The Hypothesis Testing Playbook

Stacey Hancock

January 24, 2022

Contents

1 The Big Picture

Statistical inference is the process of inferring a conclusion about a population or model using information from a random sample. It usually comes in two forms: a hypothesis test (significance test) or a confidence interval.

The logic behind **hypothesis testing** follows these three steps:

1. Define a set of competing hypotheses:

   Null hypothesis (H_0): "no effect", "no difference", "nothing going on", "status quo", etc.
   Alternative hypothesis (H_a): the research hypothesis of some sort of effect – usually what the researchers are hoping to show

2. Determine what we would expect to see in sample data *if H_0 were true*, either through simulation or using mathematical probability distributions.

3. Compare the observed data to what we would expect to see under H_0, and quantify how likely we would have seen data like ours. If our data are unusual, we have evidence that H_0 may not be a valid assumption.

A **confidence interval** is calculated in these three steps:

1. Calculate a statistic from your data that estimates the quantity of interest.

2. Determine the margin of error, the maximum amount you would expect your statistic to vary from the quantity of interest in a specified percent of all samples.

3. Add and subtract the margin of error from your statistic to produce the interval.

This guide is a "playbook" of introductory statistical inference. Designed to be concise, it serves as a "cheatsheet"—a reference to use after you have been introduced to statistical inference at a deeper level.

2 The Language of Statistical Inference

Statistical inference has its own set of specific terms, many of which have a different meaning than one might expect. This section provides a list of those terms and their common statistical definitions.

2.1 Glossary

alternative hypothesis (H_a): Usually the research hypothesis; of the form "H_a : parameter $\neq$ hypothesized value", "H_a : parameter $<$ hypothesized value", or "H_a : parameter $>$ hypothesized value". May also be expressed as a model equation.

conclusion: Statement assessing the strength of evidence for the research hypothesis (alternative hypothesis) based on the p-value, in context of the problem.

confidence interval: An interval of values such that, prior to data collection, the interval had a specified probability of capturing the parameter.

confidence level ($1 - \alpha$): The probability, prior to data collection, that a confidence interval will capture the parameter. That is, the proportion of all possible samples in which the confidence interval calculated from that sample contains the parameter.

decision: A decision about the null hypothesis based on the p-value: either "Reject H_0" (for small p-values) or "Fail to reject H_0" (for non-small p-values).

margin of error: A value that measures how far away a specified percentage, typically 95%, of statistics lie from their mean. This is typically equal to a "critical value" (quantile from a particular distribution) multiplied by the statistic's standard error.

null hypothesis (H_0): Hypothesis of no effect or no difference; of the form "H_a : parameter $=$ hypothesized value". May also be expressed as a model equation.

one-sided hypothesis: An alternative hypothesis is one-sided if it states that the value of the parameter is strictly greater than the hypothesized value, i.e., H_a : parameter $>$ hypothesized value, or if it states that the value of the parameter is strictly less than the hypothesized value, i.e., H_a : parameter $<$ hypothesized value.

parameter: A numerical summary measure of the entire population or random process of interest, e.g., population mean.

population: The entire group of individuals/units to which our research hypothesis applies.

power ($1 - \beta$): The probability of rejecting the null hypothesis. If the null hypothesis is true, then the power is equal to the significance level, α.

practical significance: Results of a study are practically significant if the observed difference or effect in the sample would have a meaningful impact in context of the discipline.

p-value: The p-value is the probability of seeing our observed statistic or one more extreme (in the direction of H_a), assuming that H_0 is true; colloquially, the probability of my data under the null.

sample: The group of individuals/units on which we collect data.

sampling distribution: A probability distribution of a statistic as it varies across all possible samples.

significance level (α): A cut-off value α for which we reject H_0 if the p-value is less than or equal to α (and fail to reject H_0 otherwise).

statistical significance: Results of a study are statistically significant if we reject H_0 based on our p-value.

statistic: A numerical summary measure of the observed data, e.g., sample mean.

standard deviation: A value that, colloquially, measures how far you might expect a variable to lie from its mean, on average. We can calculate a standard deviation of a variable across individuals (e.g., salaries), or a standard deviation of a statistic across samples (e.g., average salaries).

standard error: A value that estimates the standard deviation of a statistic.

test statistic: A numerical summary measure of the observed data that measures how far the data are away from what we would expect to see under the null hypothesis.

two-sided hypothesis: An alternative hypothesis is two-sided if it states that the value of the parameter could be either larger or smaller than the hypothesized value, i.e., H_a : parameter $\neq$ hypothesized value.

Type 1 error: A Type 1 error occurs if the test decides to reject H_0 (significant evidence for H_a), but H_0 is actually true. If H_0 is true, the probability of a Type 1 error is equal to the significance level, denoted by α.

Type 2 error: A Type 2 error occurs if the test fails to reject H_0 (no significant evidence for H_a), but H_a is actually true. If H_a is true, the probability of a Type 2 error is denoted by β, so the power of the test is $1 - \beta$.

