276) [True or False] F statistic cannot be negative.

A) TRUE

B) FALSE

Solution: (A)

F statistic is the value we receive when we run an ANOVA test on different groups to understand the differences between them. The F statistic is given by the ratio of between group variability to within group variability

Below is the formula for f Statistic.

$$\frac{\text{Sum of squared error for between group/degree of freedom of between group}}{\text{Sum of squared error for within group/degree of freedom of within group}}$$

Since both the numerator and denominator possess square terms, F statistic cannot be negative.

How it works...

The coin tossing experiment is modeled as a sequence of n independent random variables $x_i \in \{0, 1\}$ following the Bernoulli distribution $B(q)$. Each x_i represents one coin flip. After our experiment, we get actual values (samples) for these variables. A different notation is sometimes used to distinguish between the random variables (probabilistic objects) and the actual values (samples).

The following formula gives the sample mean (proportion of heads here):

$$\overline{x} = \frac{1}{n} \sum_i x_i$$

Knowing the expectancy $\mu = q$ and variance $\sigma^2 = q(1 - q)$ of the distribution $B(q)$, we compute:

$$E[\overline{x}] = \mu = q$$
$$\mathrm{var}(\overline{x}) = \frac{\sigma^2}{n} = \frac{q(1-q)}{n}$$

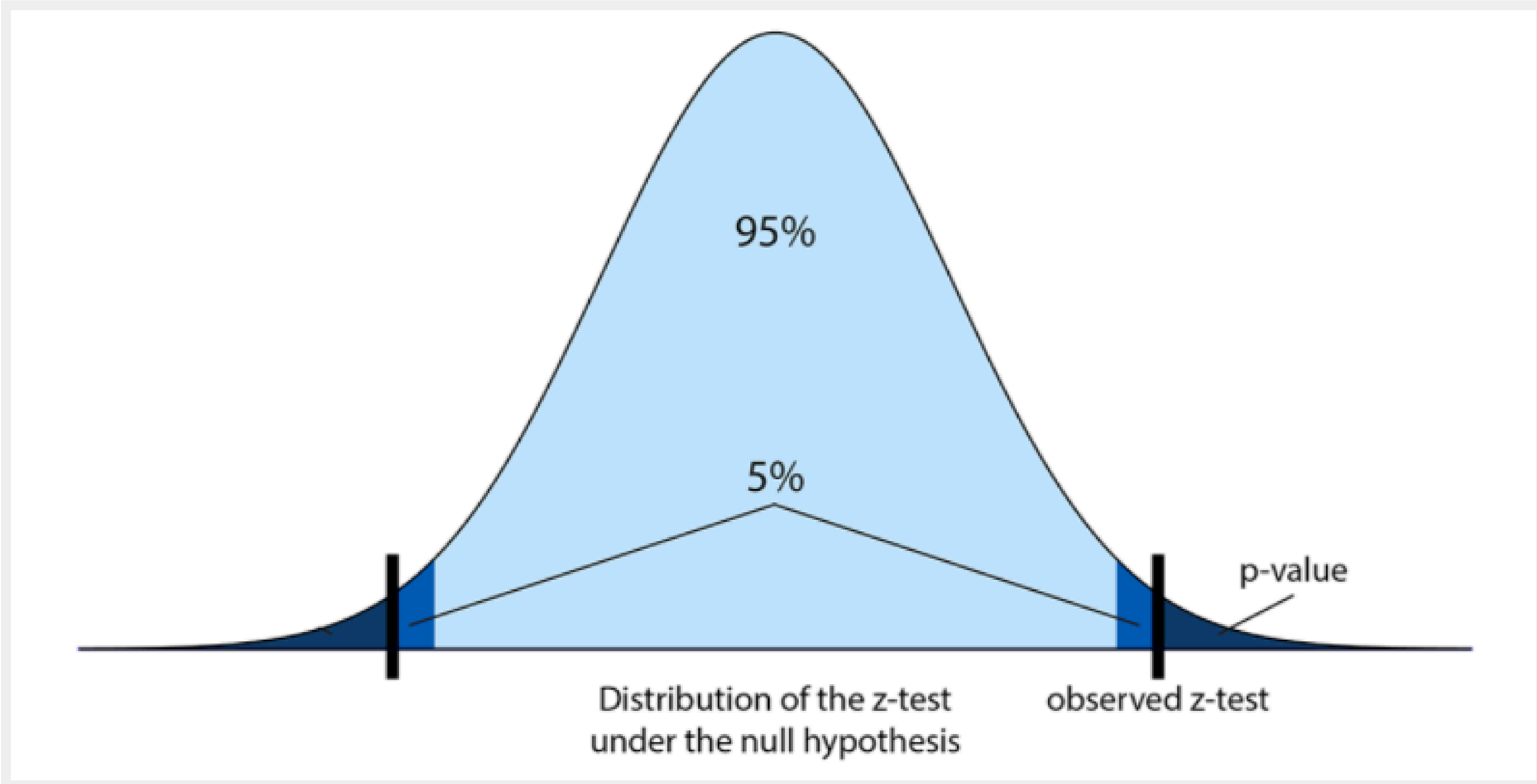
The z-test is the normalized version of $\overline{x}$ (we remove its mean, and divide by the standard deviation, thus we get a variable with mean 0 and standard deviation 1):

