Important Concepts

R-Code

Distribution of p-values under the null hypothesis for an exact test:

→ P-values should follow a U(0,1) under the null hypothesis

→ In a coin toss experiment, when testing a fair coin toss is being used, the p-values from multiple trials are U(0,1).

→ If a study with a small sample size results in p-values ≠ U(0,1), then there's an issue with the test's assumptions

```
nsim <- 500
P_val <- runif(nsim)
hist(P_val)
```

so be like this

Assumptions for performing a t-test

→ The t-test assumes the data's distribution is approximately normal, continuous, randomly sampled, homogeneity of variance (each group's variability is similar), and appropriate sample size. The lower the sample size {n}, the lower the statistical power.

→ When comparing the test scores of two groups of students, we determined that that each group is normal and variances are similar. = good results

→ You perform a t-test without checking for normality and variance equality = bad results

```
qqnorm(rnorm(500))
groupOne <- rnorm(100, mean=0, sd=1)
groupTwo <- rnorm(100, mean = 1, sd=1)
var.test(groupOne, groupTwo)
```

Assumptions for performing normal approximation to the binomial test

→ The sample size should be sufficiently large, and the probability of success in each trial should not be too close to 0 or 1. If $X \sim B(n,p)$, then $np > 5$ and $nq > 5$, it is better when they are greater than 10.

→ Use binomial test with 500 coin flips ($p = 0.5$) is good.

→ Use binomial test with 20 coin flips ($p = 0.8$) leads to unreliable results.

```
n <- 500
p <- 0.5
if (n*p > 10 && n*(1-p) > 10) {
  print("Sample size is large enough for norm")
}
if( p > 0.01 && p < 0.99) {
  print("Prob. within acceptable range")
}
```

Assumptions for interpreting the signed-rank test as a test of population mean/median

→ Assumes that the data is symmetrically distributed and the differences between pairs are independent and identically distributed.

→ When comparing differences in test scores before and after a teaching intervention, we assume the data has met the symmetrically distributed and independence.

→ Data is highly skewed and differences are not independent violates the assumptions.

```
dif <- rnorm(500)
hist(dif)
#Independence needs to be checked
 by the data collection
```

Relationship between the p-values and hypothesis test decisions

→ A smaller p-value indicates stronger evidence against the null, leading to rejecting the null at chosen confidence interval [$\alpha = 0.05$]

No code! Concept!

Relationship between confidence intervals and hypothesis tests

→ Confidence interval provides a range of values for a parameter estimate and it's construction is related to the hypothesis test. If a null hypothesis does not fall within the confidence interval, then we reject that null

No code! Concept!

Relationship between 95% and 99% confidence intervals

→ A 99% confidence interval is wider than a 95% confidence interval, as it captures more of the distribution, but results in a larger margin of error.

→ A 95% confidence interval for average height of a population might be 165-175cm while a 99% confidence interval could be 160-180cm.

No code! Concept!

Using incorrect variance in a z-test

→ A variance that is too large results in a narrower confidence interval $\overset{\text{MAY}}{\underset{\text{RESULT}}{}}$ → increasing Type I error

→ A variance that is too small results in a wider confidence interval IN → increasing Type II error

```
trueMean <- 500
sampleMean <- 510
incorrectVar <- 35
zScore <- sampleMean - trueMean/sqrt(incorrectVar/n)
pval <- 2*(1-pnorm(abs(zScore)))
```

Performing the chi-squared test for variance on non-normal data:

→ Big problem: Chi-squared assumes the data is normal. Hence, using chi-squared in non-normal dat will result in inconclusive and/or inaccurate results.

No code! Concept!

