1 Study Design

Aims of study design: draw unbiased conclusions and provide estimates.

Bias is a lack of accuracy.

Study units, cases, subjects : units where data are obtained. One observation per unit.

Response/ Dependent variable : variables of interest that are dependent

Explanatory/ Independent variables : variables used to explain and predict the response variables

Confounding variables : other dependent variables not interested, related to both exp. and resp. vars. Control or eliminate them.

1.1 Validity and biases

A biased study includes measures not accounted that influences the response variable.

To minimize biases

Comparison comparing with a control group, placebo, current best (current standard treat.), natural groups

Control restrictions limits, protocols being systematic, blinding (single and double), eliminate or holding constant confounding variables

Randomization Randomize representative subjects, randomize subject assigned to subject groups

Confounding factors (lurking variables) are related to treatment group, causal rel. with outcome, making Correlation is not causation.

To reduce confounding

Distribution Random subject groups, fair distribution of characteristics

Randomization of treatment order (practice eff.)

Restriction on experimenters

Blocking small mini-experiments of common characteristics

Block when you can, randomize otherwise.

1.2 Precision

A precise study has close and more confident estimates with small error.

To maximize precision

Blocking Divide study units into blocks of similar characteristics. Randomize treatment within block. (Stratification for observational studies)

Replication increasing total subjects sampled, or repeating measures within study group. NOT repeatability or reliability. Linked with degree of freedom. More replicates = Higher df = More complicated models

Balance equally sized study groups (minimize SE with same total samp. size)

Matched pair, twins study are extreme levels of blocking.

1.3 Study types

Observation studies have data collected through observations. Subjects decide the group they are in. Cannot generate a causal link. Evidence by observation.

Observation studies problems

Selection bias Surveys may be biased, or self selection of subjects

Reporting bias groups are biased in responding/ reporting (diag. bias)

Question wording

Confounding factors not accounted

Designed experiments have the experimenter deliberately impose treatment to study groups. The experimenter decides the group subjects are in. Can prove causation. Evidence by design.

Designed experiments can better randomize, block, to reduce biases and confounding variables.

A completely randomize design has no matching, usually done with mechanical or computer randomizers.

2 Exploratory Data Analysis

Used to: Discover important data features, Improve understanding of underlying population, Transform data into information.

2.1 Variable types

Hierarchy of information (Least ⟶ Most info)

Categorical nominal Groups

Categorical ordinal Ordered groups

Numerical discrete Ordered, scale component

Numerical continuous Most informative

Questions we can ask

Categorical Category, mode, association

Numeric Mean, variance, min, max, median, outliers

Distribution features

Shape Symmetrical, skewed, (right tail is positive skewness) or unimodal, multimodal

Location mean, median

Spread variance, IQR

Unusual outliers, groupings

Relationships correlations (for numericals), associations (for categorical)

2.2 Graphical displays

N	C	Displays
1	0	dotplot, histogram, boxplot
0	1	Table (with percentages) or bar-chart
2	0	Scatterplot, boxplot of differences
1	1	Comparative dotplot, boxplot
0	2	Contingency tables (with %) or comparative bar charts
3	0	Surface Plot
2	1	Grouped Scatterplot
1	2	Interaction plot
0	3	Cross tabulation or comparative bar chart

2.3 Categorical

Figures using tables or barplots.

Simpson's paradox is a phenomenon where the trend exhibited within each group changes when combining all groups. Caused by imbalanced group sizes.

2.4 Numerical

Figures using dotplots for individual data, boxplot to summarize, comparative boxplot for comparisons, histograms for large dataset.

Figures

Histogram Divide range into bins, height of each bin is the number of data points within the range

Boxplot Five number summary of quartiles. Outliers are represented by crosses

Location

Mean The statistical average

$$\mu = \bar{x} = \frac{1}{x}\sum x_i$$

Order statistic The observations sorted by value. $x_{(i)}$ is the ith ordered statistics. Can infer quartiles

Quartiles Lower quartile is $x_{(\frac{n+1}{4})}$, upper quartile is $x_{(\frac{3(n+1)}{4})}$. Median is $x_{(\frac{n+1}{2})}$. Use linear interpolation if the index is fractional.

Spread

Range Largest - Smallest. Sensitive to outliers

IQR Middle 50% of data (used for skewed data):

$$IQR = Q_3 - Q_1$$

Std Dev Measures the consistency that observations are to the mean. The sample standard deviation is

$$\hat{\sigma} = s = \sqrt{\frac{1}{n-1} \times \sum (x_i - \bar{x})^2}$$

with the $n-1$ accounting for samples' df

Outliers Check if: a legit data value, entry mistake, or belonging in another population group. Quantitatively, outlier if observation is 1.5 IQR from Q_1 or Q_3.