2.2 Common Statistical Symbols

Table 1: Parameters

Symbol	Description
μ	Population/model mean
σ	Population/model standard deviation
p	Population/model proportion or probability
β_0, β_1	Population regression model coefficients (intercept, slope)
ρ	Population/model correlation coefficient

Table 2: Statistics

Symbol	Description
n	sample size
$\bar{x}, \bar{y}$	Sample mean
s	Sample standard deviation
$\hat{p}$	Sample proportion
b_0, b_1	Estimated regression model coefficients (intercept, slope)
R (or r)	Sample correlation coefficient

3 Scenarios and Calculations

At an introductory level, most statistical methods you will encounter will be covered under one of the following scenarios:

1. **One proportion (one-sample z-test or exact binomial test)**: One binary categorical response variable (no explanatory variables)
2. **Difference in two proportions (two-sample z-test)**: Two binary categorical variables
3. **Two-way table (chi-squared test of independence/homogeneity or Fisher's exact test)**: Two categorical variables (with any number of levels)
4. **One mean (one-sample t-test if one variable; paired t-test if paired data)**: One quantitative response variable (no explanatory variables), or paired quantitative response variables
5. **Difference in two means (two-sample t-test)**: One quantitative response variable and one binary categorical explanatory variable
6. **Analysis of variance (ANOVA)**: One quantitative response variable and one categorical explanatory variable with more than two levels
7. **Simple linear regression**: Two quantitative variables
8. **Multiple linear regression**: One quantitative response variable and several explanatory variables (either quantitative, or categorical coded as dummy variables)
9. **Logistic regression**: One binary categorical response variable and one or several explanatory variables (either quantitative, or categorical coded as dummy variables)

A randomization/simulation-based hypothesis test can be conducted for all the above scenarios. We summarize normal-based inference here, which relies on the Central Limit Theorem (unless the response variable itself has a normal distribution), and thus requires large samples. How large of a sample you need depends on how far away your data are from a normal distribution in the first place.

General form of a **test statistic**: $\quad \text{test statistic} = \dfrac{\text{statistic} - \text{null value}}{\text{null standard error}}$

General form of a **confidence interval**:

$$\text{statistic} \pm \text{margin of error} \; = \; \text{statistic} \pm (\text{critical value}) \times (\text{standard error})$$

where the critical value is a quantile from the sampling distribution of the standardized statistic.

3.1 Summary of Normal-based Hypothesis Test Mechanics

Categorical Response Variable

Scenario	Parameter	Statistic	Test Statistic	Distribution of Test Statistic Under H_0	Standard deviation of statistic	Standard error of statistic
One-sample z-test	p	$\hat{p}$	$\dfrac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$N(0,1)$	$\sqrt{\dfrac{p(1-p)}{n}}$	$\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
Two-sample z-test	p_1-p_2	$\hat{p}_1-\hat{p}_2$	$\dfrac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}_P(1-\hat{p}_P)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ $\hat{p}_P$ = pooled sample proportion	$N(0,1)$	$\sqrt{\dfrac{p_1(1-p_1)}{n_1}+\dfrac{p_2(1-p_2)}{n_2}}$	$\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
Chi-squared test	N/A	N/A	$X^2=\sum\dfrac{(\text{obs}-\text{exp})^2}{\text{exp}}$ obs = observed count exp = expected count $=\dfrac{(\text{row sum})(\text{col sum})}{n}$	$\chi^2((r-1)(c-1))$ r = no. rows c = no. cols	N/A	N/A

Quantitative Response Variable

Scenario	Parameter	Statistic	Test Statistic	Distribution of Test Statistic Under H_0	Standard deviation of statistic	Standard error of statistic
One-sample t-test	μ	$\bar{y}$	$\frac{\bar{y}-\mu_0}{s/\sqrt{n}}$	$t(n-1)$	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$
Paired t-test	μ_{diff}	$\bar{y}_{diff}$	$\frac{\bar{y}_{diff}}{s_{diff}/\sqrt{n}}$	$t(n_{diff}-1)$	$\frac{\sigma_{diff}}{\sqrt{n}}$	$\frac{s_{diff}}{\sqrt{n}}$
	$diff$ = quantity calculated using paired *differences* in response variable					
Two-sample t-test (unpooled)	$\mu_1-\mu_2$	$\bar{y}_1-\bar{y}_2$	$\frac{\bar{y}_1-\bar{y}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	$t(n^*)$	$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
				n^* = Satterthwaite approximate df		
ANOVA	$\mu_1,\mu_2,\ldots,\mu_k$	$\bar{y}_1,\bar{y}_2,\ldots,\bar{y}_k$	$F=\frac{MS_{btwn}}{MSE}$	$F(k-1,N-k)$	–	–
	k = no. categories		MS_{btwn} = variance between groups			
	$N=\sum_{i=1}^{k} n_i$		$=\sum_{i=1}^{k}(\bar{y}_i-\bar{y})^2/(k-1)$			
			MSE = variance within groups			
			$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2/(N-k)$			
Simple linear regression	β_0,β_1	b_0,b_1	$\frac{b_i}{SE(b_i)}$	$t(n-2)$	–	–
Multiple linear regression	$\beta_0,\beta_1,\ldots,\beta_{q-1}$	$b_0,b_1,\ldots,b_{q-1}$	$\frac{b_i}{SE(b_i)}$	$t(n-q)$	–	–
	q = number of coefficients					

3.2 Normal-based Hypothesis Testing in R

Scenario	R function
One-sample z-test	`prop.test`
Two-sample z-test	`prop.test`
Chi-squared test for two-way tables	`chisq.test`
One-sample t-test	`t.test`
Paired t-test	`t.test`
Two-sample t-test	`t.test`
ANOVA	`aov`
Simple linear regression	`lm`
Multiple linear regression	`lm`