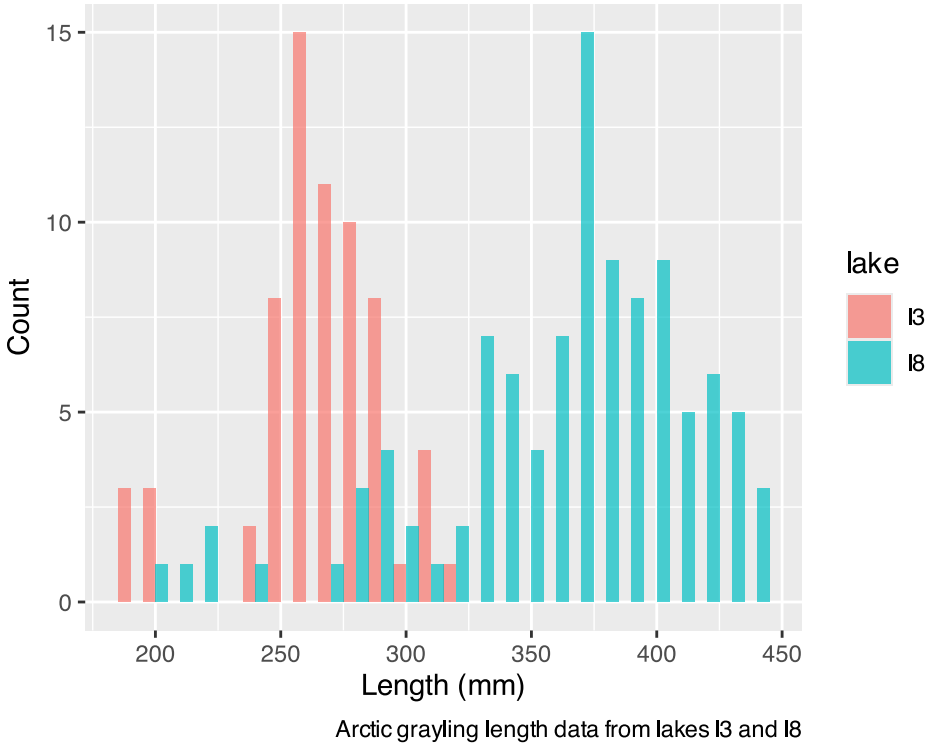
Lecture 04: Probability and Inference

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

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)
```

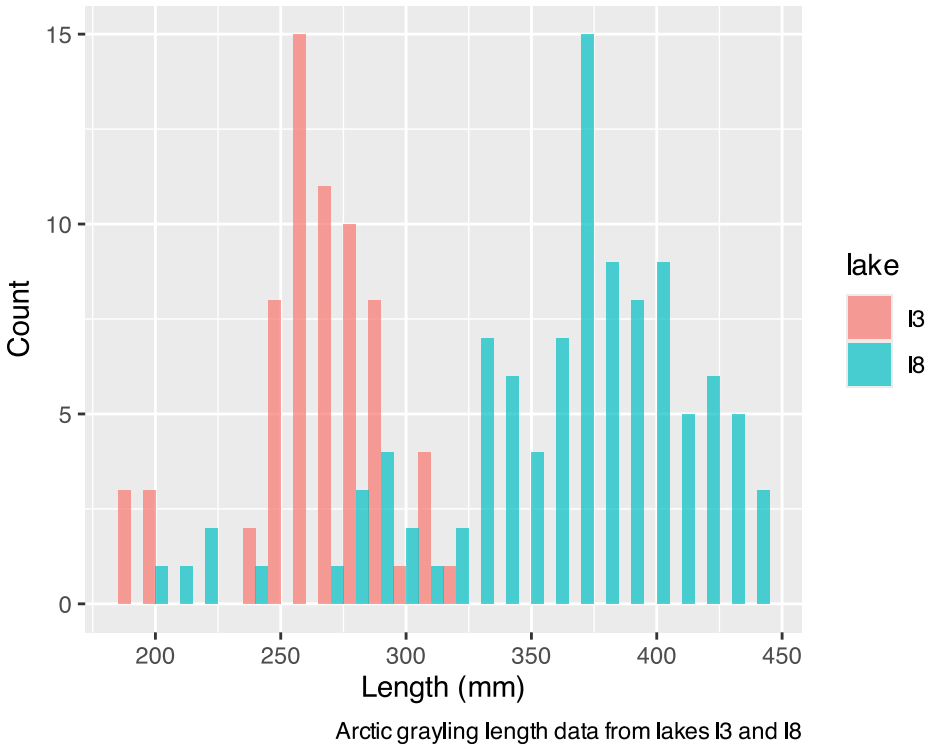
```
      site         lake             species            length_mm
 Min.   :113   Length:168         Length:168         Min.   :191.0
 1st Qu.:113   Class :character   Class :character   1st Qu.:270.8
 Median :118   Mode  :character   Mode  :character   Median :324.5
 Mean   :116                                        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