$$z = \frac{\overline{x} - E[\overline{x}]}{\mathrm{std}(\overline{x})} = (\overline{x} - q)\sqrt{\frac{n}{q(1-q)}}$$

Under the null hypothesis, what is the probability of obtaining a z-test higher (in absolute value) than some quantity z_0? This probability is called the (two-sided) p-value. According to the central limit theorem, the z-test approximately follows a standard Gaussian distribution $N(0, 1)$ for large n, so we get:

$$p = P[|z| > z_0] = 2P[z > z_0] \simeq 2(1 - \Phi(z_0))$$

The following diagram illustrates the z-score and the p-value:

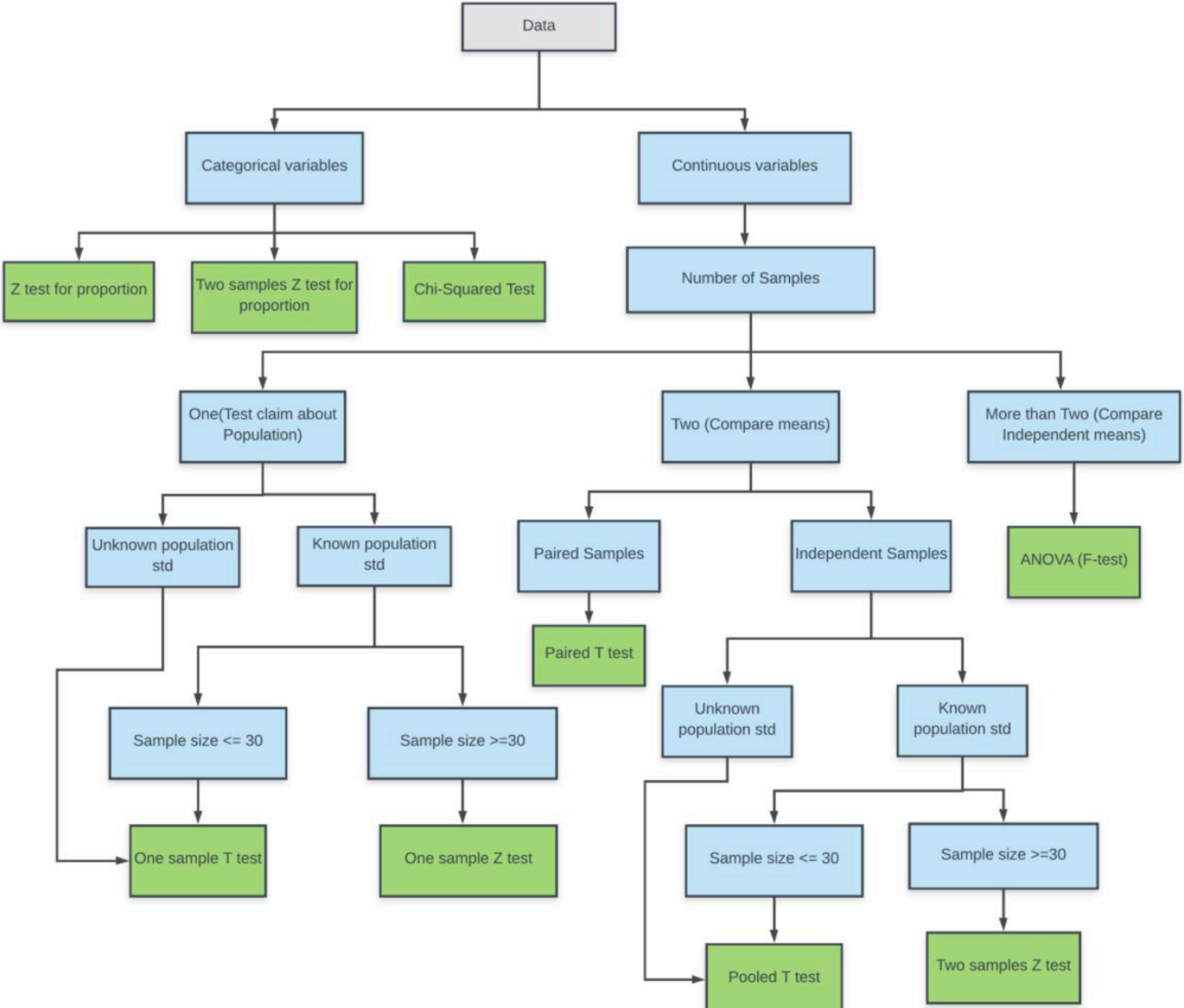


In this formula, Φ is the cumulative distribution function of a standard normal distribution. In SciPy, we can get it with `scipy.stats.norm.cdf`. So, given the z-test computed from the data, we compute the p-value: the probability of observing a z-test more extreme than the observed test, under the null hypothesis.

