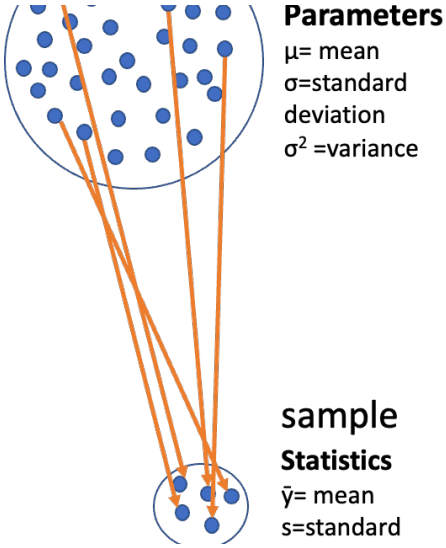
Lecture 05: Probability and Statistical Inference

Bill Perry

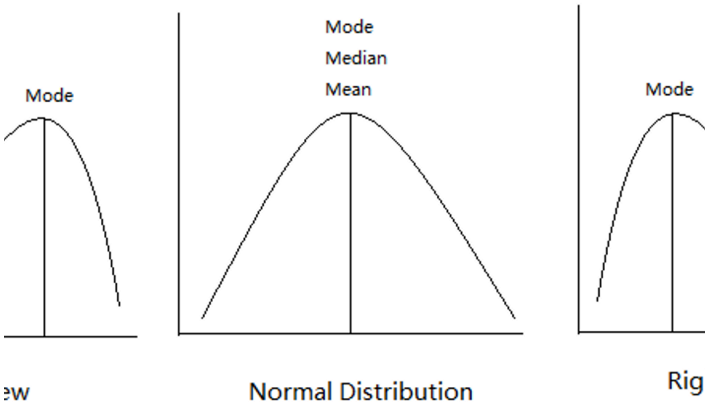
Lecture 4: Review

- Introduction to histograms or frequency distributions
- Probability Distribution Functions (PDF)
- Descriptive Statistics
  - Center - mean, median, mode
  - Spread - range, variance, standard deviation



Lecture 4: Review - Statistical Concepts

- Introduction to histograms or frequency distributions
- Probability Distribution Functions (PDF)
- Descriptive Statistics
  - Center - mean, median, mode
  - Spread - range, variance, standard deviation



Lecture 4: Review - Summary Statistics

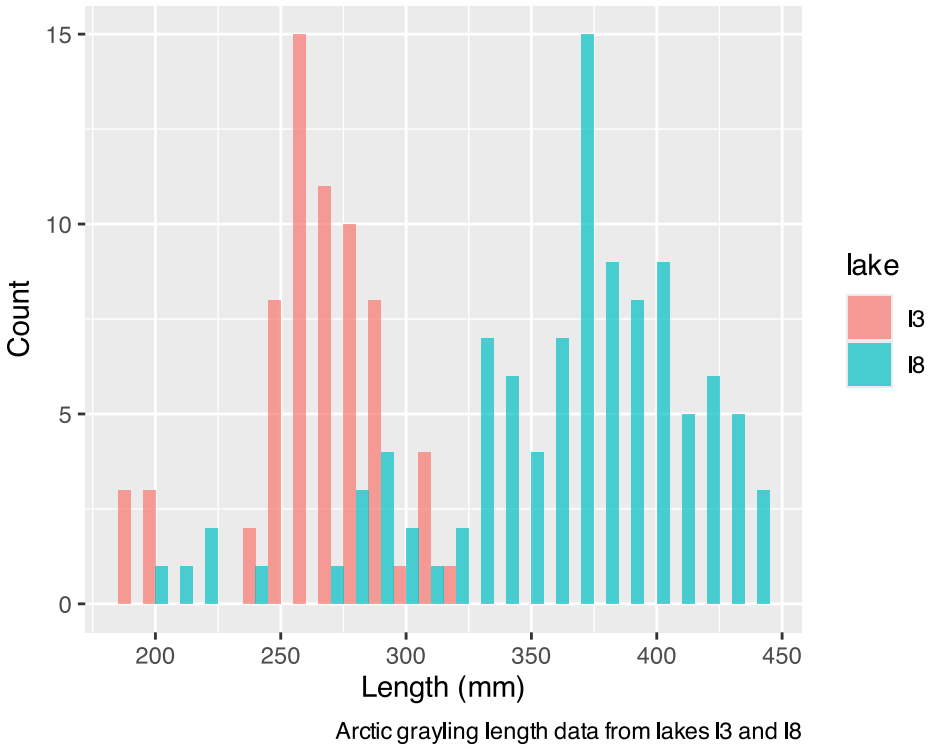
- Introduction to histograms or frequency distributions
- Probability Distribution Functions (PDF)
- Descriptive Statistics
  - Center - mean, median, mode
  - Spread - range, variance, standard deviation

lake	mean_length	sd_length	se_length	count
l3	265.6061	28.30378	3.483954	66
l8	362.5980	52.33901	5.182334	102

Lecture 5: Probability and Statistical Inference

The goals for today

- Statistical inference fundamentals
- Hypothesis testing principles
- T Distributions
- One sample T Tests
- Two sample T Test

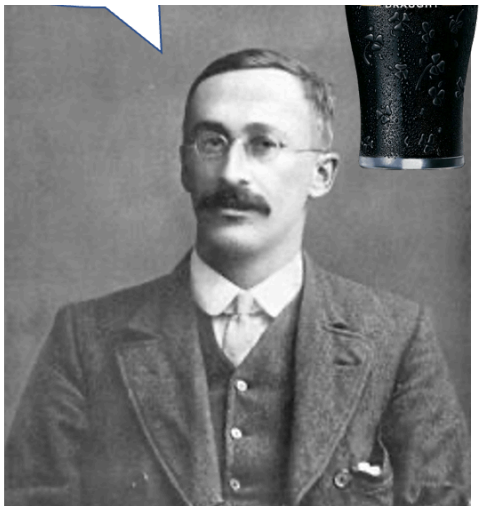


Lecture 5: Confidence intervals

In the more typical case DON'T know the population σ or standard deviation

- estimate it from the samples
- and when sample size is <~30)
- can't use the standard normal (z) distribution