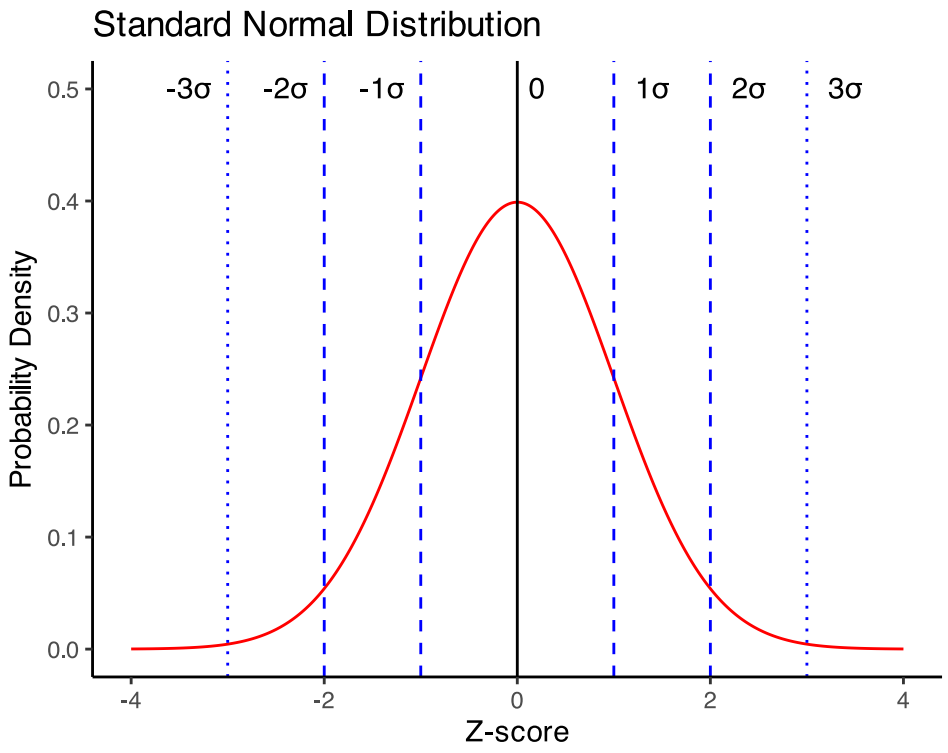


Lecture 4: The Standard Normal Distribution

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

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
```

Z-score Results

```r
# 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")
```

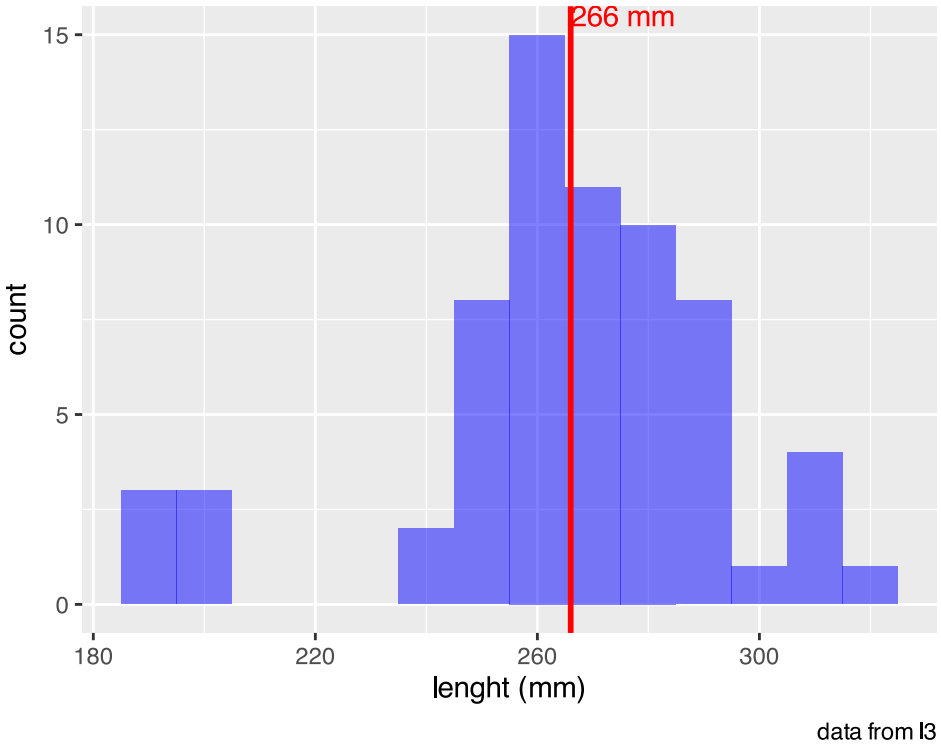
```
Proportion within 1 SD: 64.3 %
```

Lecture 4: Standard normal distribution - Fish Data

You want to know things about this population like

- probability of a fish 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.
```



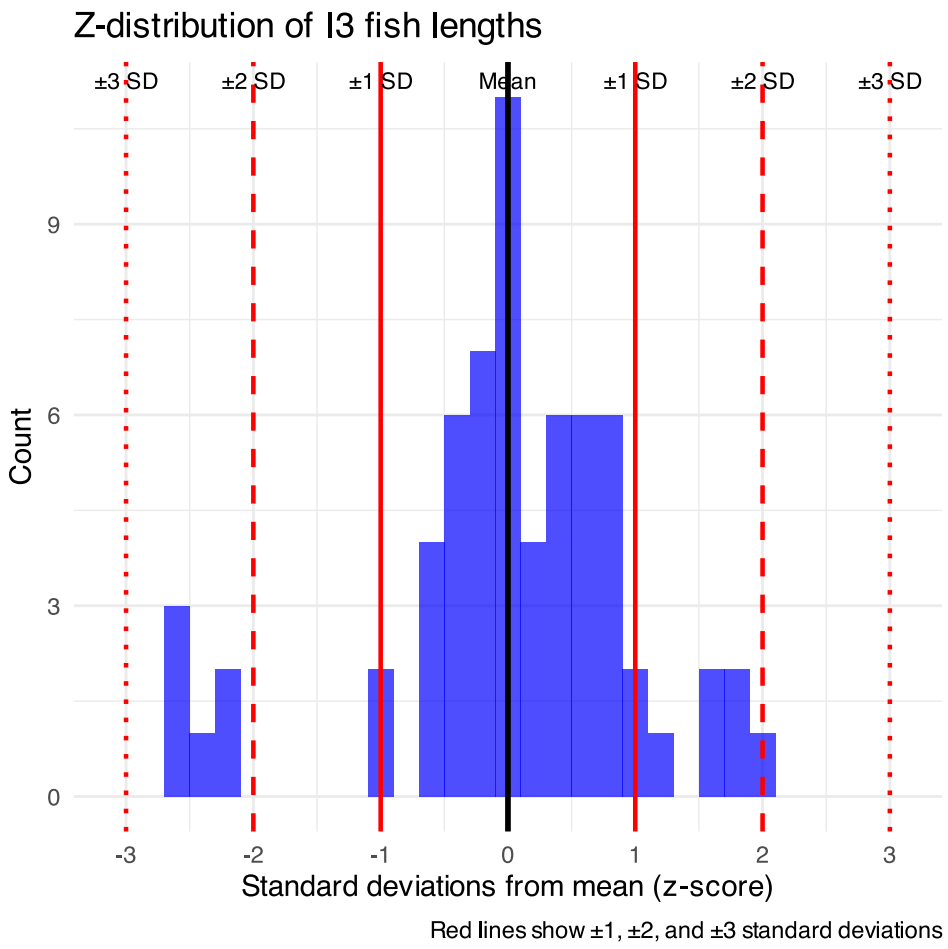
Lecture 4: Standard normal distribution properties