Chebyshev's inequality states that for most distributions, at least 75% of data points are within 2sd. For normal distributions, the 68-95-99.7 rule. Squaring std dev to get variance.

Use medians and IQR when dataset is skewed.

2.5 Several numerical

Figures using scatterplots or dotplots on differences for data pairs.

Correlation is the strength (magnitude) and direction (sign) of a **linear** relationship between two random variables:

$$r = \frac{1}{n-1}\sum \frac{x-\bar{x}}{s_x}\frac{y-\bar{y}}{s_y}$$

Attributes of r

- From -1 to 1
- Affected by outliers and unitless

r must be used in conjunction with a scatterplot, for it can lie about the plot shape.

The covariance is the unstandardized correlation

$$\begin{aligned}\text{Cov}(X,Y) = s_x s_y r &= \frac{1}{n-1}\sum (x-\bar{x})(y-\bar{y}) \\ &= E(X)E(Y) - E(XY)\end{aligned}$$

3 Randomness models

Studying how sample statistics are generated by hypothetical parent populations.

Distributions to model

Empirical distribution The variable samples we see, observed distribution

Hypothetical distribution Probability model to mimic empirical distribution

A random process is an event where a single trial outcome is unpredictable, but becomes predictable over many trials.

Probability of an event is the relative frequency of its occurrence in an infinite sequence of trials.

Symbol: $\Pr(\text{event})$

In experimental data, randomness arise from: assigning subjects to treatments, sampling of subjects, measurement errors.

3.1 Random variables

A numeric variable with value determined by the outcome of a random process. Has an unknown value before the random process (population), and observed value after the process (sample). Observation is realization of the random variable.

A random variable is completely specified by its probability distribution: summarizing probabilities associated with all possible outcomes.

Discrete Has countably many possible outcomes. Values represent counts.

Probability mass function.

Continuous Any value in an interval. Values represent measurements.

Probability density function. The area under the pdf between two values is the probability of a sample to be within the interval.

The cumulative mass (density) function is

$$F(x) = \Pr(X \leq x)$$

and the interval probabilities are

$$\Pr(a \leq X \leq b) = F(b) - F(a-1) = F(b) - F(a)$$

3.2 Distribution properties

The mean (expected value) is

$$E[X] = \sum x_i f(x_i) = \int x f(x)\, dx$$

where

- Does not have to be observable
- Point of symmetry with symmetrical distribution
- $E[aX + bY] = aE[X] + bE[y]$

The pth percentile of the distribution is the x where p% of population falls below this value:

$$\Pr(x \leq X) = p$$

The median is the 50th percentile.

The variance measures the spread of a population distribution.

$$\sigma^2 = \sum f(x)(x-\mu)^2 = \int f(x)(x-\mu)^2\, dx$$

Useful form: $\sigma^2 = E[x^2] - \mu^2$

where

- The variance must be positive
- $V[aX + b] = a^2 V[X]$
- If X and Y are independent

  $$V[aX + bY] = a^2 V[X] + b^2 V[Y]$$

  Otherwise

  $$V[X + Y] = V[X] + V[Y] + 2\operatorname{Cov}(X, Y)$$

The standardized random variable has a mean of 0, and standard deviation of 1.

$$Z = \frac{X - \mu}{\sigma}$$

3.3 IIDRVS

Consider the sum of a series of independent, identically distributed random variables (IIDRVS) X

$$S_n = \sum X$$

then

$$E[S_n] = nE[X] \qquad V[S_n] = nV[X]$$

The central limit theorem states that the sum of a large number of IIRDRVS is approximately normal

$$S_n \sim N(n\mu, \sqrt{n}\sigma)$$

hence the prevalence of the normal distribution.

The sampling distribution of sample means is

$$\overline{X} = \frac{1}{n} S_n$$

with

$$E[\overline{X}] = E[X] \qquad V[\overline{X}] = \frac{1}{n} V[X]$$

Laws of Large number: As the sample size increases, the sample mean (and std) tends to the population parameter.

3.4 Independent events

Independent events are unrelated events. Events A and B are independent when

$$\Pr(A \cap B) = \Pr(A)\Pr(B)$$

Independent events implies that

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \Pr(A)$$

Two events are positively associated if observing one event increases the probability of seeing another

$$\Pr(A|B) > \Pr(A)$$

and negatively associated when the other way around.