Instead, we use Student's t distribution

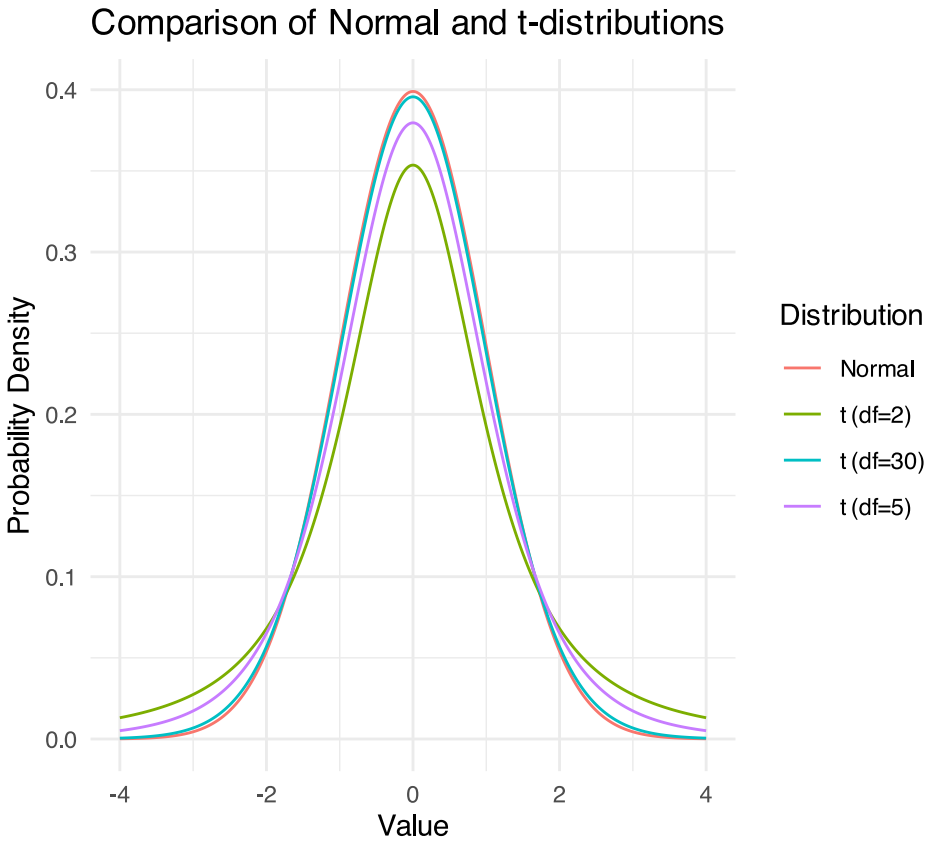


Lecture 5: Understanding t-distribution

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



Lecture 5: t-distribution Properties

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

- Similar to normal distribution (1.96 = 2.5% tails) 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

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

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: 252.8 mm
```

```
cat("Standard deviation:", round(sample_sd, 2), "mm\n")
```

```
Standard deviation: 31.59 mm
```

```
cat("Standard error:", round(sample_se, 2), "mm\n")
```

```
Standard error: 9.99 mm
```

```
cat("95% CI using z:", round(z_ci_lower, 1), "to", round(z_ci_upper, 1), "mm\n")
```

```
95% CI using z: 233.2 to 272.4 mm
```

```
cat("95% CI using t:", round(t_ci_lower, 1), "to", round(t_ci_upper, 1), "mm\n")
```

```
95% CI using t: 230.2 to 275.4 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 5: Student's t-distribution Table

Here is a t-table

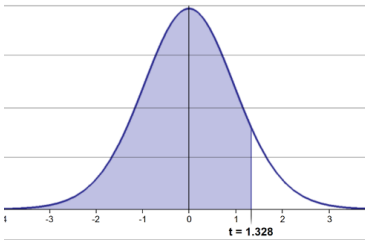
- Values of t that correspond to probabilities
- Probabilities listed along top
- Sample dfs are listed in the left-most column
- Probabilities are given for one-tailed and two-tailed "questions"

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 5: 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)

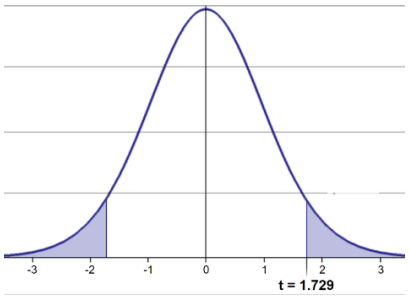


4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	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

Lecture 5: 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 5: Calculating CI Example

Let's calculate CIs again:

Use two-sided test

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

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 5: Applications of t-distribution

So:

- Can assess confidence that population mean is within a certain range
- Can use t distribution to ask questions like:
  - "What is probability of getting sample with mean = $\bar{y}$ from population with mean = μ?" (1 sample t-test)
  - "What is the probability that two samples came from same population?" (2 sample t-test)

Lecture 5: Single Sample T-Test

We want to test if the mean fish length in I3 differs from 240mm.

Activity: Define hypotheses and identify assumptions

H_0: $\mu = 240$ (The mean fish length in I3 is 240mm)

H_1: $\mu \neq 240$ (The mean fish length in I3 is not 240mm)

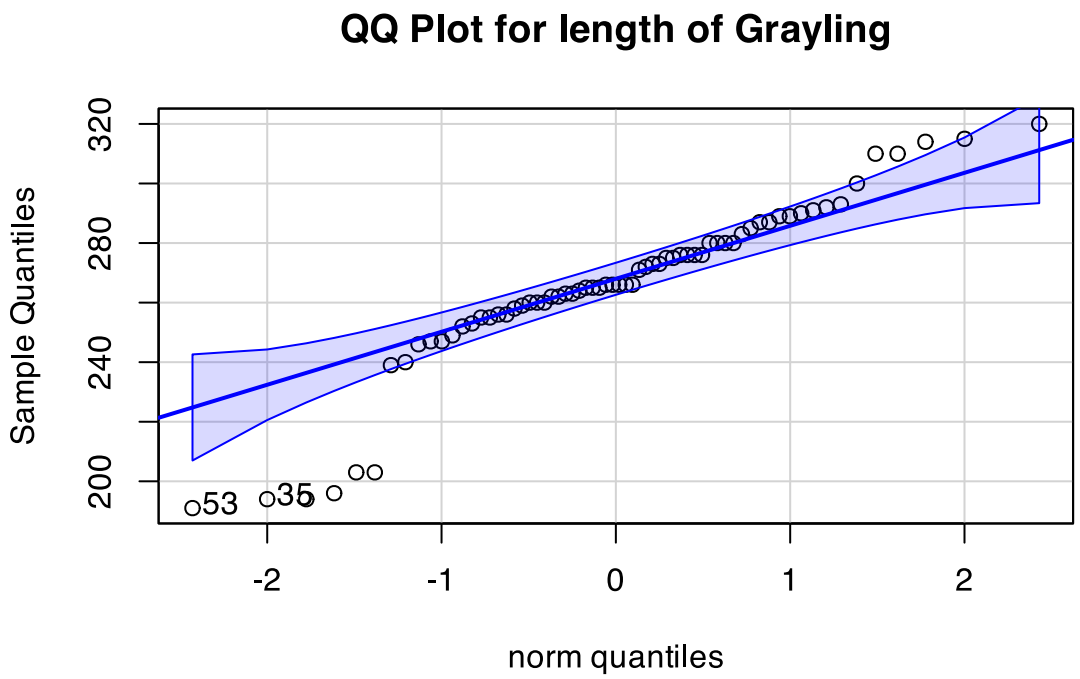
Assumptions for t-test:

1. Data is normally distributed
2. Observations are independent
3. No significant outliers

Assumptions in R - qqplots from car

```
# Filter for just windward side needles

# YOUR TASK: Test normality of windward pine needle lengths
# QQ Plot
qqPlot(i3_df$length_mm,
       main = "QQ Plot for length of Grayling",
       ylab = "Sample Quantiles")
```




```
[1] 53 35
```

Statistical Test of Normality

Shapiro-Wilk test

```
# Shapiro-Wilk test
shapiro_test <- shapiro.test(i3_df$length_mm)
print(shapiro_test)
```