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

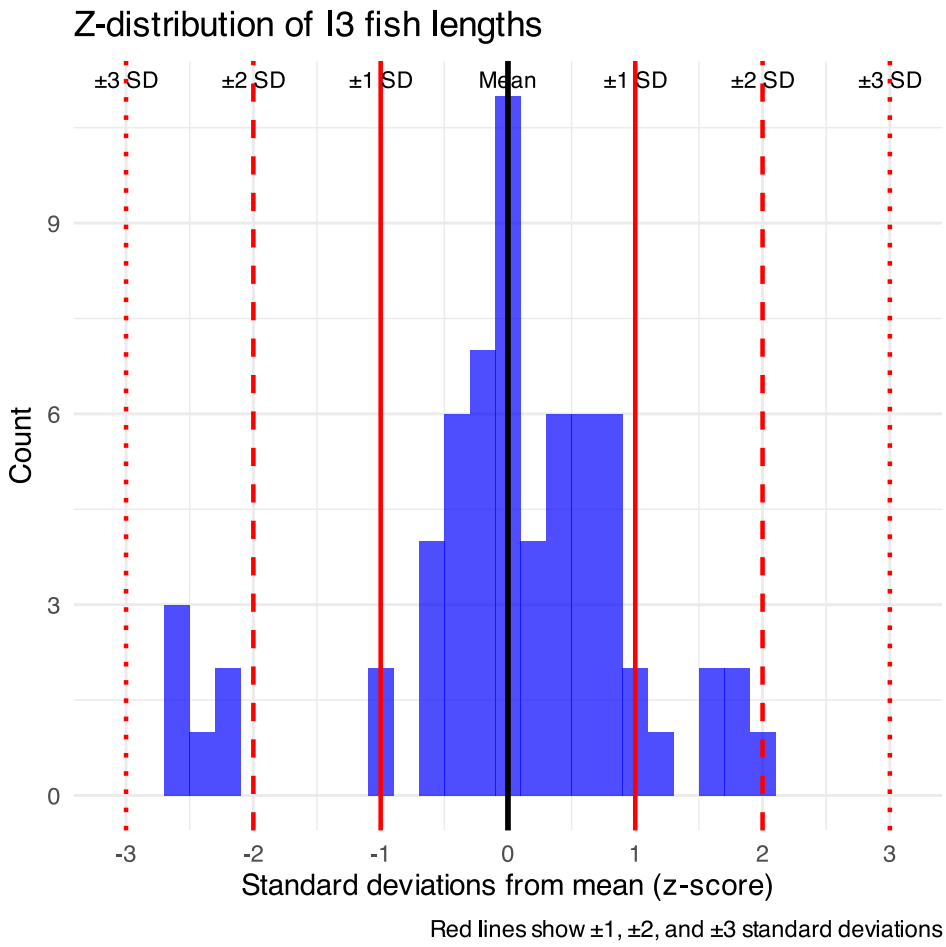
*remember σ = standard deviation



Lecture 4: Using Z-tables

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 be looked up in a table



Lecture 4: Z-score Formula

Done by converting original data points to z-scores

- Z-scores calculated as:

$$Z = \frac{X_i - \mu}{\sigma}$$

- 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: Z-score Example

Done by converting original data points to z-scores

- Z-scores calculated as:

$$Z = \frac{X_i - \mu}{\sigma}$$

- 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

0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993

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:

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

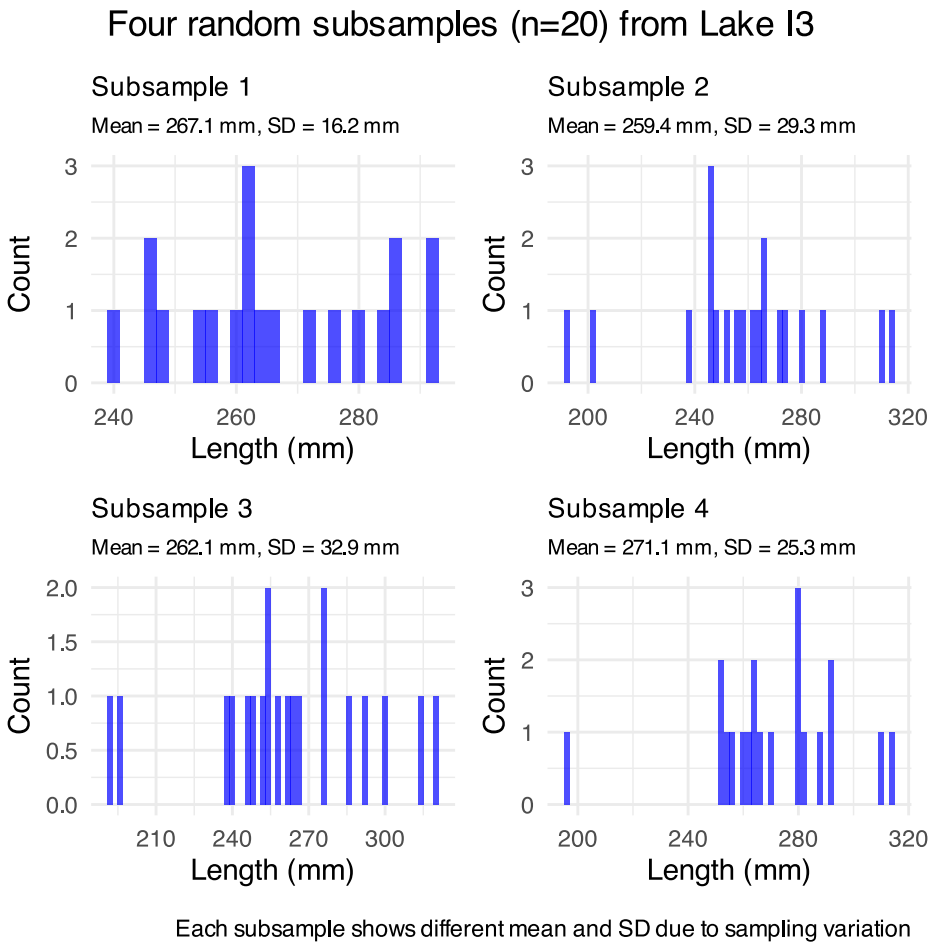
Where:

- s is the sample standard deviation
- n 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

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?

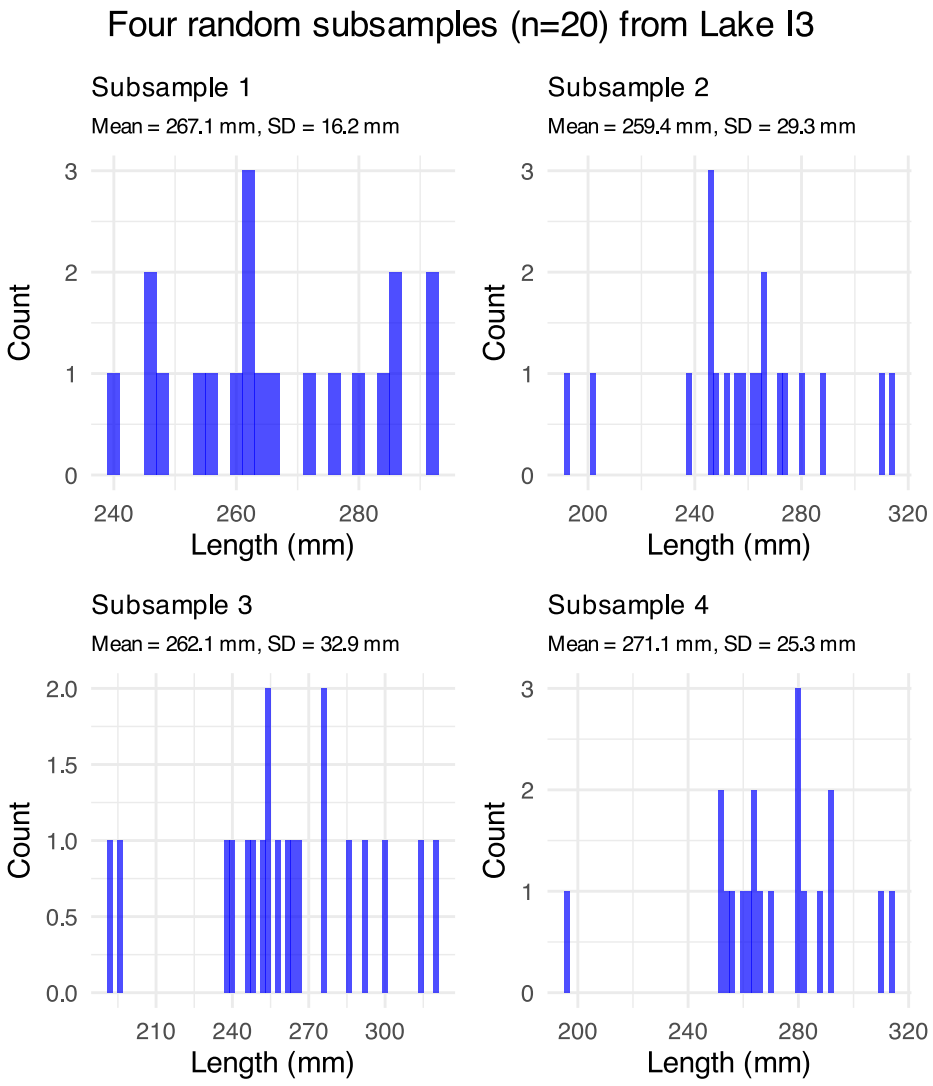
Lecture 4: Estimating μ - population mean