4 Probability distributions

Uniform distribution has equal probability outcomes.

A normal distribution has

$$X \sim N(\mu, \sigma)$$

1. Bell shaped, extends indefinitely; 2. Symmetrical around mean (= median = mode); 3. Inflection point at $\pm\sigma$; 4. Sum independent normal variables is normal

The standardized normal distribution denotes the number of std dev away from the mean:

$$Z \sim N(0, 1)$$

where for any $X \sim N(\mu, \sigma)$

$$\Pr(X \leq x) = \Pr(Z \leq \frac{x - \mu}{\sigma})$$

A Bernoulli trial is a random process where outcomes are either successes or failures. The parameter p is the probability of success.

A binomial distribution is the sum of n Bernoulli trials. Its conditions are

- Fixed number of trials
- Independent trials
- Constant probability of trial's success

Binomial properties

$X \sim Bi(n, p)$

$\mu = np \qquad \sigma = \sqrt{np(1-p)}$

Adding independent binomial variables with SAME probability of success

$X \sim Bi(n, p), Y \sim Bi(m, p)$

$X + Y \sim Bi(n + m, p)$

Normal approximations for binomial distribution applies when the expected number of success and failures are larger than 5.

Continuity corrections account for rounding to a continuous approximation of a discrete one. We take the values $x \pm 0.5$ instead, depending on the scenario.

4.1 Normality Tests

Qualitatively use the 68-95-99.7 rule.

Normal probability plots is an empirical cdf plot of the values against the scaled y-axis. If the distribution is normal, the scaled cdf should be a straight line. If the distribution is right skewed, there is a clump of values at the left; if it is left skewed, there is a clump of values at the right. A probability plot is a QQ plot with its axis flipped.

QQ plot is a normal scores plot that checks for normality. Steps are

1. Compute the ordered statistics

2. Find the empirical cdf with the formula: $F(X \leq x_(i)) = \frac{i}{n+1}$ there will be a total of $n + 1$ steps

3. Use the inverse z-table to find the z-value corresponding to the probabilities. These are the normal scores

4. Plot the points (z, x) with the ordered statistics

If the data is normally distributed, the QQ plot forms a line with equation: $x = \sigma z + \mu$

Quantitative normality checks are better because it is easier to check if points are close to a line.

5 Statistical Modeling

Data consists of signal, noise, and dirt.

1. **Signals** are relationship explaining variables
2. **Noise** are random variations
3. **Dirt** are mistakes

Statistical models have the form

$y_i = f(x_i) + e_i \qquad e \sim N(0, \sigma)$

the signal part is deterministic and explained, the noise part is random and unexplained.

Stages of statistical modeling

1. Formulate model;
2. Estimate parameters;
3. Check model assumptions;
4. Estimate quantities of interest and inference

Modeling process

The estimated error is the residual between the predicted and observed value

$\hat{e} = y - \hat{y}$

Error = population, residual = sample. Estimate s_{error} with $s_{\hat{e}} \approx s_e$.

The residual standard deviation is

$$s_{\hat{e}} = \sqrt{\frac{\sum (e_i)^2}{n - \nu}}$$

where ν is the number of estimated parameters in our deterministic equation.

Extrapolation is guessing.

Model assumptions

Independent errors, normally distributed, with zero mean and constant standard deviation. Same distribution throughout model. Mean error is ALWAYS zero.

Suitable deterministic part.

Check for normality of residuals using a probability plot, verify independence by looking at the study design. Check constant std with residual vs fit scatterplot.

Examples

No explanatory Null model Use the sample mean. $s_{residual} = s_{sample}$:

$$y_i = \bar{x} + e_i, e \sim N(0, \sigma)$$

Numeric Use a polynomial fit

$$y_i = \alpha + \beta x_i + \cdots + e_i, e \sim N(0, \sigma)$$

Categorical Apply grouped fits, an index for each category fit

$$y_{ij} = f_j(x_{ij}) + e_{ij}, e \sim N(0, \sigma)$$

6 Sampling Distribution

Modeling the inferred population parameters from a sampling distribution.

The sampling distribution of sample means $\overline{X}$ is an estimator for the population mean

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

the distribution is normal only when the sample size is large, due to the CLT.

