The Central Limit Theorem (CLT)

- The sampling distribution is centered around the true population mean.
- As the sample size increases, most sample means cluster closely around the true mean.
- The sampling distribution's shape may differ from the population's distribution but becomes more symmetrical and bell-shaped with larger samples.
- Both the figures and plots clearly demonstrate the Central Limit Theorem.

Regardless of the parent population distribution whose mean is μ and standard deviation is σ, the Central Limit Theorem (CLT) states that the sampling distribution of the sample mean $\bar{X}$ converges to a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size n increases, when sampling with replacement.

0.1 Confidence Interval for Continuous Cases

Note we need to define **SE**:Confidence Interval = Point Estimate $\pm$ $\underbrace{Z\_1 - \alpha/2 \times \text{SE}\_\text{MoE}}$

From the CLT definition, the **standard error** would be:SE $= \frac{\hat{\sigma}}{\sqrt{n}}$

CL - α - $\alpha/2$ - $z\_\alpha/2$-Critical value

90% - 0.10 - 0.050 - 1.645
95% - 0.05 - 0.025 - 1.960
98% - 0.02 - 0.010 - 2.326
99% - 0.01 - 0.005 - 2.576
99.5% - 0.005 - 0.0025 - 2.807
99.9% - 0.001 - 0.0005 - 3.291

```
#Calculate proportions
click_through %>%
  group_by(webpage) %>%
  summarize(
    prop = sum(click_target) / n(),
    successes = prop * n(),
    failures = n() * (1 - prop)
  )
#Calculate Standard Error
click_through_est <- click_through %>%
  group_by(webpage) %>%
  summarize(
    click_rate = sum(click_target) / n(),
    n = n()
  ) %>%
  mutate(se = sqrt(click_rate * (1 - click_rate) / n))
#Calculate Confidence Intervals
click_through_CLT_95_ci <- click_through_est %>%
  mutate(
    lower_95 = click_rate - (qnorm(0.975) * se), #for 90CI
        qnorm(0.95)
    upper_95 = click_rate + (qnorm(0.975) * se), #for 90CI
        qnorm(0.95)
    Method = rep("CLT", 2)
  )
```

0.2 Dif between CI from CLT and bootstrapped

CLT we are finding the CI directly from the sample. Bootstrapped (empirical) and CLT (theoritical) give almost same CI results for a sufficiently large sample.

0.3 When to use CLT or bootstrapped

If computational speed does not matter, and you do not violate any of the conditions of the CLT, then you can either use a simulation-based approach (like bootstrapping) or CLT, and you will get essentially the same results. If you are violating the conditions of the CLT, then you might want to use simulation-based approaches.

0.4 Hypothesis tests based on normal and t-distributions

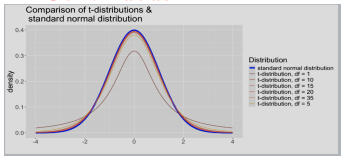
- a Normal distribution when we are testing for a difference in proportions;
- t-distribution when we are testing for a difference of continuous means.

0.5 Proportions

```
#Shortcut for tests in proportions case
click_summary <- click_through %>%
  group_by(webpage) %>%
  summarize(
    success = sum(click_target),
    n = n()
  )
prop.test(x = click_summary$success, n = click_summary$n,
  correct = FALSE, alternative = "less"
)
```

The square of a z-score is almost equal to Chi-square test. If p-value $< \alpha$, then we have enough statistical evidence to reject the null hypothesis in favour of the alternative hypothesis.

1 Continuous



1.1 One-sample t-test

1.2 two-sample t-test

```
#Shortcut for tests in continuous case
twospecies_data <- penguins |>
  filter(species %in% c("Adelie", "Gentoo")) |>
  drop_na(flipper_length_mm)
# Perform the t-test
t_test_result <- t.test(flipper_length_mm ~ species, data =
      twospecies_data, alternative = "less")
```

2 Deciding Which Test to Use

The choice of statistical test depends primarily on the type of variables being compared and the research question:

2.1 Chi-Square Tests (χ^2) : Categorical Variables

If your analysis involves comparing observed counts to expected counts for one or more categorical variables, you should use **Chi-Square Tests**.

The main assumptions of Chi-squared tests are that:
The observations are independent.
The expected counts are sufficiently large (greater than 5 is the typical standard).