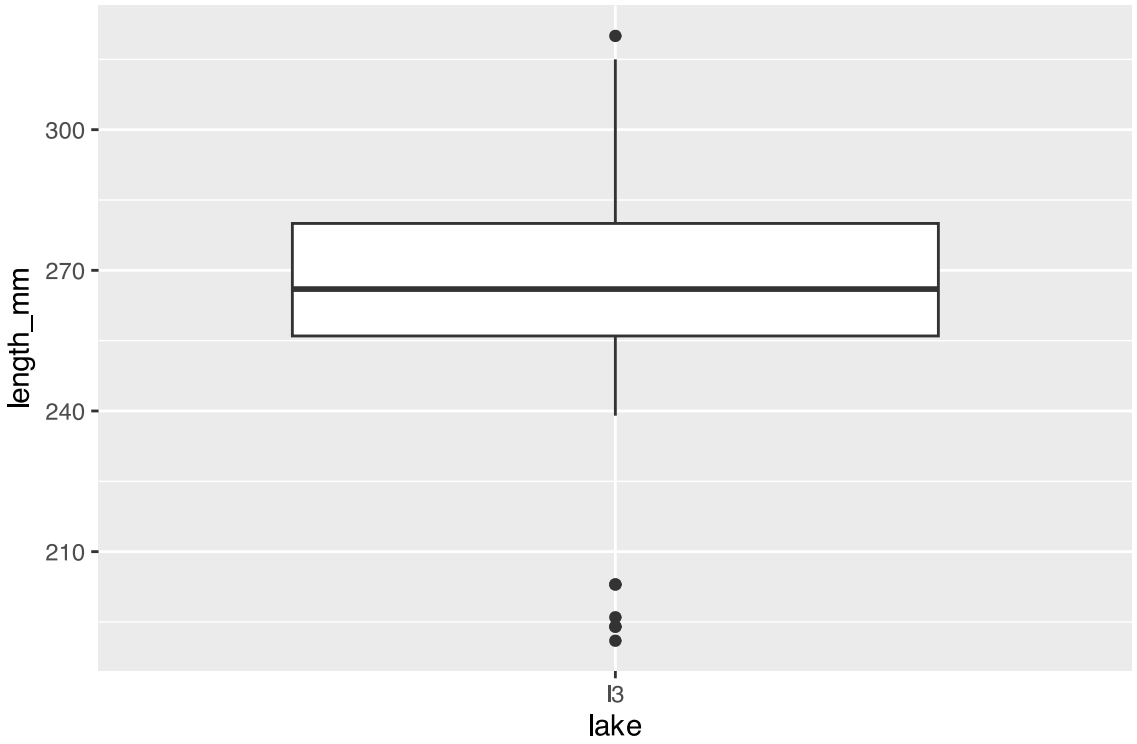
```

    Shapiro-Wilk normality test

data:  i3_df$length_mm
W = 0.91051, p-value = 0.0001623
```

Checking for Outliers

```
# Check for outliers using boxplot
# YOUR CODE HERE
i3_df %>% ggplot(aes(lake, length_mm))+geom_boxplot()
```