Terms

Parameters are popu.'s summary measures

Statistics are sample's summary measures

Estimators are random variables that are sets of all estimates of parameters

Estimate is a realization of an estimator

The sampling variability is a parameter that measures the spread of the sampling distribution. The standard error is the estimated sampling variability using the sample standard deviation:

$$SE = \frac{s}{\sqrt{n}}$$

For proportions, the estimator is $\hat{P}$, a random variable of all sample proportions of size n. Approximately normal with large sample sizes

$$\hat{P} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$$

6.1 Probability Interval

The interval where we have p% confidence that the sample mean will fall within.

With known population standard deviation, this is

$$\bar{x} \pm z\sigma_{\overline{X}}$$

where z is the level of confidence (for 95%, $z = 1.96$).

Without known population standard deviation, use standard error, and the t-distribution (n-1).

This is the same for proportions after normal approximations, except we need to use a z distribution instead.

6.2 Prediction Interval

The interval where one sample from the population is likely to be observed.

If the population parameters are known, this is

$$\bar{x} \pm z\sigma\sqrt{1 + \frac{1}{n}}$$

due to the sampling variance of the sample mean.

Without the population standard deviation, replace σ with s, z with t_{n-1}.

These are the same for proportions, except we use a z-distribution regardless.

7 Confidence Interval

Confidence interval on the population mean has the form

$$\text{estimate} \pm \text{distribution} \times \text{variability}$$

It provides a range of plausible parameter values.

Distribution refers to the certainty of estimate range and sampling distribution

Variability is the sampling variability

Margin of error is the half-width of the CI.

With known population std dev, our random interval (p% confidence)

$$\overline{X} \pm z\sigma_{\overline{X}}$$

will contain the population mean p% of the time. Its realization using the sample mean is a p% CI.

With unknown population std, random interval:

$$\overline{X} \pm t_{n-1}SE$$

This CI works exactly when the population is normally distributed. If no strong skew, t-distributions produce conservative estimates. It fails with outliers or skewness when sample size is small.

This is the same with proportions, except that we must use a z-distribution.

T-distribution rules

- Normality of population data
- When sample size less than 15, use t only when data is normal
- When sample size is at least 15, use t unless outliers or major skewness
- For larger sample size (>= 40), t is fine even for skewed populations

The t-distribution t_ν is symmetrical, mean zero, bell shaped, but more spread out. As $\nu \to \infty$, the distribution approaches normal.

Confidence interval assumptions

- Independent samples (randomisation in data collection)
- Normal population when sample size is small. For proportion, check np and n(1 - p) $\geq$ 5 and constant p.

7.1 Sample size

To compute the sample size required to get a certain MOE, we invert the algebra and solve for n with known population std dev.

With known approximate population proportions, we do the same algebra. Without any estimates, we produce a conservative sample size by using $p = 0.5$ and solve for n, for this proportion maximizes the possible CI.

8 Hypothesis Testing

H-test tests the hypothesis that our sample population has parameters conforming to the null hypothesis.

Terminologies

Null hypothesis is the current best hypothesis

Alternative hypothesis is testing hypothesis

Tails one tail tests for one-sided inequalities; two tail tests for general inequalities

Test statistic is the standardized sample mean under H_0

P-value is the level of extremeness observed under H_0

H-test Process

1. State hypothesis in terms of the parameter, state the significance level α
2. Assume null hypothesis, compute SE of std dev of sampling distribution
3. Compute test statistics, determine p-value
4. Reject or retain H_0
5. State conclusion in terms of the problem

The test statistic is

$z = \frac{\bar{x}-\mu}{\sigma_{\overline{X}}}$ $\quad$ $t = \frac{\bar{x}-\mu}{SE}$

1-tail tests: P-value = Pr getting the sample mean or more extreme. 2-tail tests: P-value = $Pr_{extreme} \times 2$.

Reject H_0 when P-value < α or test stat more extreme than critical value.

Trade-off between Type I (α) and Type II (β) errors: Changing α to decrease one will increase another.

Statistical significance does not equal to actual importance, for it depends on the significance level and true mean. Large p-value does not imply that H_1 must be false.

Statistical Power

The probability to reject H_0 when H_0 is false.

Factors increasing: Increasing sample size, α, mean difference; Decreasing estimator std dev.

1-tailed has higher power than equiv. 2-tailed.

Power curve graphs power against differences between population means. On power curve: Bottom point is α; Symmetric implies 2-sided test.

Assumptions are

- Normality of Sampling Distribution
- Samples are independent

Details included in the reports:

1. Study type, sample size;
2. Significance and p-value, distribution value;
3. Rejection or retaining H_0;
4. CI or point estimate of the true mean;
5. Conclusion in context of the study

For paired data, conduct H-test on the sample differences. H_0: $\mu_{diff} = 0$.