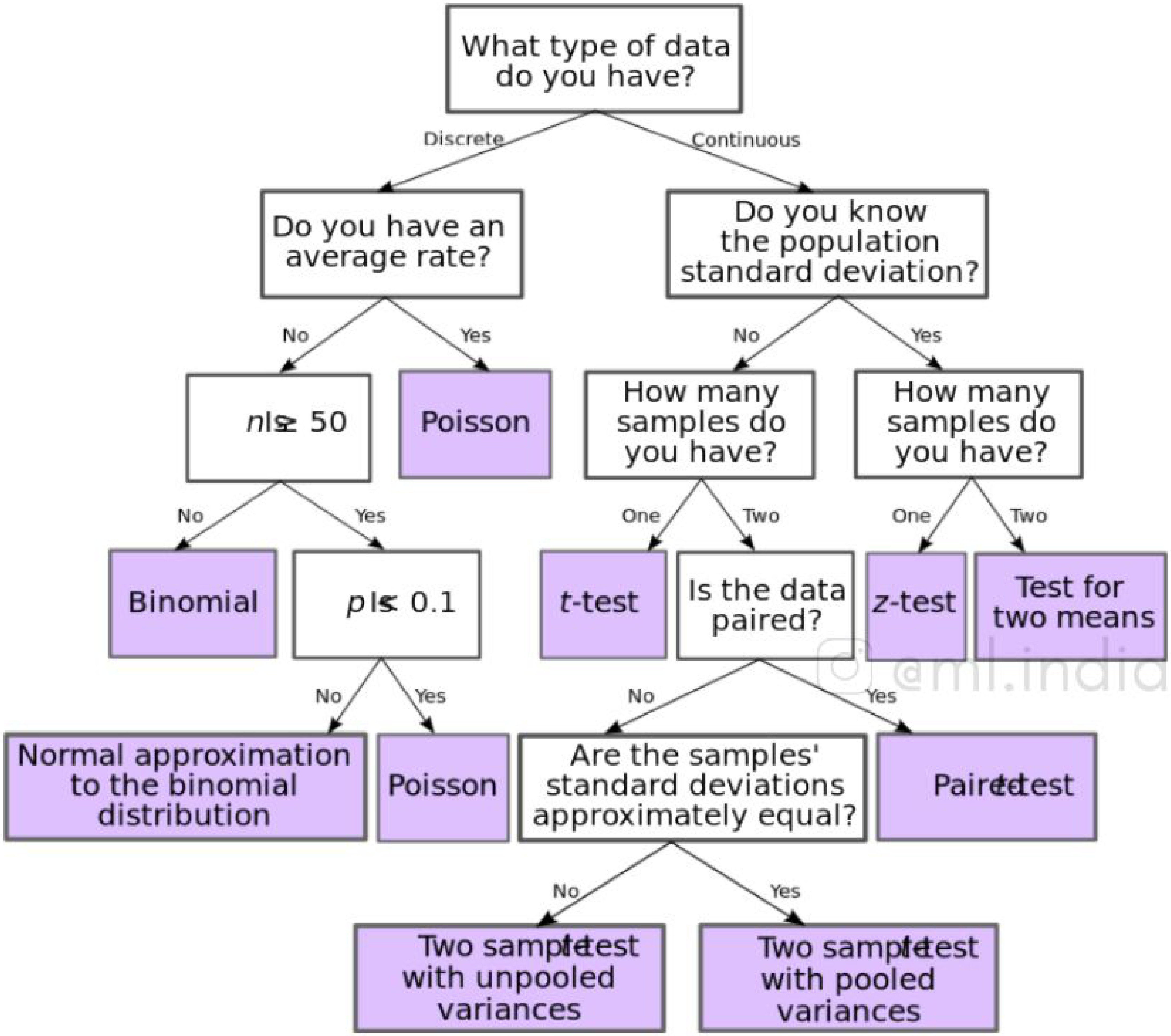
If the p-value is less than five percent (a frequently-chosen significance level, for arbitrary and historical reasons), we conclude that either:

- The null hypothesis is false, thus we conclude that the coin is unfair.
- The null hypothesis is true, and it's just bad luck if we obtained these values. We cannot make a conclusion.

We cannot disambiguate between these two options in this framework, but typically the first option is chosen. We hit the limits of frequentist statistics, although there are ways to mitigate this problem (for example, by conducting several independent studies and looking at all of their conclusions).


Data
Categorical variables
Continuous variables
Z test for proportion
Two samples Z test for proportion
Chi-Squared Test
Number of Samples
One(Test claim about Population)
Two (Compare means)
More than Two (Compare Independent means)
Unknown population std
Known population std
Paired Samples
Independent Samples
ANOVA (F-test)
Paired T test
Sample size <= 30
Sample size >=30
Unknown population std
Known population std
One sample T test
One sample Z test
Sample size <= 30
Sample size >=30
Pooled T test
Two samples Z test

A **cheatsheet** on selecting a **hypothesis** test:

