|  |  |  |
| --- | --- | --- |
|  | **Population statistics** | **Sample statistics** |
| Total Number | N | n |
| Mean | μ (given) | x̅ = |
| Median | - | x̃ (middle) |
| Standard deviation | or given |  |
| Pearson Coefficient of linear correlation | ρ: |  |
| Proportion | p (given) | p̂ = |
| Calculated test statistic | - | z, t, , Binomial |

|  |  |  |
| --- | --- | --- |
| **Rules** | **Formula** | **Remark** |
| 68-95-99.7 rule | ,2,3 | Can be used to find extreme data |
| Interquartile Range | IQR = Q3 - Q1 | Can be used to test spread |
| Central location | Median | Compare between two datasets |
| 5 Numbers Box Plot | Min|Q1|Median|Q3|Max | Spread & Outliers |
| Outliers | >Q3 + 1.5IQR  <Q1 - 1.5IQR | If outliers' are present, distribution is less likely to be normal |
| Influential Point | Recalculation of r value | Is an outlier and removing it affect the best fit line |
| Simpson's Paradox | An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. Due to the difference in weightage | |
| Restricted Range | Association changes when data range changes | |
| Average data | Correlation based on average data most likely resulting in higher association | |
| Central Limit Theorem | For sample with large sample size, the sampling distribution of the sample mean is approximately a normal distribution | Note it is the sample mean, not the sample |
| Law of Large Numbers | With a large number of experiments, the average will tends towards the expected value | |
| Binomial probability |  |  |
| Normality Test | Anderson Darling Normality  Normality probability plot | P≥0.05 when normal  Linear plot |
| Examine graph | Overall pattern (Form/ Direction/ Strength)  Deviations (Outliers) | |
| Confidence Interval | α% of the confidence intervals constructed in this way would contain the true value for the population parameter of .  We estimate with α% confidence that between CI of | |
| Conclusion statement | There is sufficient/insufficient evidence to support | |

|  |  |  |
| --- | --- | --- |
| **Studies** | **Method** | **Remarks** |
| Observational Studies | Observe samples without modification | Contrast to experimental studies  Placebo/Control Group/Blinding |
| Simple Random Sampling | Every sample has equal probability | Essential for N, Z, T tests |
| Stratified Random Sampling | Divide sample into representative groups before SRS | Useful if the sample contains different groups |
| Cluster Random Sampling | Divide sample into groups then randomly select a few groups | Low cost and no need sampling frame  Large sample size required to reduce margin of error |
| Voluntary response Sampling | Require sample to voluntarily response | Subject to response/ non-response bias |
| Multistage Sampling | Conduct the sampling in different stages | Easily mistaken with stratified sampling |

|  |  |  |  |
| --- | --- | --- | --- |
| **CI (α/2) sample mean** | 90% | 95% | 99% |
| CI = ± Z\* × | 1.645 | 1.960 | 2.576 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Biasness** | **Problem** | | |
| Response Bias | Misleading questions, incorrect response | | |
| Non-Response Bias | Subjects cannot be reached or refused to participate | | |
| Sampling Bias | Under coverage/ Non-random sample | | |
| Anecdotal data | Anecdotal evidence is based on haphazardly selected individual cases | | |
| **Bias** | Affects Mean | **Variability** | Affects Spread |

|  |  |  |
| --- | --- | --- |
| **Linear Transformation** | **Center, x̅** | **Spread, or** |
| Addition (a) | a x̅ | no change |
| Multiplication (b) | b x̅ |  |

|  |  |  |
| --- | --- | --- |
|  | **Assumption** | **Implication** |
| Z, T test | Normality | All formulas require the sample to be normally distributed |
| Z, T test | Simple Random Sample (Independence) | Result can only apply to selected samples and not the population |

|  |  |  |
| --- | --- | --- |
| **Interpret Association** | **Quantitative** | **Categorical variable** |
|  | Scatter diagram | Two-Way Table |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| **Outliers** | Left  Mean < Median | No  Mean = Median | Right  Mean > Median |

