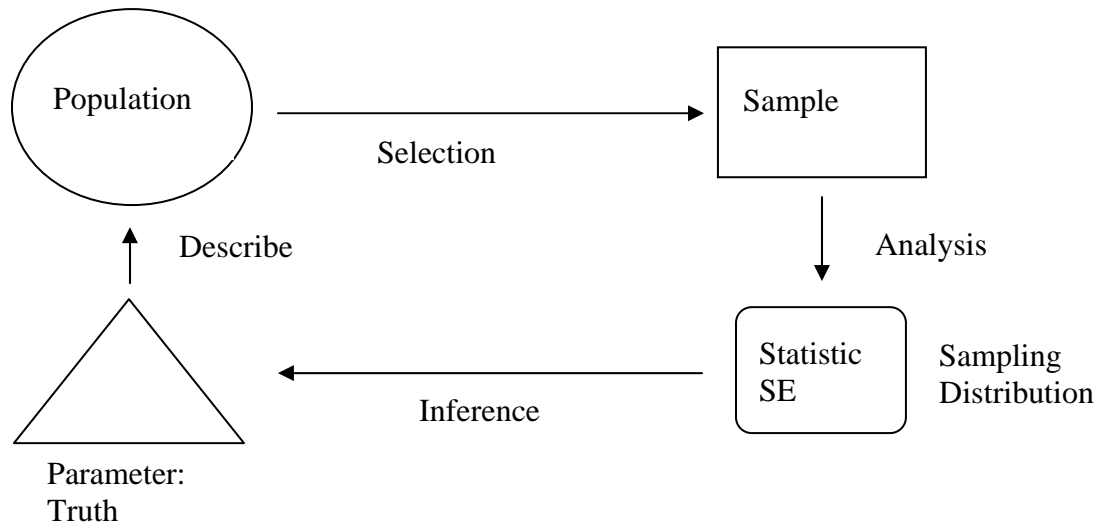


Research Paradigm

A model for research



The Basis of Inferential Statistics

Sampling distribution theory – Central limit theorem

Based on the known properties of the Normal Distribution $\sim N(\mu, \sigma^2)$

$$\text{Estimate: } \bar{X} = \frac{\sum X_i}{n}$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad SE(\bar{X}) = \sqrt{\frac{\sigma^2}{n}}$$

Parameter: θ

Sample: n

Estimator: $\hat{\theta}$

$$SE(\hat{\theta}) = \sigma_{\hat{\theta}}$$

$$\text{CI for } \theta: \hat{\theta} \pm c_{\alpha} \sigma_{\hat{\theta}}$$

A statistical paradigm

Model Building

Specify the model for your situation; the most important step

Estimation of Parameters

Getting results; the third most important step

Testing Fit of the Model

Consistency between the data and the model; the second most important step

Building Linear Models for Data Analysis

1. Theory

Always begin with theory. Develop an argument, supported by previous literature (could combine several different sources) and add a personal touch.

2. Model specification (outcome[s], explanatory variables)

Define all factors/variables involved in the theory. Draw a diagram of the relationships among the variables. Specify the outcome(s), explanatory variables, mediating/moderating variables, potentially confounding variables. Argue causality based on the design; beware of the term “predictor.”

3. Measuring Variables (reliability and validity)

Using standardized instruments versus self-constructed instruments. Standardization population should be recent, representative, and relevant. Self-constructed instruments must be piloted and evaluated. Most direct measurement possible is best.

4. Data collection: sampling (random, convenient, purposive)

Affects some statistical manipulations; most assume samples are randomly drawn from an identifiable population. A given statistic may not be dependent on sampling method, but the inference is always dependent on sampling method and research design.

Estimation of Parameters

1. Factors in the model can be fixed or random

Fixed factors are variables in which the data in your sample represent all possible levels (scores, groups, treatments, behaviors, conditions) in the population to which you generalize.

Random factors are variables in which the data in your sample represent a subset of levels from the population (which is infinite) sampled with a know model – and you wish to generalize to the entire population of all possible levels.

2. General Linear Model Assumptions

- a. Structural assumptions allow us to interpret the results
 - i. Observations are independent
 - ii. Variables are linearly related
 - iii. Explanatory variables are independent
 - iv. Explanatory variables are measured without measurement error
 - v. The right variables are in the model (argumentation: confounds, misspecification)
- b. Stochastic assumptions allow us to test parameter estimates
 - i. Errors have a mean of zero
 - ii. Errors have constant variance; homogeneity of variance
 - iii. Errors are normally distributed
 - iv. Errors are independent
 - v. Errors are independent of explanatory variables

If the X variables are thought of as random, then we make the general assumption that the joint distribution of Y and the Xs is multivariate normal. If the X variables are fixed, we assume the conditional distributions of Y (given the Xs) are independently and normally distributed. Moderate departures of either set of assumptions are tolerable.