Practice Exercise 1: One-Sample t-Test

Practice Exercise 1: One-Sample t-Test

Let's perform a one-sample t-test to determine if the mean fish length in I3 Lake differs from 240 mm:

```
# 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 = 240)

# View the test results
t_test_result
```

```

    One Sample t-test

data:  i3_df$length_mm
t = 7.3497, df = 65, p-value = 4.17e-10
alternative hypothesis: true mean is not equal to 240
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?

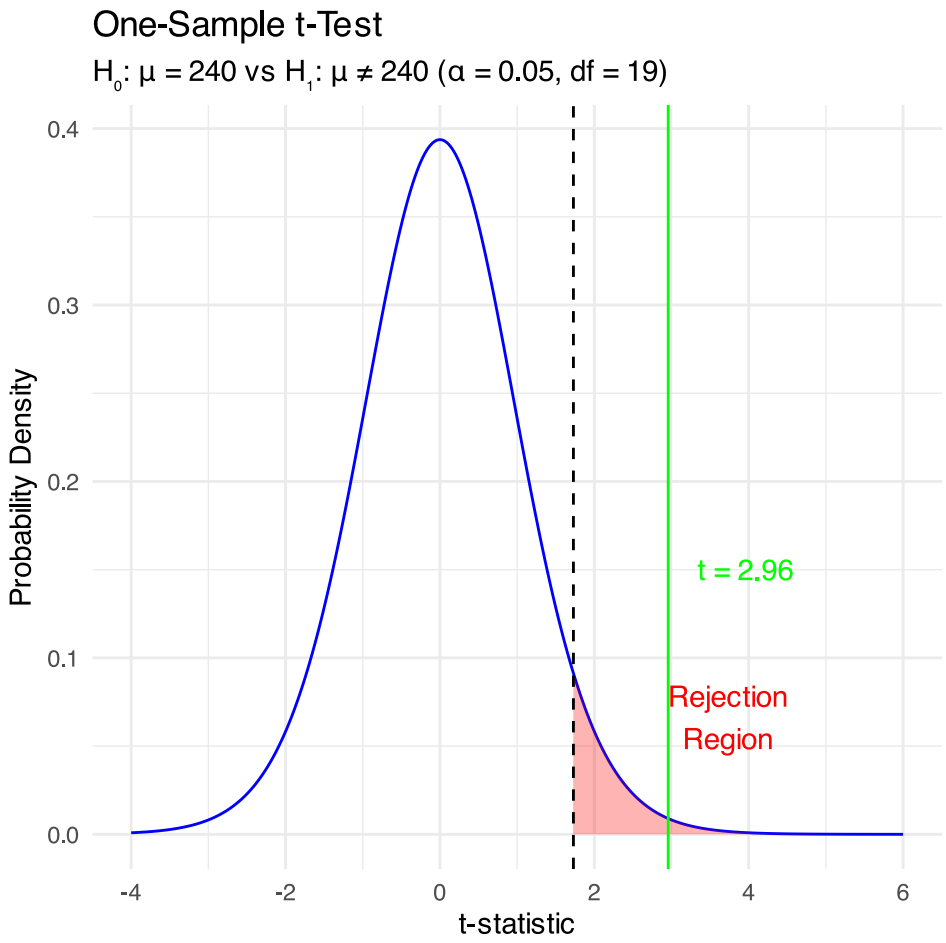
Lecture 5: Hypothesis Testing Framework

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

Key components:

1. **Null hypothesis (Ho)**: 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 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 Ho if p-value < α == p < 0.05



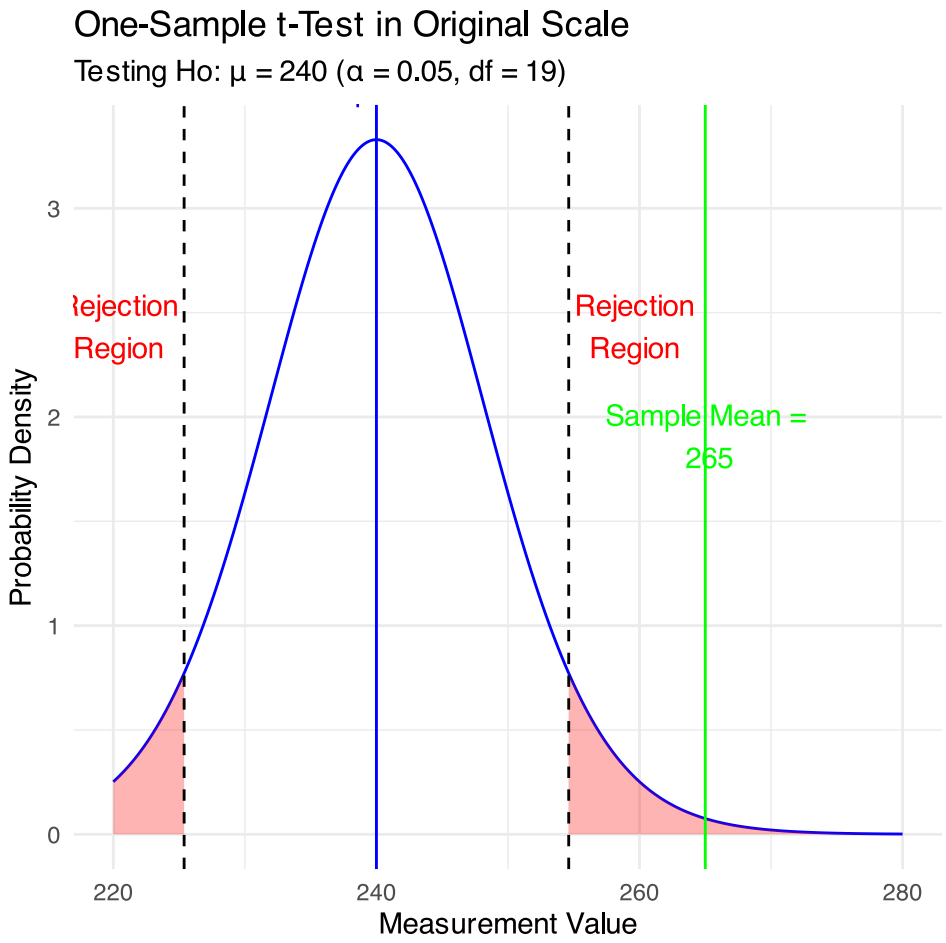
Lecture 5: Hypothesis Testing - Original Scale

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



Lecture 5: Interpreting One-Sample T-Test Results

Activity: Interpret the t-test results

- What does the p-value tell us?
- Should we reject or fail to reject the null hypothesis?

How to report this result in a scientific paper:

"A two-tailed, one-sample t-test at α=0.05 showed that the mean pine needle length on the windward side (... mm, SD = ...) [was/was not] significantly different from the expected 55 mm, t(...) = ..., p = ..."

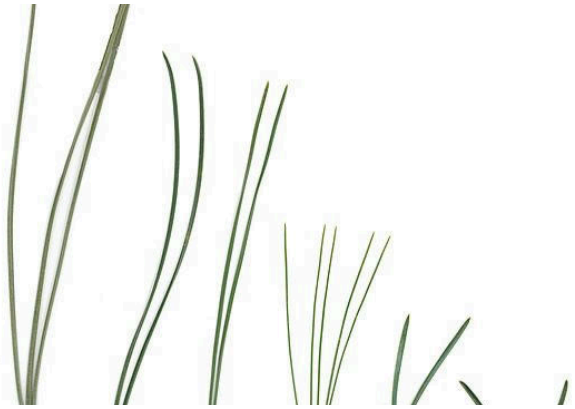
Lecture 5: Two Sample T-Tests Introduction

For example

- what is probability that population X is the same as population Y?

How would you assess this question using what we learned?

This is what we will do with the pine needles...



Lecture 5: Comparing Two Samples

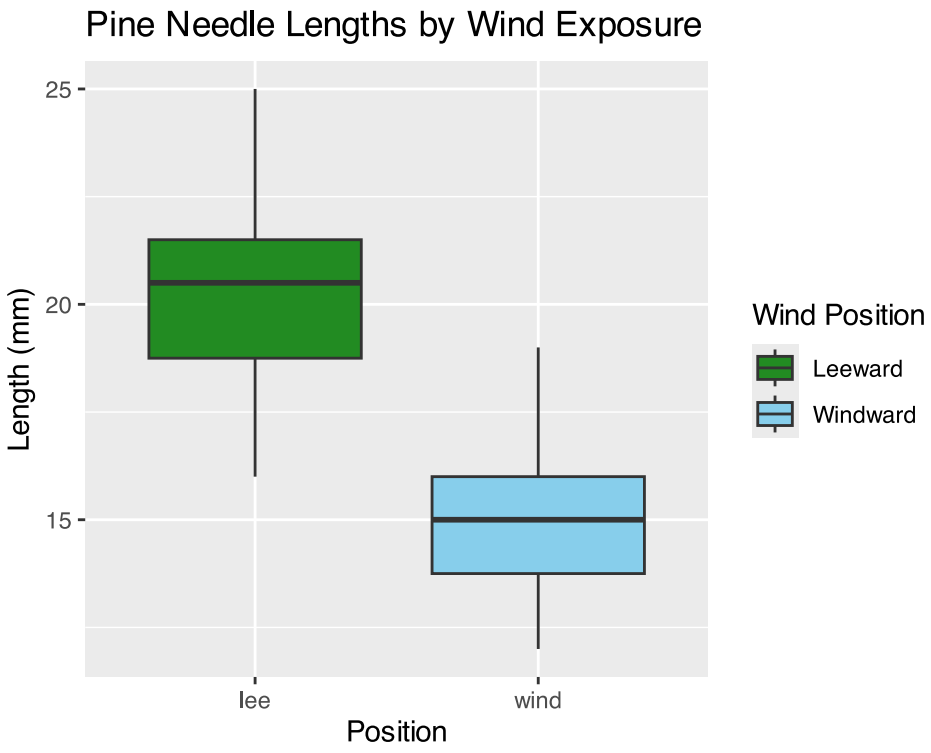
For example

- what is probability that population X is the same as population Y?

How would you assess this question using what we learned?

```
# Now create a boxplot to visualize the difference in fish lengths between these lakes:
pine_df <- read_csv("data/pine_needles.csv")

# Create a boxplot comparing the two lakes
pine_wind_plot <- pine_df %>%
  ggplot(aes(x = wind, y = length_mm, fill = wind)) +
  geom_boxplot() +
  labs(title = "Pine Needle Lengths by Wind Exposure",
       x = "Position",
       y = "Length (mm)",
       fill = "Wind Position") +
  scale_fill_manual(values = c("lee" = "forestgreen", "wind" = "skyblue"),
                    labels = c("lee" = "Leeward", "wind" = "Windward"))
pine_wind_plot
```




