Assumptions, Guidelines, and Rules-of-Thumb for Hypothesis Tests and Confidence Intervals

The following assumptions and rules-of-thumb (marked with *) should be examined and assessed before performing any inference procedures.

---

1. Investigating a mean μ with one sample:

   A. Use the normal distribution if

      I. you have a properly gathered random sample (simple, systematic, stratified, clustered, etc.)

      II. the population is normal

      III. sigma (σ) is known

      Please note that this situation (A) NEVER occurs in real problems. We are only using it pedagogically to help understand the logic and set up of hypothesis tests and confidence intervals.

   B. Use t distribution with $df = n - 1$ if

      I. you have a properly gathered random sample (simple, systematic, stratified, clustered, etc.)

      II. at least one of the following is true:

         a. sample size is small ($n < 15$) and there is no skewness and there are no outliers*

         b. sample size is medium ($15 \leq n \leq 40$) and there is little skewness and no extreme outliers*

         c. sample size is large ($n > 40$)*

2. Investigating a difference of means (μ_D) with two dependent samples:

   A. Use t-distribution with $df = n - 1$ if

      I. you have a properly gathered random sample (simple, systematic, stratified, clustered, etc.)

      II. the samples are 'matched'

      III. the list of differences satisfies at least one of the following:

         a. the list is small ($n < 15$) and there is no skewness and there are no outliers*

         b. the list is medium ($15 \leq n \leq 40$) and there is little skewness and no extreme outliers*

         c. the list is large ($n > 40$)*

3. Investigating a difference of means $(\mu_1 - \mu_2)$ with two independent samples:
    A. Use t-distribution with df = Satterthwaite's formula or $df = \min\{n_1 - 1, n_2 - 1\}$ if
        I. BOTH samples satisfy all conditions in part 1 B above

4. Investigating a proportion p with one sample:
    A. Use the normal distribution if
        I. you have a properly gathered random sample (simple, systematic, stratified, clustered, etc.)
        II. $np > 10$ and $nq > 10$*
        III. $N > 10n$*

5. Investigating a difference of proportions $(p_1 - p_2)$ with two samples:
    A. Use the normal distribution if
        I. BOTH samples are properly gathered random samples (simple, systematic, stratified, clustered, etc.)
        II. $np > 10$ and $nq > 10$ for BOTH samples (NOTE: For confidence intervals use $\hat{p}_1, \hat{q}_1, \hat{p}_2, \hat{q}_2$ to check conditions. For a hypothesis test you must use the pooled $\hat{p}$ to check conditions.)*
        III. $N > 10n$ for BOTH samples*

6. Investigating Goodness-of-Fit (1 sample, 1 categorical variable) OR Independence (1 sample, 2 categorical variables) OR Homogeneity (2 or more samples, 1 categorical variable):
    A. Use χ^2 distribution if
        I. you have a properly gathered random sample (simple, systematic, stratified, clustered, etc.)
        II. you are working with counts
        III. at least one of the following is true
            a. all expected counts are ≥ 5*
            b. all expected counts are > 1 AND no more than 20% of expected counts are < 5*

7. Investigating the slope (β) of a linear regression model:
    A. Use t distribution with $df = n - 2$ if
        I. residuals are normally distributed and homoscedastic
        II. residual plot does not show a curved or fan-shaped pattern*
        III. residual plot shows little skewness and no extreme outliers*

Here is a suggested format for doing hypothesis tests:

1. define your variables and clearly state your hypotheses in words or symbols
2. assess the credibility of the assumptions when possible and state that you are worried about the assumptions if you cannot assess them
3. state the formula and calculate your test statistic (z, t, χ^2, with df if applicable) - it is usually helpful to sketch a picture of the appropriate PDF
4. calculate the p-value using the appropriate table or make your decision based on the level of significance (α) provided in the problem
5. state your decision and conclusion in the context of the problem

Here is a suggested format for constructing confidence intervals:

1. assess the credibility of the assumptions when possible and state that you are worried about the assumptions if you cannot assess them
2. construct the interval using the following:

$$\begin{pmatrix}\text{best estimate of}\\ \text{your parameter}\end{pmatrix} \pm \begin{pmatrix}\text{critical value from}\\ \text{your distribution}\end{pmatrix} \cdot \begin{pmatrix}\text{standard error of}\\ \text{your estimate}\end{pmatrix}$$