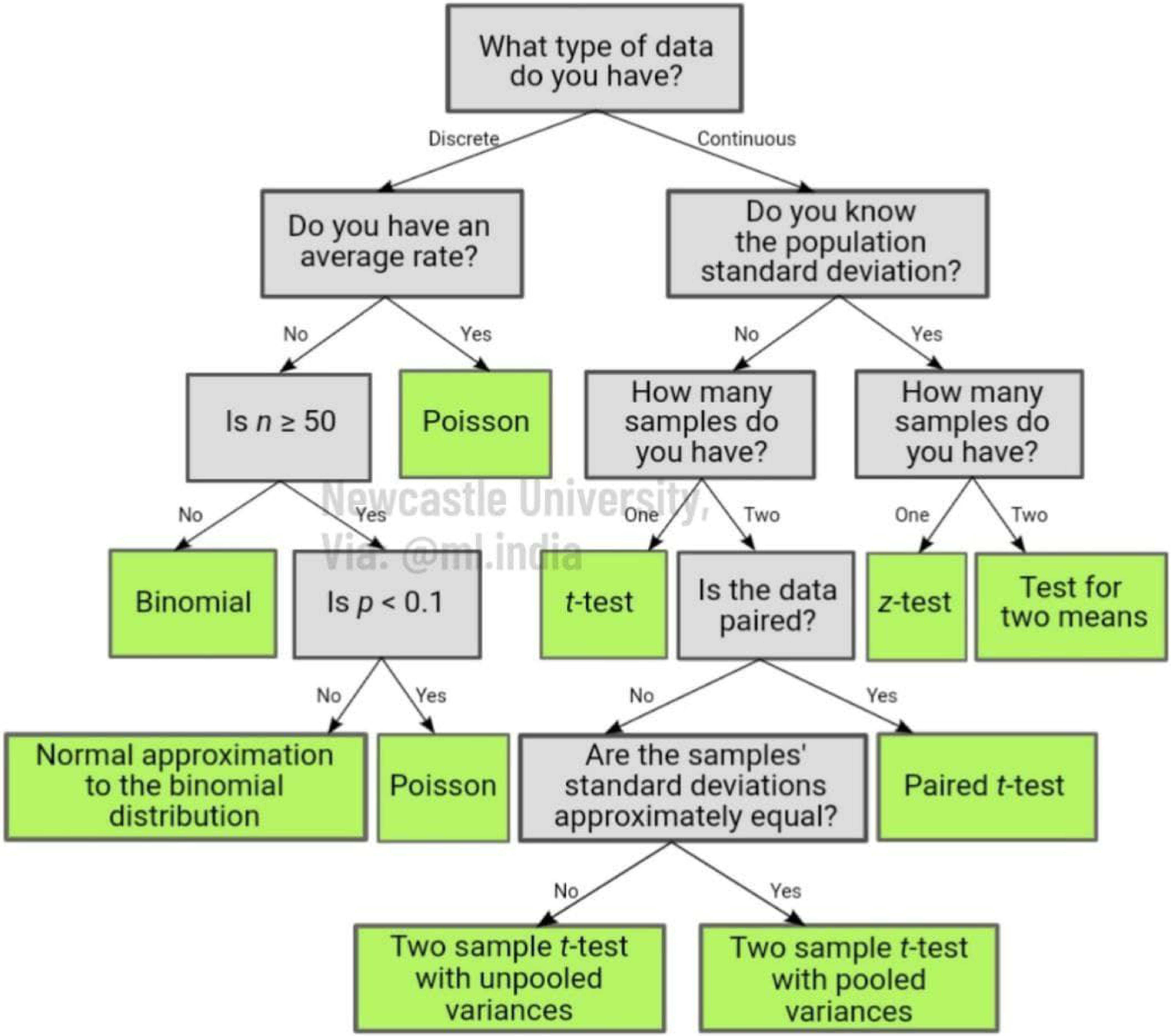


Source: Newcastle University

Hit to support!

Save for later!

Hypothesis Testing:

