "two.sample",  alternative = "two.sided")  # Display results power  Two-sample t test power calculation    n = 66  delta = 0.6741298  sd = 1  sig.level = 0.05  power = 0.9702076  alternative = two.sided  NOTE: n is number in \*each\* group |

# **Lecture 4:** Summary

**Key concepts covered:**

1. **Probability distributions** model random phenomena
   * Normal distribution is especially important
   * Z-scores standardize measurements
2. **Standard error** measures precision of estimates
   * Decreases with larger sample sizes
   * Used to construct confidence intervals
3. **Confidence intervals** express uncertainty
   * Provide plausible range for parameters
   * 95% CI: mean ± 1.96 × SE
4. **Hypothesis testing** evaluates claims
   * Null vs. alternative hypotheses
   * P-values quantify evidence against H₀
   * Consider both statistical and practical significance