```
# Based on the t-test results and the boxplot
#
# what can you conclude about the fish populations in these two lakes?
```

Practice Exercise 2: Formulating Hypotheses

Practice Exercise 2: Formulating Hypotheses

For the following research questions about pine needles write the null and alternative hypotheses:

1. Are needles on the lee side longer than the needles on the windy side?

What are the hypotheses?

Ho =

Ha =

Lecture 5: Two-Sample T-Test Framework

Now, let's compare pine needle lengths between windward and leeward sides of trees.

Question: **Is there a significant difference in needle length between the windward and leeward sides?**

This requires a two-sample t-test.

Two-sample t-test compares means from two independent groups.

$$t = \frac{\bar{x_1} - \bar{x_2}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where:

- $\bar{x}_1$ and $\bar{x}_2$: These represent the sample means of the two groups you're comparing

- s^2_p: This is the pooled variance, calculated as: $s^2_p = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2] / (n_1 + n_2 - 2)$, where s_1^2 and s_2^2 are the sample variances of the two groups.
- **n_1 and n_2:** These are the sample sizes of the two groups.
- **$\sqrt{(1/n_1 + 1/n_2)}$:** This represents the pooled standard error.

$$t = \frac{SIGNAL}{NOISE}$$

Practice Exercise 3: Summary Statistics

Practice Exercise 3: **Calculate summary statistics grouped by wind exposure**

Before conducting the test, we need to understand the data for each group.

1. You need this and the graph to see what is going on ....

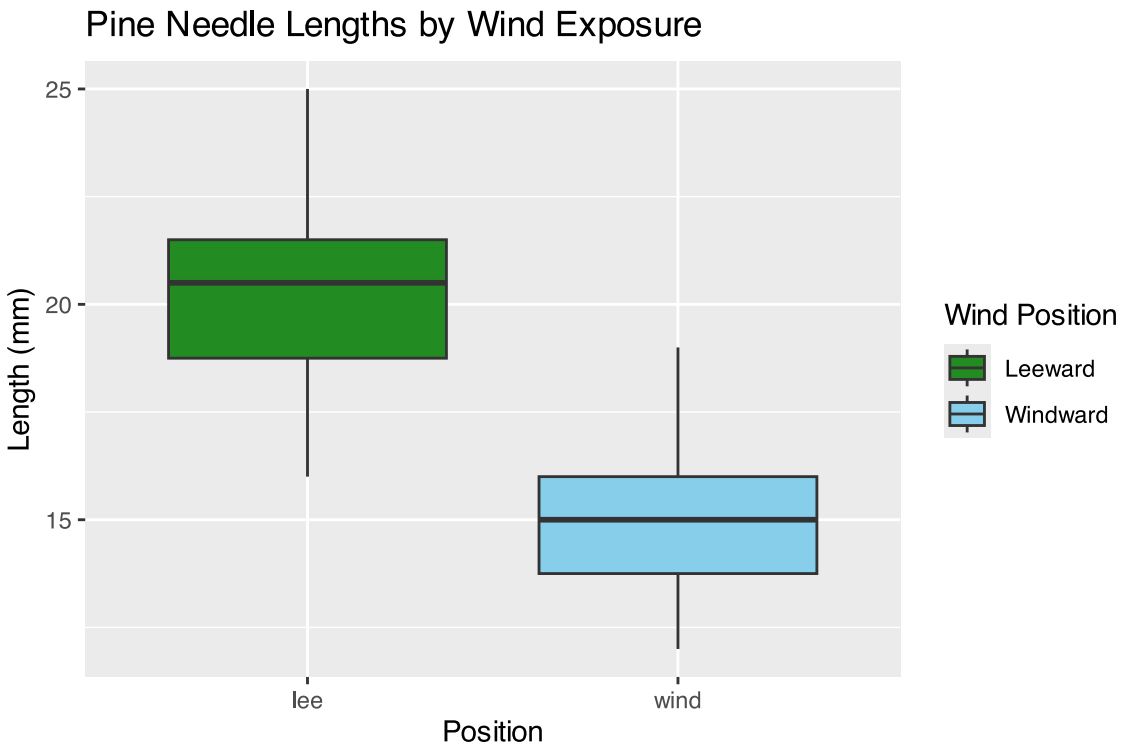
```
group_summary <- pine_df %>%
  group_by(wind) %>%
  summarize(
    mean_length = mean(length_mm),
    sd_length = sd(length_mm),
    n = n(),
    se_length = sd_length / sqrt(n)
  )

print(group_summary)
```

```
# A tibble: 2 × 5
  wind  mean_length sd_length     n se_length
  <chr>       <dbl>     <dbl> <int>     <dbl>
1 lee          20.4      2.45    24     0.500
2 wind         14.9      1.91    24     0.390
```

Visualizing Group Differences

```
# Create a boxplot comparing the two sides
pine_wind_plot
```



Practice Exercise 4: Effect Size

Practice Exercise 4: Effect size

We could also look at the difference in means... some cool code here

```
# Assuming your dataframe is called df
group_summary %>%
  summarize(difference = mean_length[wind == "wind"] - mean_length[wind == "lee"])
```

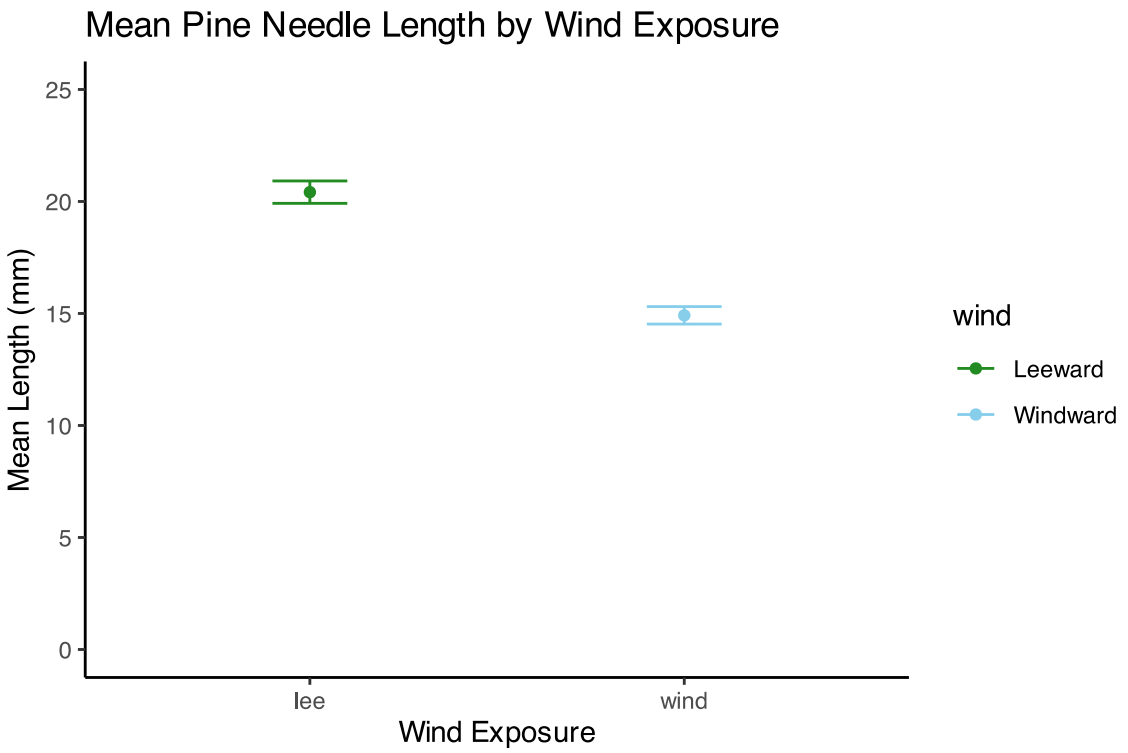
```
# A tibble: 1 × 1
  difference
       <dbl>
1       -5.5
```

Practice Exercise 5: ggplot Summary Statistics

Practice Exercise 5: Using GGPLOT to get summary stats

GGplot also has code to make the mean and standard error plots we are interested in along whit a lot of others