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:

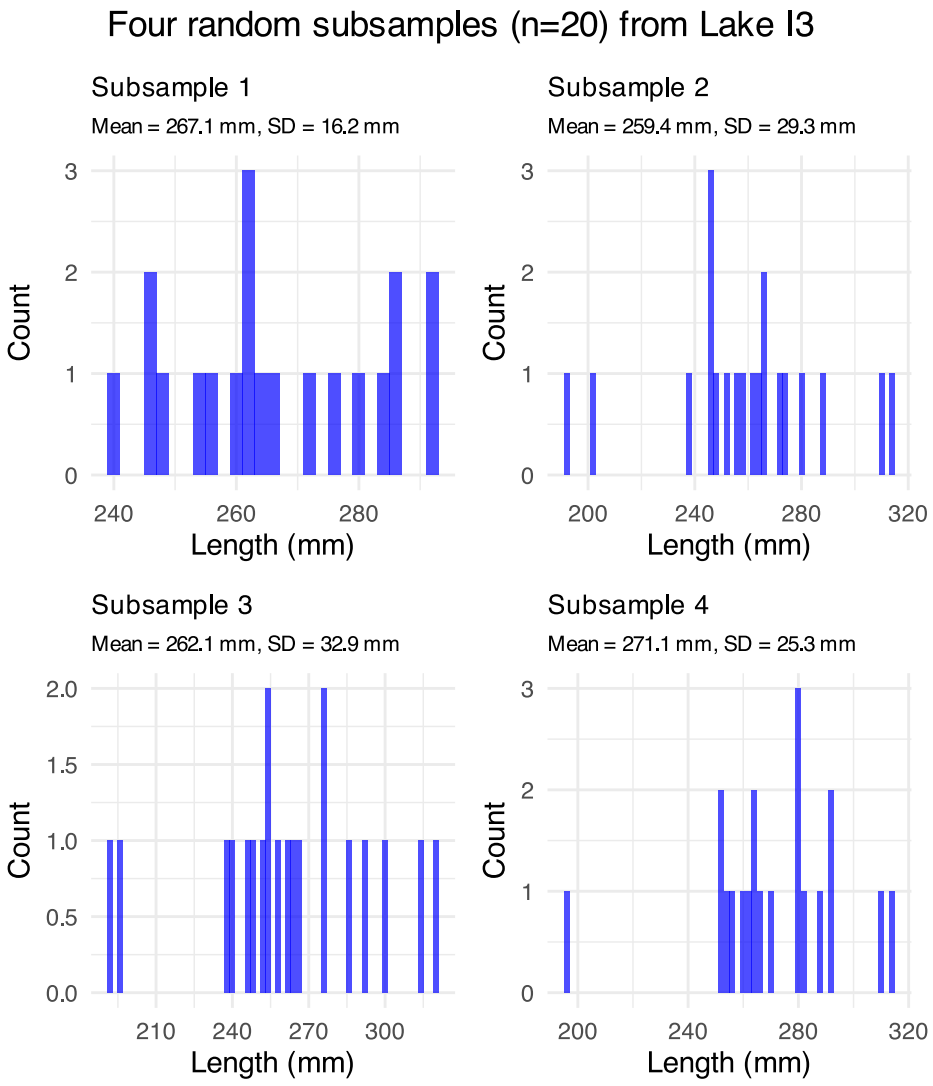
- Sampling distribution of means (SDM) from normal population will be normal
- Large Sampling distribution of means from any population will be normal (Central Limit Theorem)
- The mean of Sampling distribution of means will equal μ or the mean



Lecture 4: Standard Error Properties

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 $\bar{y}$ to estimate μ...



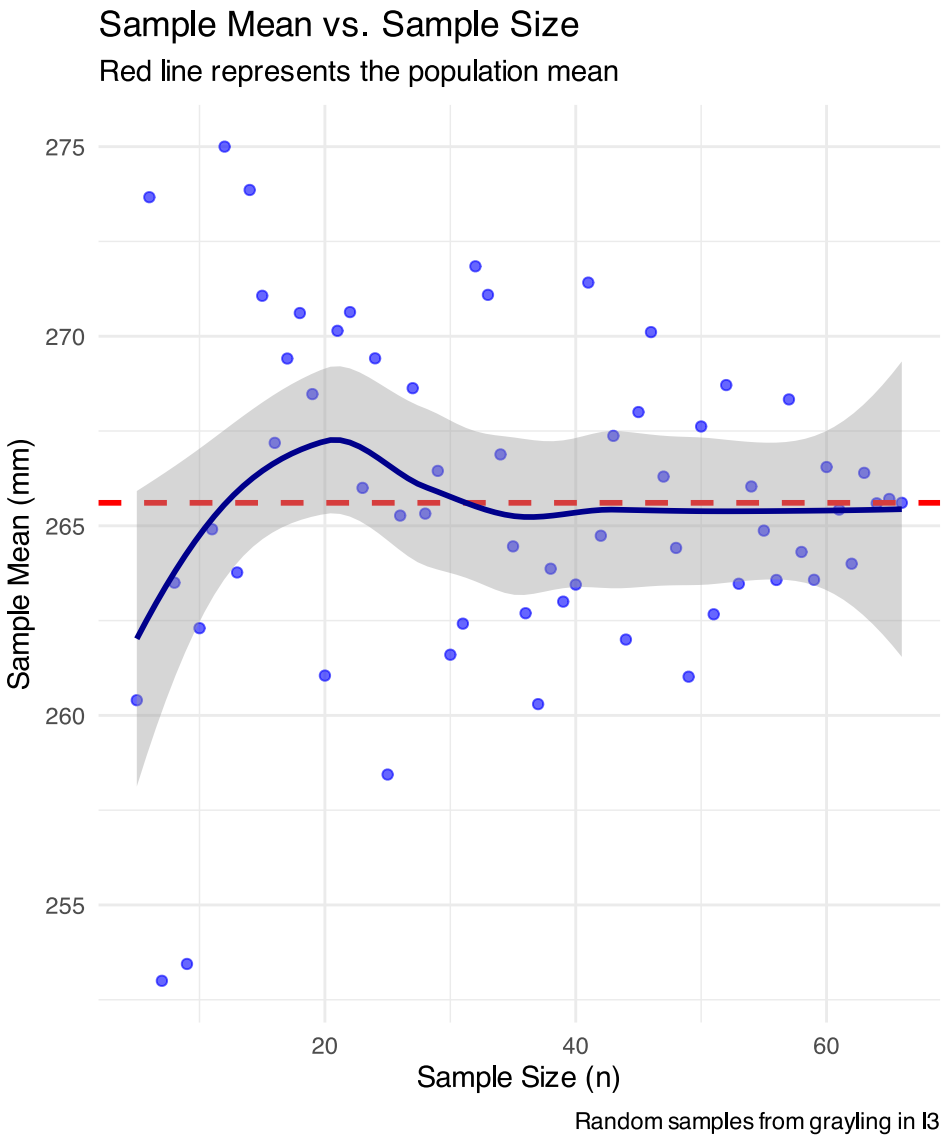
Each subsample shows different mean and SD due to sampling variation