For proportions, we use the std dev of sampling distribution from H_0 always and H_0 p for normality approximations. The rest is the same. We can also use the exact binomial distribution to do h-test, for the normal approximation does not account for continuity corrections. Worse approximation error with extreme p and smaller samples.

H-test on median

For skewed data, median is chosen as centre. Simply convert the hypothesis of median: H_0: median = v, H_1: median > v, to the corresponding proportion hypothesis: H_0: $p = 0.5$, H_1: $p > 0.5$ (p is the proportion > v), and apply H-Test.

9 Comparative Inference

Paired (dependent) samples

Subtract for the difference (eliminate confounding vars) and carry out the usual H-test and CI.

Two independent populations

Assumptions: 1. Independent observations and groups (randomized experiment); 2. Normality of sample means (normal pops or large sample size) (Prob plots for both groups); 3. Unequal/ equal variances ($0.5 < (\frac{s_1}{s_2})^2 < 2$).

Equal variances

Test equal assumption: $0.5 < (\frac{s_1}{s_2})^2 < 2$.

Use 1 std (same as s_{residual} for cat. mean model):

$$s_{\text{pooled}} = \sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}}$$

$$SE(\bar{X}_1 - \bar{X}_2) = s_{\text{pooled}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$CI = \bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2}SE$$

Degree of freedom $= n_1 + n_2 - 2$. Estimate $= \bar{x}_1 - \bar{x}_2$. Apply H-test as usual to test the difference with $t_{df above}$

Unequal variances

Used only when fail the equal test. Use both stds to find the SE:

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

Degree of freedom $= \min(n_1 - 1, n_2 - 1)$. Use t_{df} with the SE above. Note: As t is used with 1 estimated s only, this becomes an acceptable approximate, not exact. The df is lower, showing higher variability.

Reliable for similar sample sizes and distribution shapes. Same **assumptions** as 1-sample-t, using combined sample sizes.

NEVER use 2 independent pop samples for paired data (assumptions don't hold + no removal of confounding vars as in paired data).

Two proportions Always z-distributed. Assumptions: 1. Independent groups and observe.; 2. Binomial distribution; 3. Normal approx. : Large sample sizes, with +5 cases with and without quality in each samples.

CI with SE:

$$SE(\hat{P_1} - \hat{P_2}) = \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

H-Test with SE, where H_0: $p_1 = p_2 = p_0$ (Equivalent to p_1 - $p_2 = 0$

$$\hat{p_0} = \frac{p_1n_1 + p_2n_2}{n_1+n_2}$$

$$SE = \sqrt{\hat{p_0}(1-\hat{p_0})(\frac{1}{n_1}+\frac{1}{n_2})}$$

Note: SE (and not sd) as we estimate p_0

10 ANOVA: +3 populations

One-way ANOVA The comparison between single mean model and multiple mean model.

Assumptions: 1. Normality of residuals (Probability plot: all groups at 1 or each group); 2. Independence of observations/ groups (Study design); 3. Constant residual variances ($0.5 < \frac{\max s}{\min s} < 2$)

Explained variability is the variability between. Unexplained variability is the variability within.

For H-test, $H_0 : \mu_1 = \mu_2 = ... = \mu_n$ and H_1: At least 1 μ different (H_0 is the Total row on table, rest is H_1). j: different group, i: element in each group, n: number of observations, k: number of treatments/params

Source	DF	SS
Group	$k-1$	$SS_g = \sum n_j(\bar{y}_j - \bar{\bar{y}})^2$
Error	$n-k$	$SS_e = \sum(y_{ij} - \bar{y}_j)^2$
Total	$n-1$	$SS_g + SS_e = \sum(y_{ij} - \bar{\bar{y}})^2$

$MS = SS/DF$, $f = \frac{MS_g}{MS_e}$, $R^2 = \frac{SS_g}{SS_{total}}$

$s_r = \sqrt{MS}$ (use Error row for separate mean, Total row for single mean)

Distribution $f \sim F_{g,\,e}$, always check larger tail for P-value = Pr(F > f)

Fisher Intervals

The pooled std dev is $s_p = \sqrt{MS_e}$. The df is always df_e.