```
# Assuming your dataframe is called df
pine_mean_se_plot <- ggplot(pine_df, aes(x = wind, y = length_mm, color = wind)) +
  stat_summary(fun = mean, geom = "point") +
  stat_summary(fun.data = mean_se, geom = "errorbar", width = 0.2) +
  labs(title = "Mean Pine Needle Length by Wind Exposure",
       x = "Wind Exposure",
       y = "Mean Length (mm)") +
  coord_cartesian(ylim = c(0,25))+
  scale_color_manual(values = c("lee" = "forestgreen", "wind" = "skyblue"),
                     labels = c("lee" = "Leeward", "wind" = "Windward"))+
  theme_classic()
pine_mean_se_plot
```



Lecture 5: Testing Assumptions for Two-Sample T-Test

For a two-sample t-test, we need to check:

1. Normality within each group
2. Equal variances between groups (for standard t-test)
3. Independent observations

If assumptions are violated:

- Welch's t-test (unequal variances)
- Non-parametric alternatives (Mann-Whitney U test)

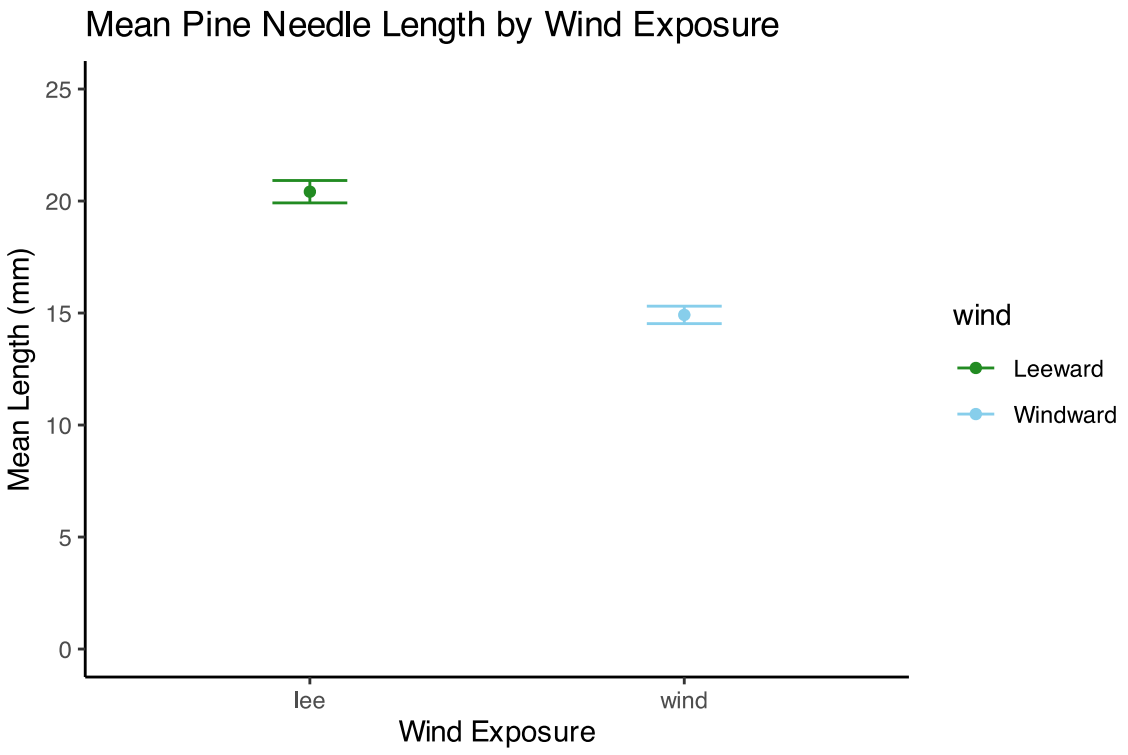
Practice Exercise 6: Creating Group Data

Practice Exercise 6: Test normality of windward pine needle lengths

qqplots

Note you need to test each groups separately...

```
# Assuming your dataframe is called df
pine_mean_se_plot
```



Practice Exercise 7: Separate Group Data

Practice Exercise 7: Test normality of windward pine needle lengths

qqplots

Note you need to test each groups separately...

```
# how do you make separate dataframes to do this on?
# Separate data by groups
windward_data <- pine_df %>% filter(wind == "wind")
leeward_data <- pine_df %>% filter(wind == "lee")
head(leeward_data)
```

```
# A tibble: 6 × 6
  date    group       n_s   wind  tree_no length_mm
  <chr>   <chr>       <chr> <chr>   <dbl>     <dbl>
1 3/20/25 cephalopods n     lee         1        20
2 3/20/25 cephalopods n     lee         1        21
3 3/20/25 cephalopods n     lee         1        23
4 3/20/25 cephalopods n     lee         1        25
5 3/20/25 cephalopods n     lee         1        21
6 3/20/25 cephalopods n     lee         1        16
```

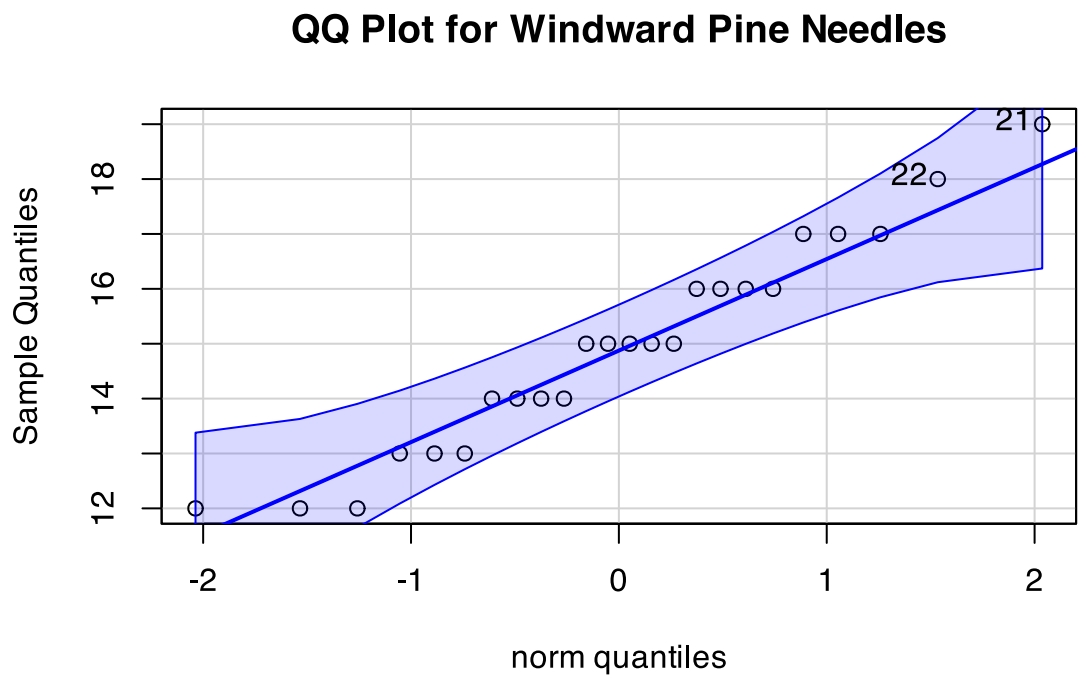
Practice Exercise 8: QQ Plot for Windward Data

Practice Exercise 8: Test normality of windward pine needle lengths

qqplots

Note you need to test each groups separately...