Lecture 4: Standard Error and Sample Size

Notice: - $s_{\bar{y}}$ depends on - sample s (standard deviation) - sample n - ($s_{\bar{y}} = \frac{s}{\sqrt{n}}$)

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

- Large sample, low s = greater confidence in estimate of μ



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:

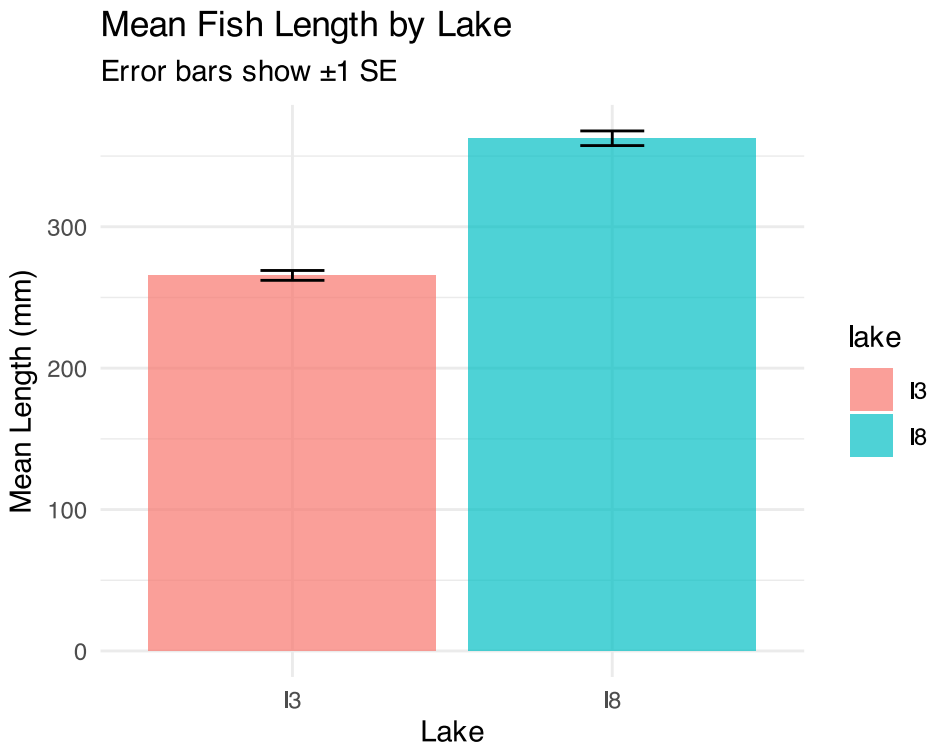
$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

Where:

- s is the sample standard deviation
- n 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 - Basic Formula

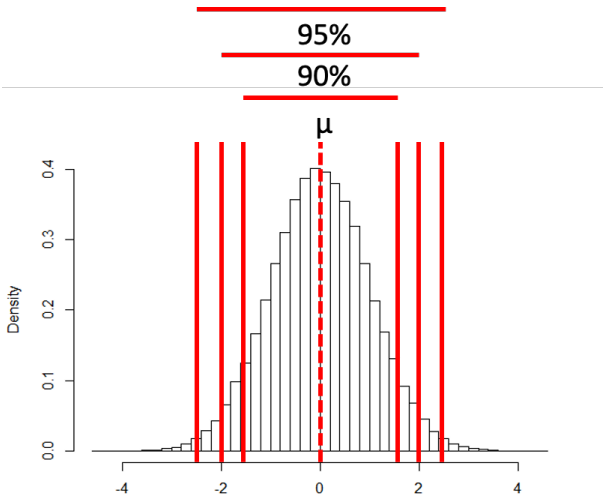
A **confidence interval** is a range of values that is likely to contain the true population parameter.

95% Confidence Interval Formula:

$$95\%\ \text{CI} = \bar{y} \pm z \cdot \frac{\sigma}{\sqrt{n}}$$

Where:

- ȳ is the sample mean
- [?] is the sample size
- σ is the population standard deviation
- z is the z-value corresponding the probability of the CI



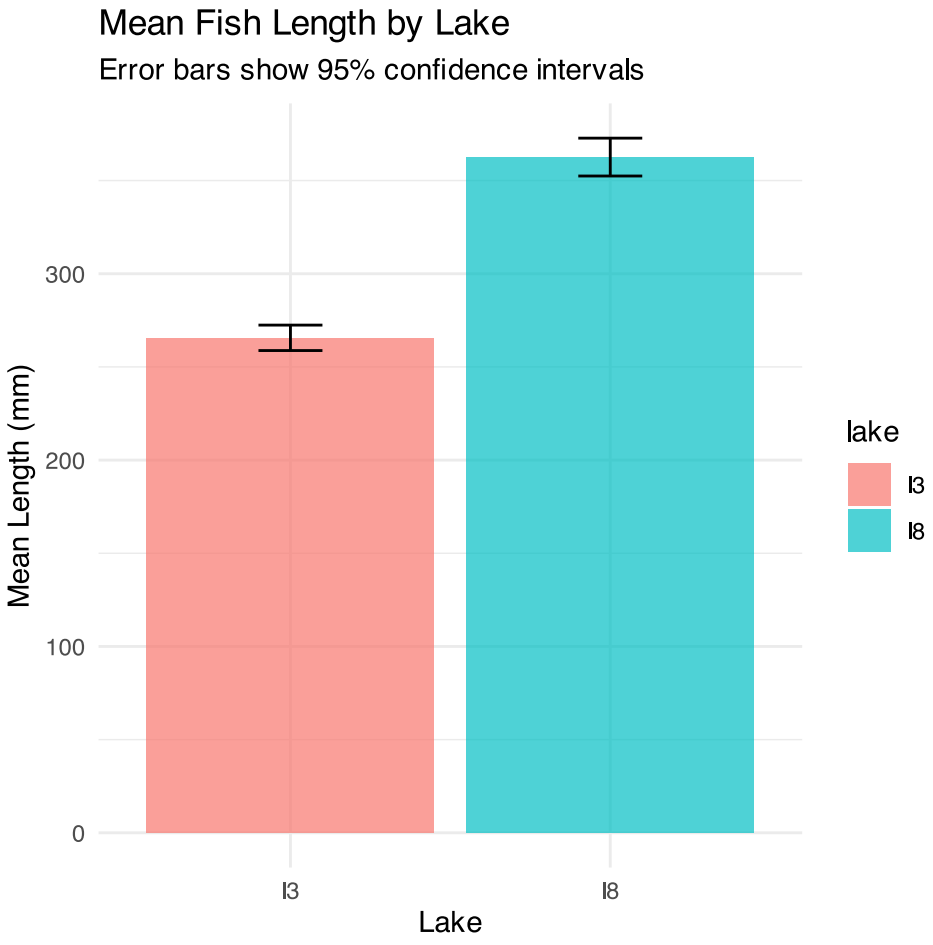
Lecture 4: Confidence Intervals - Interpretation

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.