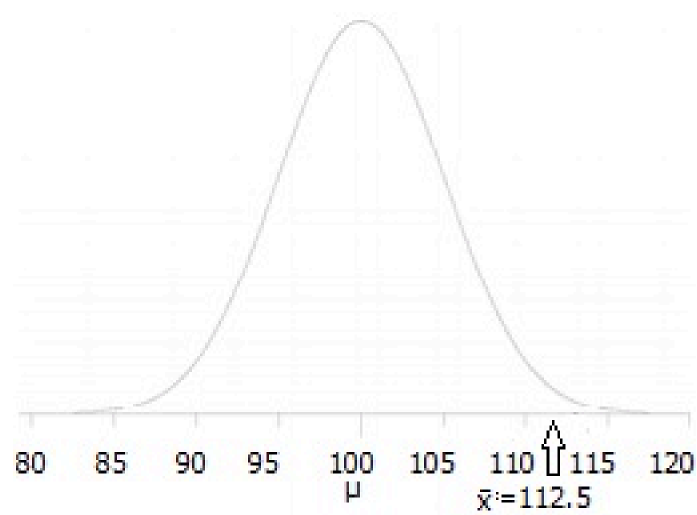


A principal at a certain school claims that the students in his school are above average intelligence. A random sample of thirty students IQ scores have a mean score of 112.5. Is there sufficient evidence to support the principal's claim? The mean population IQ is 100 with a standard deviation of 15.

Step 1: State the Null hypothesis. The accepted fact is that the population mean is 100, so: H_0: $\mu=100$.

Step 2: State the Alternate Hypothesis. The claim is that the students have above average IQ scores, so:
H_1: $\mu > 100$.
The fact that we are looking for scores "greater than" a certain point means that this is a one-tailed test.

Step 3: Draw a picture to help you visualize the problem.



Step 4: State the alpha level. If you aren't given an alpha level, use 5% (0.05).

Step 5: Find the rejection region area (given by your alpha level above) from the z-table. An area of .05 is equal to a z-score of 1.645.

Step 6: Find the test statistic using this formula:

$$Z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

For this set of data: $z = (112.5 - 100) / (15/\sqrt{30}) = 4.56$.

Step 6: If Step 6 is greater than Step 5, reject the null hypothesis. If it's less than Step 5, you cannot reject the null hypothesis. In this case, it is greater ($4.56 > 1.645$), so you can reject the null.

Overview of statistical tests

Type of dependent variable	Type of independent variable						
	Ordinal/categorical				Normal/interval (ordinal)	More than 1	None
	Two groups		More groups				
	Paired	Unpaired	Paired	Unpaired			
2 categories	McNemar Test, Sign-Test	Fisher Test, Chi-squared-Test	Cochran's Q-Test	Fisher Test, Chi-squared-test	(Conditional) Logistic Regression	Logistic Regression	Chi-squared-Test
Nominal	Bowker Test	Fisher Test, Chi-squared-Test		Fisher Test, Chi-squared-test	Multinomial logistic regression	Multinomial logistic regression	Binomial Test
Ordinal	Wilcoxon Test, Sign-Test	Wilcoxon-Mann-Whitney Test	Friedman-Test	Kruskal-Wallis Test	Spearman-rank-test	Ordered logit	Median Test
Interval	Wilcoxon Test, Sign-Test	Wilcoxon-Mann-Whitney Test	Friedman-Test	Kruskal-Wallis Test	Spearman-rank test	Multivariate linear model	Median Test
Normal	t-Test (for paired)	t-Test (for unpaired)	Linear Model (ANOVA)	Linear Model (ANOVA)	Pearson-Correlation-test	Multivariate Linear Model	t-Test
Censored Interval	Log-Rank Test		Survival Analysis, Cox proportional hazards regression				
None	Clustering, factor analysis, PCA, canonical correlation						

268) A researcher concludes from his analysis that a placebo cures AIDS. What type of error is he making?

A) Type 1 error

B) Type 2 error

C) None of these. The researcher is not making an error.

D) Cannot be determined

Solution: (D)

By definition, type 1 error is rejecting the null hypothesis when its actually true and type 2 error is accepting the null hypothesis when its actually false. In this case to define the error, we need to first define the null and alternate hypothesis.

269) What happens to the confidence interval when we introduce some outliers to the data?

A) Confidence interval is robust to outliers

B) Confidence interval will increase with the introduction of outliers.

C) Confidence interval will decrease with the introduction of outliers.

D) We cannot determine the confidence interval in this case.

Solution: (B)

We know that confidence interval depends on the standard deviation of the data. If we introduce outliers into the data, the standard deviation increases, and hence the confidence interval also increases.