```
# QQ Plot for windward group
qqPlot(windward_data$length_mm,
       main = "QQ Plot for Windward Pine Needles",
       ylab = "Sample Quantiles")
```




```
[1] 21 22
```

Practice Exercise 9: Shapiro-Wilk Test

Practice Exercise 9: Test normality of windward pine needle lengths

Shapiro-Wilk test

Note you need to test each groups separately…

```
# Shapiro-Wilk test for windward group
shapiro_windward <- shapiro.test(windward_data$length_mm)
print("Shapiro-Wilk test for windward data:")
```

```
[1] "Shapiro-Wilk test for windward data:"
```

```
print(shapiro_windward)
```

```

    Shapiro-Wilk normality test

data:  windward_data$length_mm
W = 0.96062, p-value = 0.451
```

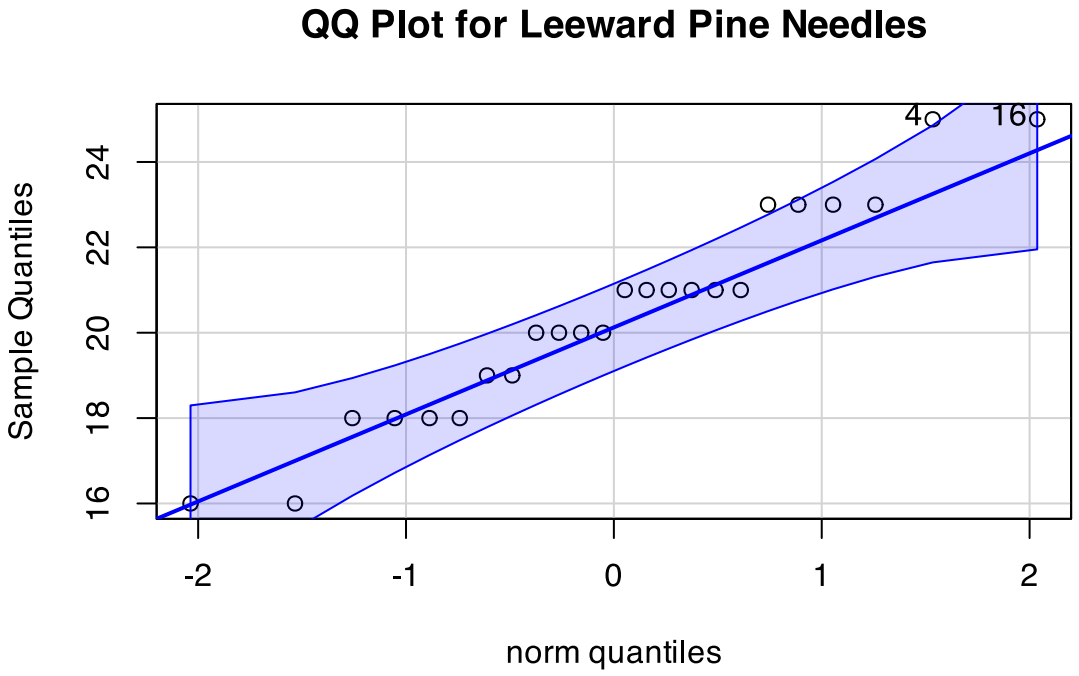
Practice Exercise 10: QQ Plot for Leeward Data

Practice Exercise 10: Test normality of leeward pine needle lengths

qqplots

Note you need to test each groups separately...

```
# You can also test the leeward group
# QQ Plot for leeward group
qqPlot(leeward_data$length_mm,
       main = "QQ Plot for Leeward Pine Needles",
       ylab = "Sample Quantiles")
```




```
[1]  4 16
```

Practice Exercise 11: Shapiro-Wilk for Leeward

Practice Exercise 11: Test normality of leeward pine needle lengths

Shapiro-Wilk test

Note you need to test each groups separately...

```
# Shapiro-Wilk test for leeward group
shapiro_lee <- shapiro.test(leeward_data$length_mm)
print("Shapiro-Wilk test for leeward data:")
```

```
[1] "Shapiro-Wilk test for leeward data:"
```

```
print(shapiro_lee)
```

```

    Shapiro-Wilk normality test

data:  leeward_data$length_mm
W = 0.95477, p-value = 0.3425
```

Practice Exercise 12: Combined Normality Test

Practice Exercise 12: Test Normality at one time

There are always a lot of ways to do this in R

```
# there are always two ways
# Test for normality using Shapiro-Wilk test for each wind group
# All in one pipeline using tidyverse approach
normality_results <- pine_df %>%
  group_by(wind) %>%
  summarize(
    shapiro_stat = shapiro.test(length_mm)$statistic,
    shapiro_p_value = shapiro.test(length_mm)$p.value,
    normal_distribution = if_else(shapiro_p_value > 0.05, "Normal", "Non-normal")
  )

# Print the results
print(normality_results)
```

```
# A tibble: 2 × 4
  wind  shapiro_stat shapiro_p_value normal_distribution
  <chr>        <dbl>           <dbl> <chr>
1 lee          0.955           0.343 Normal
2 wind         0.961           0.451 Normal
```

Practice Exercise 13: Test Equal Variances

Practice Exercise 13: Test equal variances

Levenes test can be done on the original dataframe

```
# Method 1: Using car package's leveneTest
# This is often preferred as it's more robust to departures from normality
levene_result <- leveneTest(length_mm ~ wind, data = pine_df)
print("Levene's Test for Homogeneity of Variance:")
```

```
[1] "Levene's Test for Homogeneity of Variance:"
```

```
print(levene_result)
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  1.2004 0.2789
      46
```

Lecture 5: Conducting the Two-Sample T-Test

Now we can compare the mean pine needle lengths between windward and leeward sides.

Ho: $\mu_1 = \mu_2$ (The mean needle lengths are equal)

Ha: $\mu_1 \neq \mu_2$ (The mean needle lengths are different)

Deciding between:

- Standard t-test (equal variances)
- Welch's t-test (unequal variances)

Note the Levenes Test should be NOT SIGNIFICANT - What is the null hypothesis

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  1.2004 0.2789
      46
```

Lecture 5: Running the Two-Sample T-Test

Now we can do a two sample TTEST

Calculate t-statistic manually (optional)

YOUR CODE HERE:

t = (mean1 - mean2) / sqrt((s1^2/n1) + (s2^2/n2))

Tip

```
# YOUR TASK: Conduct a two-sample t-test
# Use var.equal=TRUE for standard t-test or var.equal=FALSE for Welch's t-test

# Standard t-test (if variances are equal)
t_test_result <- t.test(length_mm ~ wind, data = pine_df, var.equal = TRUE)
print("Standard two-sample t-test:")
```

```
[1] "Standard two-sample t-test:"
```

```
print(t_test_result)
```

```

    Two Sample t-test

data:  length_mm by wind
t = 8.6792, df = 46, p-value = 3.01e-11
alternative hypothesis: true difference in means between group lee and group wind is not
equal to 0
95 percent confidence interval:
 4.224437 6.775563
sample estimates:
 mean in group lee mean in group wind
          20.41667           14.91667
```

```
# Welch's t-test (if variances are unequal)
# YOUR CODE HERE
```