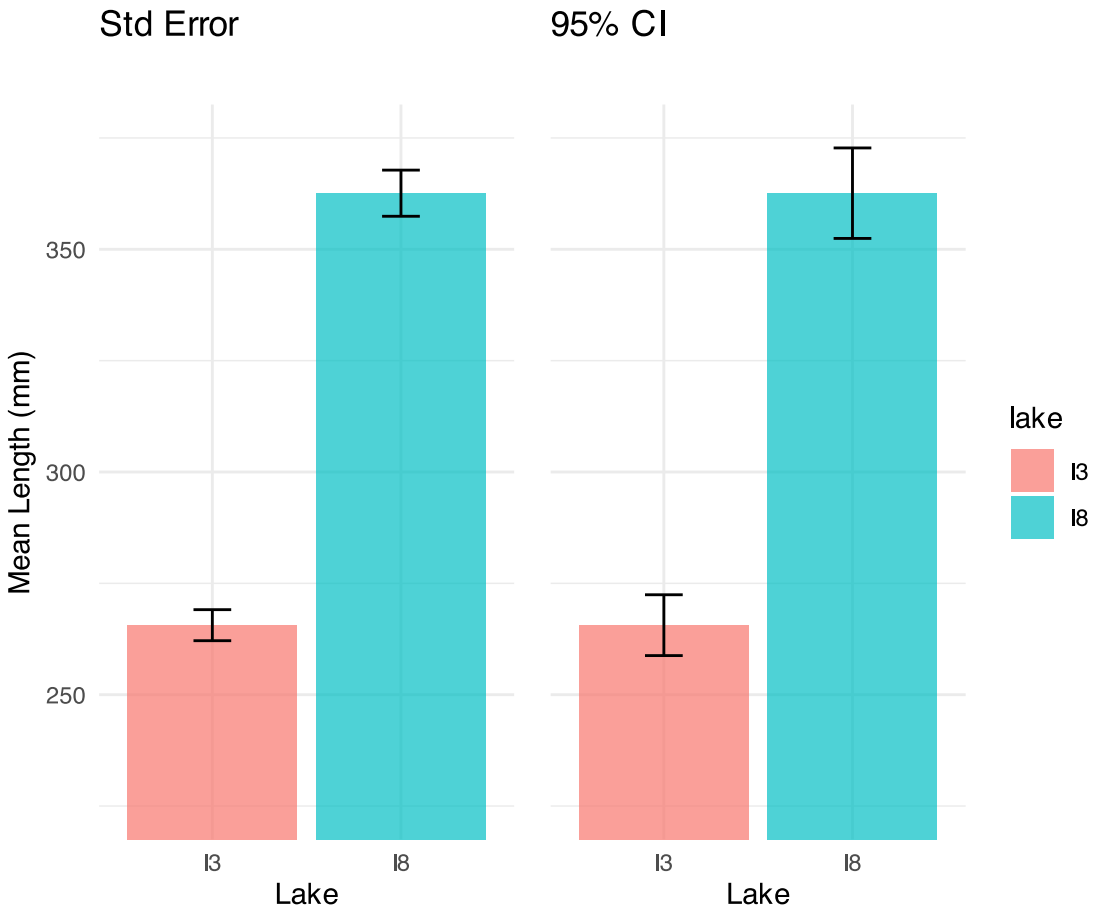
Common misinterpretation: "There is a 95% probability that the true mean is in this interval."