Performing the Kolmogorov-Smirnov test using estimated parameter values

→ Kolmogorov-Smirnov test compares an empirical distribution to a theoretical distribution with specified parameters. By using estimated parameters can lead to an inaccurate test.

```
data <- rnorm(500, mean=0, sd=1)
ks.test( data "pnorm", mean=mean(data), sd = sd(data))
```

CLT: sample distribution of the sample mean (or other stats) approaches $N(\mu, \sigma)$. The larger samples → $N(\mu, \sigma)$

Type 1: null rejected, when actually true

Type 2: null not rejected, when actually false

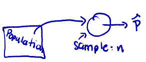
Significance level: prob of making Type I error (α)

Power: $(1-\beta)$ prob of correctly rejecting false null

Proportions

$H_0: p = p_1$

$H_a: p \neq p_1$



$\hat{p}$ ← sample proportion/statistic

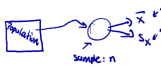
$Z = \dfrac{\hat{p} - p_0}{\sigma_{\hat{p}}}$ $\quad \sigma_{\hat{p}} = \sqrt{\dfrac{p_0(1-p_0)}{n}}$

Z-statistics

Means

$H_0: \mu = \mu_1$

$H_a: \mu \neq$



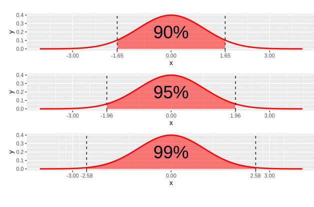
$\bar{x}$ ← sample mean

S_x ← sample standard deviation

If pop $\{\sigma_{\bar{x}}\}$ is known

$Z = \dfrac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$ ← $\dfrac{\sigma}{\sqrt{n}}$

t-statistic → $t = \dfrac{\bar{x} - \mu_0}{\frac{S_x}{\sqrt{n}}}$ ← estimate of $\sigma_{\bar{x}}$



Z-test: Is often used to determine whether the mean of a sample is significantly different from a known distribution mean when the population standard deviation is known [$\bar{x}$ = sample mean] [μ = population mean] [σ = population standard deviation] [n = sample size]

Assumptions: σ is known, sample should be randomly selected from the population, and the data should follow a normal distribution. $X \sim N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

Confidence interval: $\bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$, where z = z-score (90% = 1.645, 95% = 1.96, 98% = 2.33, 99% = 2.575)

T-Test: Is used to determine if the means of two data are significantly different from each other. There are two types: one sample and two sample t-tests. where [$\bar{x}$ = sample mean] [μ = population mean] [s = sample standard deviation] [n = sample size]

Assumptions: Data should be approximately $N(\mu, \sigma)$ and the sample should be randomly selected from the population and homogenity is present (and independent and identically distributed)

Confidence interval: Depends on t-distribution table where degrees of freedom and significance level matter: Degrees of freedom ($df = n-1$), confidence level (95%, 96%, 98%, 99%)

One sample t-test: $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$, **Two-sample t-test**: $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$

$\bar{x}_1, \bar{x}_2$ = sample mean of each group
s_p^2 = the pooled variance, common variance of each group
n_1, n_2 = sample sizes for each group

Exact Binomial Test: We use this test to determine if the proportion of successes in a fixed number of trials differ from a hypothesized value.

$P = \binom{n}{x} p^x (1-p)^{n-x}$ [n = number of trials] [x = number of successes] [p = hypothesized probability of success in a single trial]

Assumptions: Each trial is independent, there are a fixed number of trials, and each trial has the same probability of success.

Normal Approximation Binomial Test (Z-test): This test approximates the binomial distribution when the sample size is sufficiently large.

See first page!

$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

z = z-score $\hat{p}$ = sample proportion of success
p_0 = hypothesized population proportion n = sample size.

Assumptions: The sample size {n} is sufficiently large. Follow the rule that both np_0 and $n(1-p_0)$ should be greater than 5 to ensure that the normal approximation is valid. The samples should be randomly selected or obtained through a process that can be treated as random. Lastly, the outcomes should be independent.

Sign Test: Used to determine whether the median of a set of differences between paired observations is significantly different from hypothesis value (e.g. zero). Useful, when data does not follow a normal distribution.

Assumptions: paired differences are independent, can be ordinal, null usually assumes no difference.

Signed-Rank Test (Wilcoxon Signed-Rank Test): used to determine whether the median of a set of paired observations is significantly different from a hypothesized value. It is an extension of the sign test that considers the magnitude of differences not their signs.

Assumptions: same as sign test.

Chi-Squared Test for variance: used to determine whether variance of a population differs from a hypothesised value.

Assumptions: follow a normal distribution and independence.

Alternative T-test for variance: compares variances of two independent samples

Assumptions: same as chi-squared

Kolmogorov-Smirnov Test for Distributions: Used to compare the cumulative distribution function of a sample with a reference distribution.