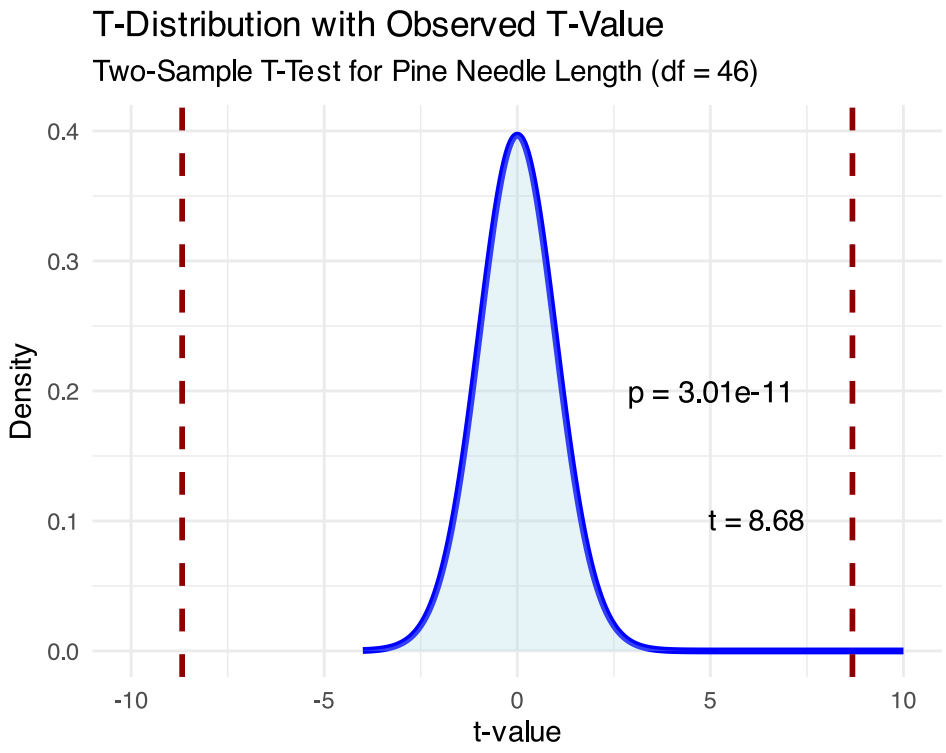
Lecture 5: Interpreting Two-Sample T-Test Results

Interpret the results of the two-sample t-test

What can we conclude about the needle lengths on windward vs. leeward sides?

How to report this result in a scientific paper:

“A two-tailed, two-sample t-test at α=0.05 showed [a significant/no significant] difference in needle length between windward (M = ..., SD = ...) and leeward (M = ..., SD = ...) sides of pine trees, t(...) = ..., p = ...”



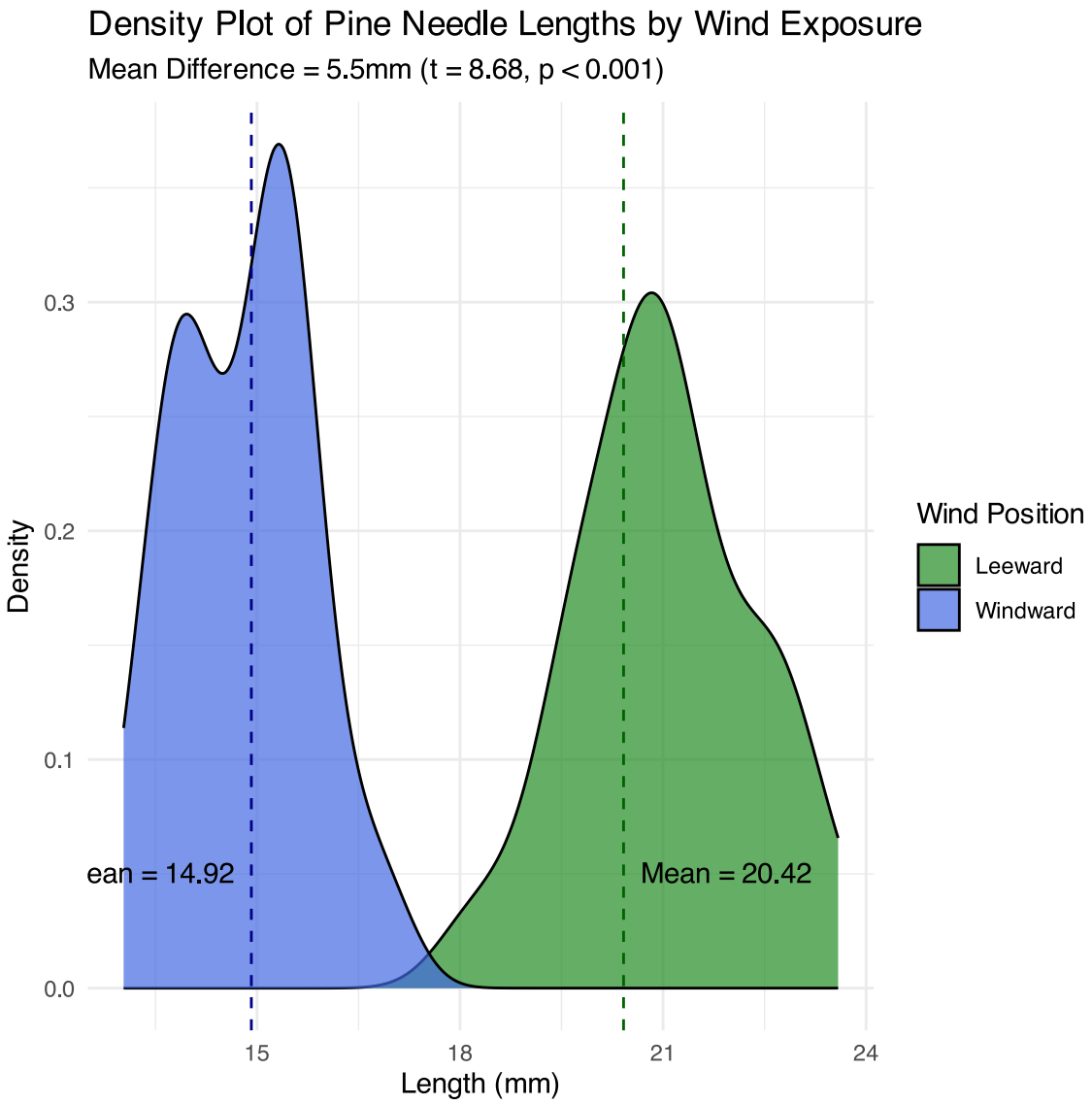
Lecture 5: Visualizing the Results

Interpret the results of the two-sample t-test

What can we conclude about the needle lengths on windward vs. leeward sides?

How to report this result in a scientific paper:

"A two-tailed, two-sample t-test at α=0.05 showed [a significant/no significant] difference in needle length between windward (M = ..., SD = ...) and leeward (M = ..., SD = ...) sides of pine trees, t(...) = ..., p = ..."



Lecture 5: Assumptions of Parametric Tests

Common assumptions for t-tests:

1. Normality: Data comes from normally distributed populations
2. Equal variances (for two-sample tests)
3. Independence: Observations are independent
4. No outliers: Extreme values can influence results

What can we do if our data violates these assumptions?

Alternatives when assumptions are violated:

- Data transformation (log, square root, etc.)
- Non-parametric tests
- Robust statistical methods

Lecture 5: Summary and Conclusions

In this activity, we've:

1. Formulated hypotheses about pine needle length
2. Tested assumptions for parametric tests

3. Conducted one-sample and two-sample t-tests
4. Visualized data using appropriate methods
5. Learned how to interpret and report t-test results

Key takeaways:

- Always check assumptions before conducting tests
- Visualize your data to understand patterns
- Report results comprehensively
- Consider alternatives when assumptions are violated