- 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: Standard Error and CI

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.
```

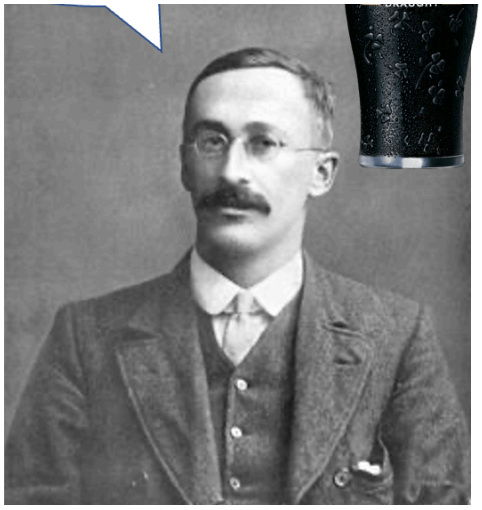
What do these confidence intervals tell us about the difference between lakes?

Lecture 4: When Population σ is Unknown

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

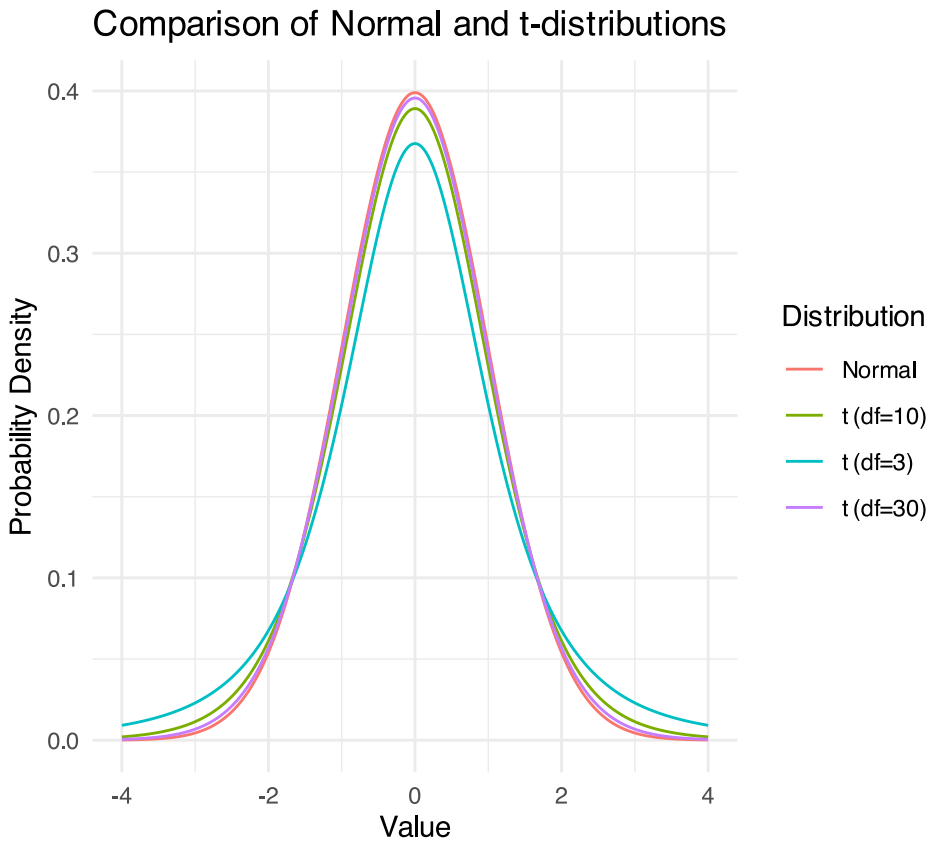
Instead, we use Student's t distribution



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

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 Formula

To calculate CI for sample from "unknown" population:

$$\text{CI} = \bar{y} \pm t \cdot \frac{s}{\sqrt{n}}$$

Where:

- ȳ is sample mean
- [?] 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)

two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Lecture 4: Student's t-distribution Table

Here is a t-table

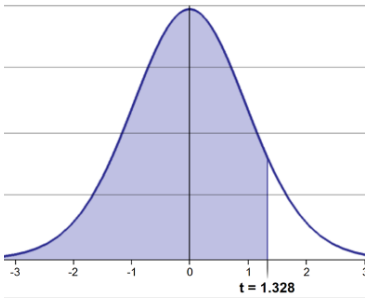
- Values of t that correspond to probabilities
- Probabilities listed along top
- Sample dfs are listed in the left-most column
- Probabilities are given for one-tailed and two-tailed "questions"

two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Lecture 4: One-tailed Questions

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)

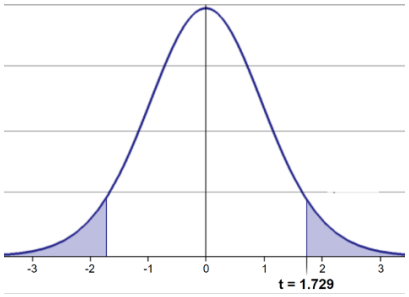


2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390

Lecture 4: Two-tailed Questions

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)



5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551

Lecture 4: t-distribution CI Example

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"

df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	[illegible]	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291

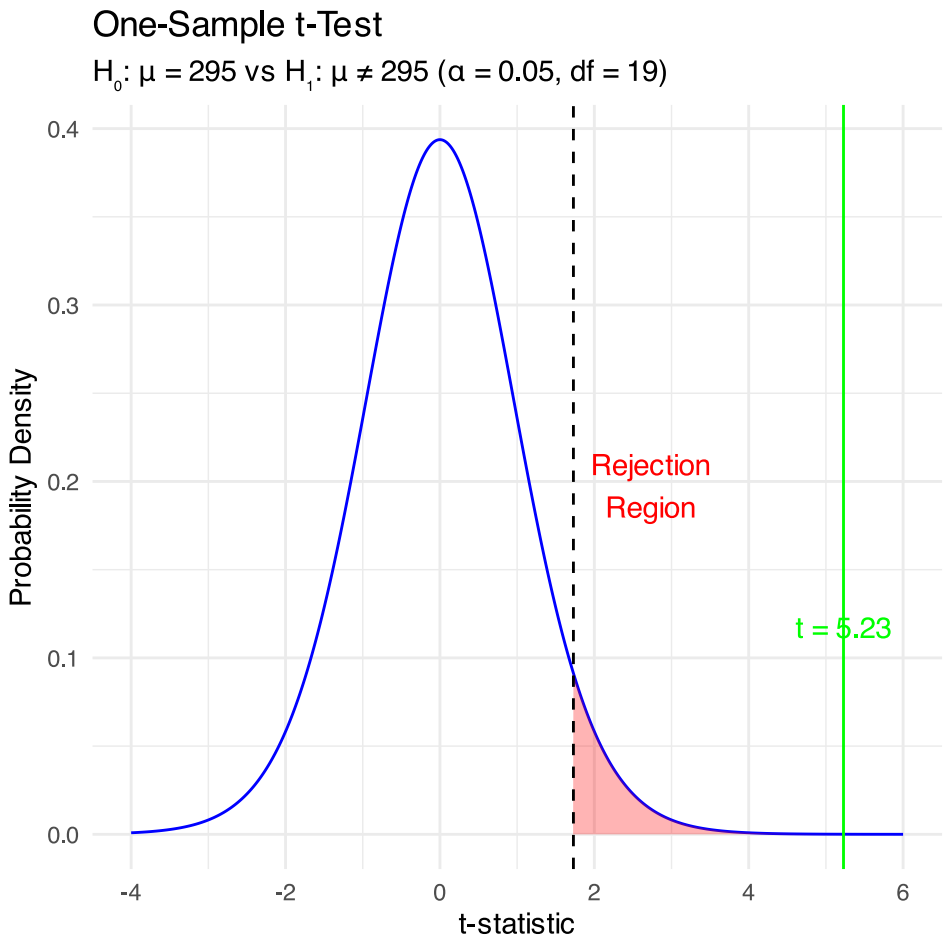
Lecture 4: Intro to Hypothesis Testing

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

Key components:

1. **Null hypothesis (H_0)**: Typically assumes "no effect" or "no difference"
2. **Alternative hypothesis (H_a)**: The claim we're trying to support
3. **Statistical test**: Method for evaluating evidence against H_0
4. **P-value**: Probability of observing our results (or more extreme) if H_0 is true
5. **Significance level (α)**: Threshold for rejecting H_0, typically 0.05

Decision rule: Reject H_0 if p-value < α



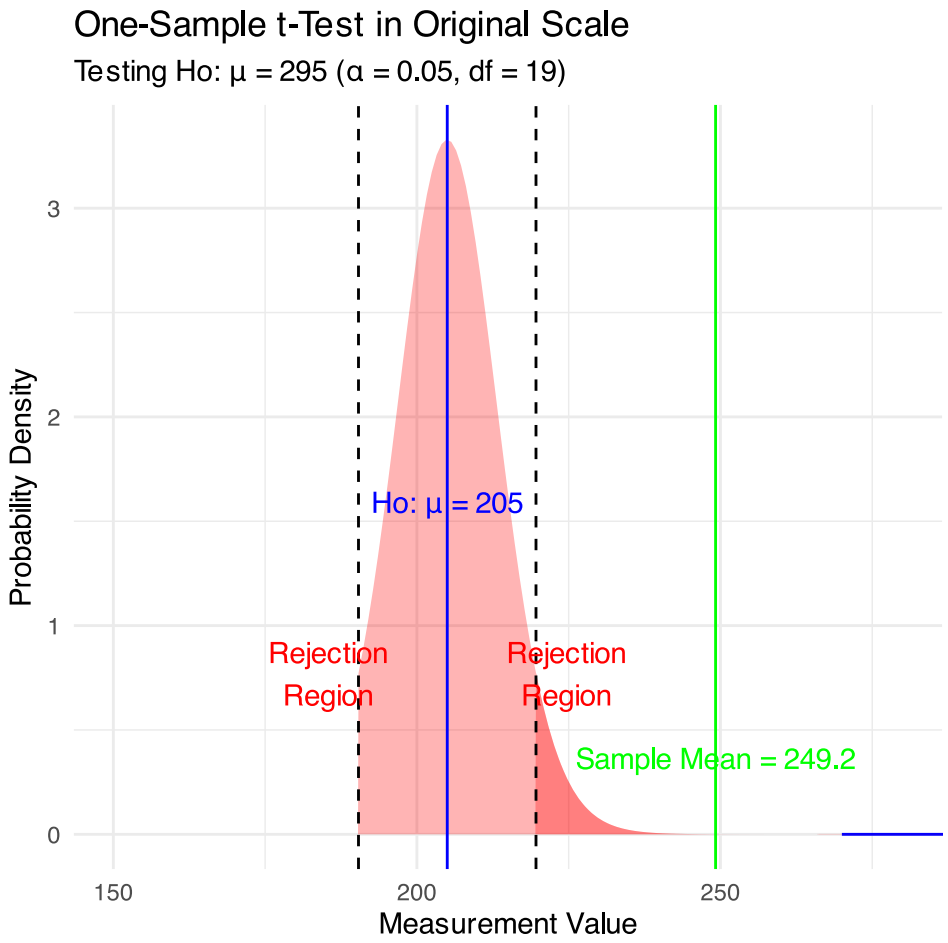
Lecture 4: Hypothesis Testing in Original Scale

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

Key components:

1. **Null hypothesis (H_0)**: Typically assumes "no effect" or "no difference"
2. **Alternative hypothesis (H_a)**: The claim we're trying to support
3. **Statistical test**: Method for evaluating evidence against H_0
4. **P-value**: Probability of observing our results (or more extreme) if H_0 is true
5. **Significance level (α)**: Threshold for rejecting H_0, typically 0.05

Decision rule: Reject H_0 if p-value < α



Practice Exercise 5: One-Sample t-Test

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 I3
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 $\alpha = 0.05$?
5. What is the practical interpretation of this result for fish biologists?

Practice Exercise 6: Formulating Hypotheses

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
```