Testing the Fit of the Model & Related Issues

1. Parsimony

The simpler model is better

2. Correlations

Squared correlations tell you the % of variance explained (coefficient of determination)

3. Simple Regression

R is the correlation between the outcome and the explanatory variable.

R^2 is the same as the squared correlation between the outcome and explanatory variable. It is a variance accounted for statistic.

Remember, $R^2 = \frac{SS_{Regression}}{SS_{Total}}$. To test the hypothesis that $R^2 = 0$, $F = \frac{\hat{R}^2 / k}{(1 - \hat{R}^2) / (n - k - 1)}$,

with k and $n-k-1$ *df*. This is equivalent to $F = \frac{MS_{Regression}}{MS_{Residual}}$.

Also: check the size of the Standard Error of the Estimate (standard deviation of residuals).

4. Multiple Regression

Here, R is the multiple correlation between the model-predicted scores \hat{Y} and the observed scores Y .

R^2 is the squared multiple correlation; the percent of variance explained in the outcome by the linear combination of explanatory variables. Standard Error is analogous to the size of the average error of prediction, the standard deviation of the residuals as above.

5. Analysis of Variance

Eta-squared, η^2 , is an estimate of the maximum squared correlation between the independent variable and the dependent variable – it can be treated as any squared correlation, the proportion of variation accounted for.

$$\eta^2 = \frac{SS_{\text{between}}}{SS_{\text{total}}}$$

Eta-squared is generally biased upward when based on sample data. Omega-squared is an adjusted value that is better for most purposes.

In Analysis of Variance, the test of mean differences is one of whether the variation between groups is greater than the variation remaining within groups

$$H_0 = \sum_{j=1}^J (\mu_j - \mu)^2 = 0 \quad F = \frac{MS_B}{MS_W}$$

6. Controlling overall Type-I error rate (α)

Compute a test-wise α to control the overall study-wise α when conducting multiple tests on the same data: $1 - \sqrt[c]{1 - \alpha} \leq \alpha$; where c is the number of statistical tests or contrasts conducted, and α is the test-wise Type-I error rate used to determine statistical significance for each statistical test on the same data set.

Statistical Modeling

Observations $X_i = T_i + E_i$ true (systematic) & error (random) components

The notation of Regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_Q X_{Qi} + \varepsilon_i$$

where:

Y is the outcome

$\beta_0 \dots \beta_Q$ are parameters

$X_{1i} \dots X_{Qi}$ are known constants, explanatory variables

ε_i are iid

$i = 1, \dots, n$

In terms of estimation:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\varepsilon}_i$$

$$Y_i = \hat{Y}_i + \hat{\varepsilon}_i$$

**Graph of regression – conditional distributions...

**Graph of single observation and relations with predicted and mean value...

$$SS_{\text{total}} = SS_{\text{regression}} + SS_{\text{residual}}$$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

The notation of Analysis of Variance

$$Y_{ij} = \mu_j + \varepsilon_{ij} \qquad Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

where:

Y is the outcome

α_j are parameters

ε_i are iid

$i = 1, \dots, n$

In ANOVA we condition Y on X partitioning the distribution by X

**Graph of conditional distributions...

$$SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}}$$

$$\sum_{i=1}^n (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^n (\bar{Y}_{\cdot j} - \bar{Y}_{..})^2 + \sum_{i=1}^n (Y_{ij} - \bar{Y}_{\cdot j})^2$$

The notation of the General Linear Model

$$\mathbf{y} = \mathbf{X} \mathbf{b} + \mathbf{e}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

When we employ a design matrix in Analysis of Variance, the model is parallel to the regression model with a matrix of explanatory variables.

A data matrix will contain rows of cases and columns of variables; for example:

ID	SAT	GPA	Gender	IQ
1	560	3.0	1	112
2	780	3.9	0	143
3	620	2.9	0	124
4	600	2.7	1	129

$$\mathbf{y} = \begin{bmatrix} 560 \\ 780 \\ 620 \\ 600 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 3.0 & 1 & 112 \\ 3.9 & 0 & 143 \\ 2.9 & 0 & 124 \\ 2.7 & 1 & 129 \end{bmatrix}$$

Multivariate Regression, ANOVA

Multiple Outcomes, generally correlated
Assumption: multivariate normality

$$\mathbf{Y} = \mathbf{X} \mathbf{B} + \mathbf{E}$$