Hypothesis Testing — Detailed Explanation

1. Overview of Hypothesis Testing
Hypothesis testing is a formal procedure for deciding whether data provide enough evidence to reject a stated claim about a population. The main elements are:

- **Null hypothesis (H_0):** the default claim (e.g., "$\mu = \mu_0$", "no effect").
- **Alternative hypothesis (H_1 or H_a):** what we consider if H_0 is rejected (e.g., "$\mu \neq \mu_0$", "$\mu > \mu_0$", "$\mu < \mu_0$").
- **Test statistic:** a function of sample data that has a known sampling distribution under H_0.
- **Significance level (α):** pre-chosen probability of making a Type I error (common choices: 0.05, 0.01).
- **P-value:** probability, under H_0, of observing a test statistic as extreme or more extreme than the one observed.
- **Decision rule:** either (a) compare p-value to α (reject H_0 if $p \le \alpha$), or (b) compare test statistic to critical value(s) (reject H_0 if statistic falls in rejection region).

---

2. Type I and Type II Errors, Power

- **Type I error (α):** rejecting H_0 when it is true. Probability = α (set by researcher).
- **Type II error (β):** failing to reject H_0 when H_1 is true. Probability depends on true effect, sample size, variance, α.
- **Power:** $1 - \beta$ = probability of correctly rejecting H_0 when H_1 is true. Increasing sample size, effect size, or α raises power.

Decision-outcome table:

- True H_0 & accept → Correct.
- True H_0 & reject → Type I error (α).
- False H_0 & accept → Type II error (β).
- False H_0 & reject → Correct (power).

Sample-size relation (two-sided z test, known σ):
$n \approx [(z_{1-\alpha/2} + z_{1-\beta}) * \sigma / \Delta]^2$
where Δ = minimum effect size you want to detect, z quantiles from standard normal.

---

3. Rejection Regions (Critical Regions)

- **Two-tailed test:** H_0: parameter = value; H_1: parameter ≠ value. Rejection if |test statistic| > critical value (e.g., $z > 1.96$ for $\alpha = 0.05$).
- **One-tailed test (right):** H_1: parameter > value. Reject if statistic > $z_{1-\alpha}$. Example: $\alpha=0.05 \Rightarrow$ critical $z \approx 1.645$.

- **One-tailed test (left):** H_1: parameter < value. Reject if statistic < z_{α}.
  Critical values depend on the sampling distribution (z, t, χ^2, F) and degrees of freedom.

---

4. Z-test
Purpose: Test means or proportions when sampling distribution is approximately normal and population variance is known (or n large so CLT applies).

One-sample mean (known σ):
Test statistic: $Z = (\bar{x} - \mu_0) / (\sigma / \sqrt{n})$
Decision: Compare Z to z critical or compute p-value from standard normal.

Two-sample mean (known σs):
$Z = (\bar{x}_1 - \bar{x}_2 - \Delta_0) / \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$

Proportion test (one sample):
$Z = (\hat{p} - p_0) / \sqrt{[p_0(1-p_0) / n]}$
(Use pooled proportion for two-sample proportion tests when H_0: $p_1 = p_2$.)

Assumptions: independent samples, normality (or large n), known population σ (or large n).

---

5. T-test (Student's t)
Used when population variance is unknown and/or sample size is small. The sampling distribution is Student's t with specific degrees of freedom (df).

One-sample t-test:
$t = (\bar{x} - \mu_0) / (s / \sqrt{n})$
$df = n - 1$

Paired t-test (dependent samples):
Compute differences d_i, $\bar{d}$ = mean difference, s_d = sd of differences.
$t = (\bar{d} - \mu_{d0}) / (s_d / \sqrt{n})$
$df = n - 1$

Independent two-sample t-tests:

- **Pooled t (equal variances assumed):**
  $sp^2 = [(n_1-1)s_1^2 + (n_2-1)s_2^2] / (n_1 + n_2 - 2)$
  $t = (\bar{x}_1 - \bar{x}_2 - \Delta_0) / [sp * \sqrt{1/n_1 + 1/n_2}]$
  $df = n_1 + n_2 - 2$
- **Welch's t (unequal variances, recommended):**
  $t = (\bar{x}_1 - \bar{x}_2) / \sqrt{s_1^2/n_1 + s_2^2/n_2}$
  $df \approx (s_1^2/n_1 + s_2^2/n_2)^2 / [(s_1^4 / (n_1^2 (n_1-1))) + (s_2^4 / (n_2^2 (n_2-1)))]$ (Welch–Satterthwaite approx.)

Assumptions: samples are independent (except paired test), underlying populations approximately normal (t robust for moderate n), variance assumption differs by test.

---

6. F-test
Purpose (simple): Compare two population variances. Also used in ANOVA to compare means across multiple groups.

Two-sample variance test:
$F = s_1^2 / s_2^2$
$df_1 = n_1 - 1$, $df_2 = n_2 - 1$
Reject $H_0: \sigma_1^2 = \sigma_2^2$ if F is too large (or too small depending which variance is numerator). Because F is positive and asymmetric, use appropriate one- or two-sided critical values.

ANOVA (one-way) — relation to F:

- H_0: all group means equal.
- Between-group variability: MS_between = SS_between / (k − 1)
- Within-group variability: MS_within = SS_within / (N − k)
- F = MS_between / MS_within
  Reject H_0 if F > F_{α, k−1, N−k}.

Assumptions (ANOVA/F-test): independent observations, normality in each group, homogeneity of variances (equal variances).

---

7. Chi-Square (χ^2) Tests
a) Goodness-of-fit test — checks if observed categorical frequencies follow a specified distribution.
Statistic: $\chi^2 = \Sigma (O_i - E_i)^2 / E_i$
df = k − 1 − m (k categories, m parameters estimated from data)
Reject H_0 if χ^2 large.

b) Test of independence (contingency table) — checks if two categorical variables are independent.

- Build contingency table with r rows and c columns, observed counts O_{ij}.
- Expected counts: E_{ij} = (row_i_total * col_j_total) / N.
- χ^2 = Σ_{i=1..r} Σ_{j=1..c} (O_{ij} − E_{ij})² / E_{ij}
- df = (r − 1)(c − 1)
  Reject H_0 if χ^2 large.

Assumptions: expected counts E_i typically ≥ 5 (rule of thumb); observations independent.

---

8. Bayesian Testing
Philosophy: Treat hypotheses or parameters as random and use prior beliefs + observed data to compute posterior beliefs. Decision-making uses posterior probabilities or Bayes factors instead of p-values.

Bayes' theorem (for hypotheses H_0 and H_1):
Posterior odds = Prior odds × Bayes factor (BF)
BF_{10} = P(data | H_1) / P(data | H_0)

Posterior probability of H_1:
$P(H_1 \mid \text{data}) = [\, P(\text{data} \mid H_1)\, P(H_1) \,] \,/\, [\, P(\text{data} \mid H_0)\, P(H_0) + P(\text{data} \mid H_1)\, P(H_1) \,]$

Bayes factor interpretation (rough):

- $BF < 1/10$: strong evidence for H_0
- $BF \approx 1$: data do not prefer either hypothesis
- $BF > 10$: strong evidence for H_1
  (Thresholds vary—these are conventional guidance.)

Advantages: direct probability statements about hypotheses, incorporate prior information, naturally penalizes model complexity (in many settings).
Disadvantages: requires choosing priors (can be subjective), computation can be intensive, results depend on prior choice.

Bayesian credible interval vs frequentist confidence interval: credible interval gives direct probability that parameter lies in interval (given prior and data); confidence interval has a different frequentist interpretation.