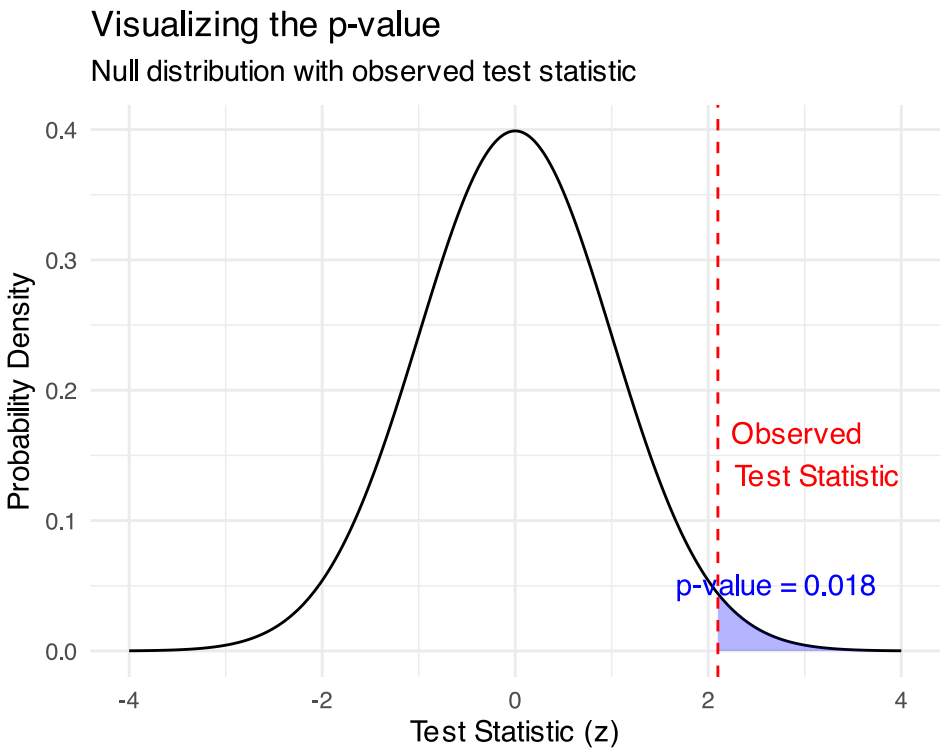
Based on this t-test, what can we conclude about the difference in fish length between the two lakes?

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 H_0 - $0.05 \leq p < 0.10$: Moderate evidence against H_0 - $p \geq 0.10$: Insufficient evidence against H_0

Common misinterpretations: - p-value is NOT the probability that H_0 is true - p-value is NOT the probability that results occurred by chance - Statistical significance ≠ practical significance



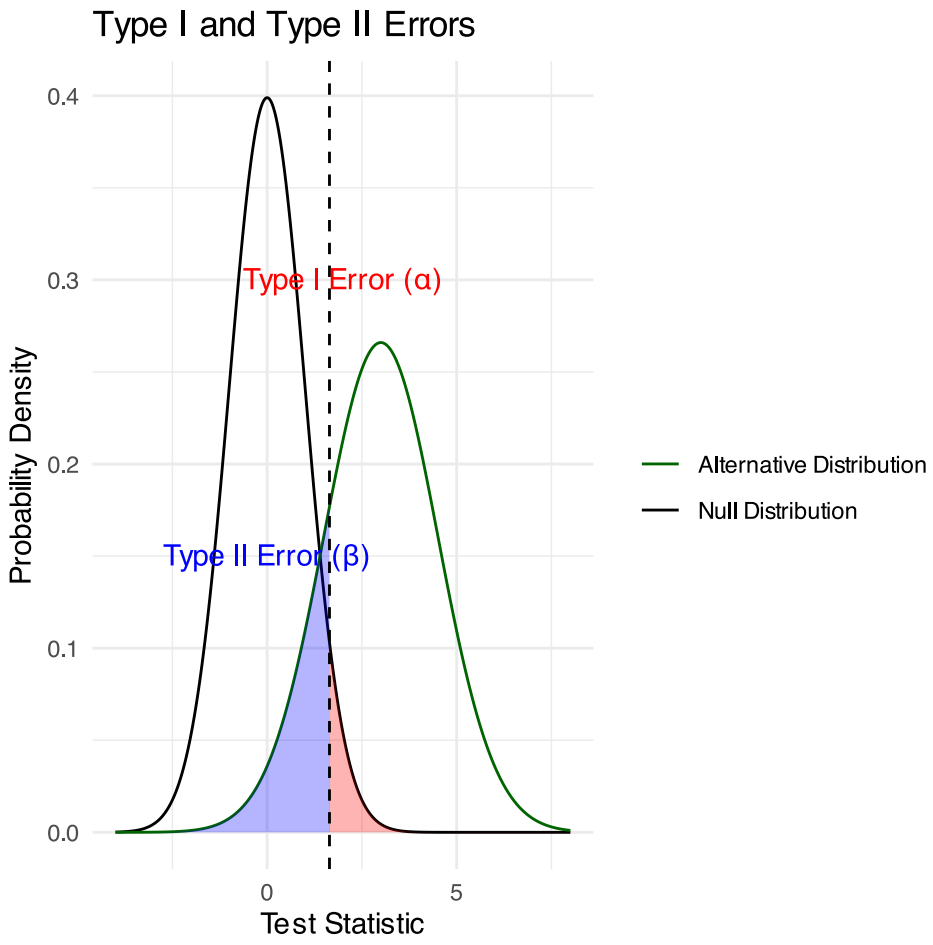
Lecture 4: Type I and Type II Errors

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

Type I Error (False Positive) - Rejecting H_0 when it's actually true - Probability = α (significance level) - "Finding an effect that isn't real"

Type II Error (False Negative) - Failing to reject H_0 when it's actually false - Probability = β - "Missing an effect that is real"

Statistical Power = 1 - β - Probability of correctly rejecting a false H_0 - Increases with: - Larger sample size - Larger effect size - Lower variability - Higher α level



Practice Exercise 7: Interpreting Errors and Power

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 H_0
   - Consider both statistical and practical significance