|  |  |
| --- | --- |
| **Type I Error** | **Type II Error** |
| P(Reject is true) = significant level (α)  α times there will be an error | P(Accept is true) = β  Power = 1 - β = P(Reject is true),  Higher power = better |
| A Type I error occurs when the researcher rejects a null hypothesis when it is true | A Type II error occurs when the researcher accepts a null hypothesis that is false |
|  | Both errors are calculated using **Z test**  Despite their initial distributions  The probability of *not* committing a Type II error is called the Power of the test. Power helps to determine if sample size is large enough. If your sample size is too small, your results may be inconclusive when they may have been conclusive with a large enough sample.  It's better to commit Type II error than Type I error |

|  |  |  |
| --- | --- | --- |
| **Regression** | **Implication** | **Remark** |
| R-Squared | Fraction of the variation in the values of y that is explained by the least-squares regressions of y on x | |
| Least-squares | Least square is when the minimum sum of residual^2 | |
| Correlation, r | Strength and direction of the linear association between two quantitative variables  Non linear r/s: Circular r/s, Curve r/s | Independent of response/explanatory variable  Not resistant to outliers (e.g. influential observation)  Independent of unit of variables |
| Residual | Sum of residual = 0, above 0 = overestimated, under 0 = underestimated  Residual = , vertical distance between actual and predicted response variable  Regression Line should contains both positive and negative residuals (uniform residual plot) | |
| Regression Equation | , regression line always pass through | |
| Lurking Variable | An unobserved variable that influences the association between variable of primary interest | Confounding effect |
| Common response | Similar to confounding, difference in x and y has no relationships | |

|  |  |
| --- | --- |
| **Probability rules** | |
| False Negative: P(NEG|P) | False Positive: P(POS|A) |
| Sensitivity: P (POS|P) | Specificity: P (NEG|A) |
| P(Outcome occurring as first n events) = P(outcome as last n events) | |
| **Conditional Probability** | |
| Addition rule for disjoint events  If events are disjoint, then events are dependent (if dice is odd, the chance of dice is even can never happen) | General Addition rule |
| Multiplication Rule for independent events | Complement |
| Conditional probabilities | Intersection |
| Independence |  |

|  |  |
| --- | --- |
| **Calculator function** | **Implication** |
| Normal Calculation |  |
| Binomial Calculation |  |

|  |  |  |
| --- | --- | --- |
| **Standard Normal** |  | 1. Unimodal 2. Bell shaped 3. Symmetric |
| **Normal Proximation** | Binomial distribution  Sample Proportion |  |
| **Continuity Correction (cc)** | Required for Binomial approximation to normal distribution |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Distribution** | **Usage** | **Mean** | **Standard deviation** | **Hypothesis Testing** | **Standard Error/**  **Sample Size** | **Confidence Interval**  𝑃𝑜𝑖𝑛𝑡 𝑒𝑠𝑡𝑖𝑚𝑎𝑡𝑒 ± 𝑚𝑎𝑟𝑔𝑖𝑛 𝑜𝑓 𝑒𝑟𝑟𝑜𝑟(m) | **Remarks** |
| Population *Only if given population parameters* | | | | | | | |
| Normal | Known population s.d. |  |  |  | *Known true parameters* | *Known true parameters* | Must be Normal |
| Binomial | Known pop. probability  Two outcomes |  |  | [extreme cases] | *Known true parameters* | *Known true parameters* | Approx. Normal w. condition |
| Sample *Must be Normal, or large sample through CLT* | | | | | | | |
| Sample proportion | Unknown pop. probability |  |  |  |  |  | Approx. Normal w. condition |
| Sample Mean | Normal distribution or CLT |  |  |  |  |  |  |
| Student's T | 1. Unknown population s.d.  2. Small sample size |  |  |  |  |  | df = (n-1)  Degree of freedom: |
| Two samples *Must be Normal, or large sample through CLT* | | | | | | | |
| Matched Paired | Before & After |  |  |  |  |  | df = (n-1) |
| 2 Samples Mean | Z test, Compare Differences |  |  |  |  |  |  |
| 2 Samples Mean | t test, Compare Differences |  |  |  |  |  | Must be between two independent samples  df:  min |
| Pooled Test |  |  |  |  |  |  | df = |
| Two sample proportions | Compare Differences |  |  |  |  |  |  |