The group population mean CI is

$$CI(\mu_j) = \bar{x}_j \pm t_{df_e} \times \frac{s_p}{\sqrt{n_j}}$$

The pairwise mean CI is (j1, j2 = any 2 groups)

$$CI(\mu_{j_1} - \mu_{j_2}) = \bar{x}_{j_1} - \bar{x}_{j_2} \pm t_{df_e} s_p \sqrt{1/n_{j_1} + 1/n_{j_2}}$$

Groups Comparison

Fisher sets Individual Confidence Level (CL): CL for population difference between 2 groups within the CI. Higher Type I errors.

Tukey sets Simultaneous Confidence Level (CL): CL that all paired CI contains the population differences simultaneously. Higher Type II errors.

Tukey Intervals has the benefits of: 1. Guaranteed SCL, so take into account number of groups; 2. More conservative (wider interval); 3. More appropriate with > 2 populations; 4. Tukey is consistent with the ANOVA H-test. Fisher might not.

Same as Fisher when there is only one pair.

Fisher Individual Test: If 0 is not inside the CI, the 2 groups are significantly different.

For treatment with same samples size (balanced), let the Least Significant Difference be the pairwise $t_{df}SE$. Two means are significantly different if they differ by more than LSD.

Summary diagram: SORT all the mean groups, draw underlines connecting groups that are NOT significantly different.

11 Linear Regression

Models a linear relationship between two continuous numeric variables. A type of ANOVA.

Simple linear regression tests the validity of the model: $y_i = \alpha + \beta x_i + e_i \qquad e_i \sim N(0, \sigma)$

Assumptions: normal residuals around line; independent random residuals/ observ.; equal variances of residuals; linear relationship (no pattern (funnel = log transform, curved: wrong relationship, unusual points) + straight boundary in residuals).

The regression line is

$$E(\hat{Y}|x) = \hat{\alpha} + \hat{\beta}x$$

$$\hat{\beta} = r\frac{s_Y}{s_X} = \frac{Cov(x,y)}{s_x^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

The line always crosses $(\bar{x}, \bar{y})$. SS_e is minimized. ANOVA table is constructed the same way with $df_g = 1$, $df_e = n - 2$. The model residual s_r is the spread of data around the line. $R^2 = r^2$, r CAN be +/-.

Parameter/Response Inference

The coefficients α and β are t_{n-2} distributed. We can conduct h-test and CI on them.

H_0: $\beta = \rho = 0$. H_1: H_0 is false. The f-ratio test checks for a significant linear trend.

The confidence and prediction interval on the response is

$$SE(fit) = \sqrt{\frac{s_r^2}{n} + (x - \bar{x})^2SE(\beta)^2}$$

$$CI = \hat{\alpha} + x\hat{\beta} + t_{n-2}SE(fit)$$

$$PI = \hat{\alpha} + x\hat{\beta} + t_{n-2}\sqrt{s_r^2 + SE(fit)^2}$$

Note: Since s_r is large compared to SE(fit), the PI is usually around max ±2 × s_r

Correctness

The lower the s_r the better, the higher the R^2 the better.

Outliers have high standardized residuals with large y, leverage with large x. Report both the analysis with and without the outliers.

For 2-samples equal varianced data, the 2-sample t, ANOVA, and regression produce identical: model residuals, p-value and conclusion, error df, and grouped mean CI and PI (also $t - val^2 = F$).

12 χ^2 Association Test

The χ^2 test tests for association between two categorical variables.

The null H_0 model is that the two variables are independent, the alternative H_1 model is that they are associated. Assuming H_0:

$$X_{ij} \sim B(n,p) \qquad p = \frac{cr}{n^2}$$

$$E[X_{ij}] = np = \frac{\text{row total} \times \text{col total}}{\text{grand total}}$$

The test statistic is

$$U^2 = \sum\frac{(\hat{x}_{ij} - x_{ij})^2}{x_{ij}} \qquad U^2 \sim \chi^2_{(r-1)(c-1)}$$

$$U^2 = \sum\frac{(observed - expected)^2}{expected}$$

The summation fraction part is each cell's contribution to the chi-squared. The t-test is performed on the test statistic using ONE-TAIL (>) p-values.

Associations

T-Test conclusions must identify cells with and without associations. Only report cells with high contribution and difference's directions.

Assumptions include: all expected frequencies ≥ 1; at least 80% of cells has ≥ 5 expected frequencies. If the assumptions fail, merge two similar columns or rows such that it holds again.

Goodness of Fit

Use χ^2 to test if the observed data fits expected proportion, with H-test as above. Find expected value based on H_0 (expected val = expected proportion × n) and df = num groups - 1

H_0: All expected proportion; H_1: Not H_0, there is a change