The notion of chi-squared tests is to compare observed counts with expected counts from a population or distribution. In general, the hypotheses can be formulated as:
H_O: The observed (O) and expected (E) counts are equal.
H_A: The observed (O) and expected (E) counts are not equal.

One Categorical Variable:
- Use the *Chi-Square Goodness of Fit Test* to see if the distribution of that single variable matches a hypothesized distribution (e.g., equal distribution).

```
observed <- mm$Count
expected <- rep(sum(mm$Count) / 6, 6)
X_sq <- sum((observed - expected)^2 / expected) # Chi-
    Square
pchisq(X_sq, df=8, lower.tail=FALSE) # p-value
chi_test_result <- chisq.test(x = mm$Count, correct = FALSE
    ) #To get all the chi test values
chi_test_result <- chi_test_result$statistic
```

Two Categorical Variables:
- Use the *Chi-Square Test of Independence/Homogeneity* to determine if the distribution of one variable depends on the other (e.g., is click-through rate independent of the webpage version?).

Note: When comparing two categories, this test is mathematically equivalent to a two-sample proportion test.

```
count_table_AB <- click_through |>
  tabyl(webpage, click_target)
chisq.test(cont_table_AB, correct = FALSE)
```

2.2 One-Way Analysis of Variance (ANOVA): Continuous Variables

If you want to compare the means of a continuous outcome across

(ANOVA).
ANOVA Hypotheses:
- Use this if the assumptions of *Normality (of residuals)* and *Equal Variance* across groups are met.
H_O: All means are equal ($\mu_1 = \mu_2 = ... = \mu_k$)
H_A: Not all means are equal.

```
res <- aov(flipper_length_mm ~ species, data = penguins)
summary(res)
p_value <- summary(res)[[1]]$'Pr(>F)'[1]
```

The F-test:
- The test statistic is an F-ratio that compares the variance between groups (Treatment Sum of Squares, SST) to the variance within groups

2.3 ANOVA Assumptions

Normality: Residuals should be approximately normally distributed (checked via Q-Q plots or Shapiro-Wilk test).
Equal Variance (Homoscedasticity): Population variances should be equal across groups (checked via Levene's test).
Independent Observations.

```
qqnorm(res3$residuals)
shapiro.test(res$residuals)
leveneTest(flipper_length_mm ~ species, data = penguins)
```

2.4 Nonparametric Alternative: Kruskal-Wallis Test

```
kruskal.test(bill_length_mm ~ species, data = penguins)
```

When the normality assumption is violated, the Kruskal-Wallis Test serves as a robust alternative, comparing group medians based on the ranks of the data rather than the raw values.
When comparing two groups in ANOVA, the p-value matches the two-sample t-test, and the F-statistic equals the square of the t-statistic ($F \approx t^2$).
Since the Kruskal-Wallis test avoids assumptions about data distribution and variability, we make interpretations on the median (not mean), which better reflects the central tendency in the case of outliers or skewness.

Type I Error (α)

- We commit a **Type I error** if we **reject** H_0 when, in fact, it is **true** (a **false positive**).
- The Type I error rate is represented by the symbol α, known as the **significance level**.
- It's the primary error rate we control (e.g., using the rule p-value $< \alpha$).
- **Conditional Probability:** $\alpha = P(\text{reject } H_0 | H_0 \text{ true})$

Type II Error (β)

- We commit a **Type II error** if we **fail to reject** H_0 when, in fact, it is **not true** (a **false negative**).
- The Type II error rate is denoted by the symbol β.
- Its complement, $(1-\beta)$, is the **power of the test** (our ability to correctly reject a false null hypothesis).
- **Conditional Probability:** $\beta = P(\text{fail to reject } H_0 | H_0 \text{ false})$

The Power of a Test

The power of a test $(1-\beta)$ is the true positive rate. In other words, it is the probability of correctly rejecting H_O when it should be rejected because it is false. pnorm(qnorm(0.025), mean=-3, sd=1)

Cohen's Effect Size

There are different ways to compute effect size, and a popular one is called Cohen's effect size d

The p-value

Intuitively, a small p-value indicates that your observed test statistic is highly unlikely under H_O

The p-hacking

The misuse of p-values in scientific research include cherry-picking data in order to get a "significant result" in a given statistical study.

Double Dipping

if we look at all of the data and then use that same data to test our hypotheses it can lead to an issue in inference called double dipping. The consequence of this is our p-values are no longer controlling for the type I error rate, and results could be misleading.