1 Study Design

Principles of good study design: draw unbiased conclusions and provide precise estimates.

1. Study units, cases, subjects, units where data are obtained. One observation per unit.
2. Response/Dependent variable, variables of interest that are dependent
3. Explanatory/Independent variables, variables used to explain and predict the response variables
4. Confounding variables, variables related/causes both the response and explanatory variables. We wish to control or eliminate them.

1.1 Validity and biases

Bias is a lack of accuracy. A biased study includes variables not accounted for that influences the response variable.

Principles to minimize biases

1. Comparison, comparing with a control group, placebo, current best treatment, natural groups
2. Control, restrictions/limits, protocols being systematic/consistent, blinding (single and double), eliminate or holding constant confounding variables
3. Randomization, randomize representative subjects, randomize subject assigned to subject groups

Confounding factors (lurking variables when unobserved) are related to treatment group, causation with outcome. It is the reason that correlation does not imply causation.

To reduce confounding, make the subject groups similar in respect to the variables

1. Random subject groups, fair distribution of characteristics
2. Randomization of treatment order
3. Restriction on experimenters
4. Blocking, creating 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

1. Blocking/Stratification, divide study units into blocks of similar characteristics (confounding variables). Randomize treatments within block.
2. Replication, increasing total subjects sampled, or repeating measures within study groups. Not repeatability or reliability. Increases degree of freedom, allows more complicated models
3. Balance, equally sized study groups, to minimize SE with similar sample sizes

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

1. Selection bias, surveys may be selection biased, or self-selection of subjects
2. Reporting bias, groups are biased in responding/ reporting
3. 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.

1. Display/Graph sample data
2. Summarize distribution of sample data
3. Describe stats, graph information, and summarize
4. Conjecture about the population

2.1 Variable types

Hierarchy of information (Least to Most info)

1. Categorical nominal, groups
2. Categorical ordinal, ordered groups
3. Numerical discrete, scale component
4. Numerical continuous, most informative

Questions we can ask

1. Categorical: Category, mode, association
2. Numeric: Mean, variance, min, max, median, outliers

Distribution features

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

Center mean, median

Spread variance, IQR

Unusual outliers, groupings

Relationships numeric correlations, categorical associations

To describe some data points, consider its shape, center, spread and outliers.

2.2 Graphical displays

N	C	Displays
1	0	Dotplot, histogram, boxplot
0	1	Table (with %) or barchart
2	0	Scatterplot
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 Variables

Figures: tables or barplots.

Simpson's paradox: a phenomenon where the trend exhibited within each group changes when combining all groups. Caused by imbalanced group sizes, where the most frequently sampled values get over-proportionally valued.

2.4 Numerical Variables

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

1. Histogram, divide range into bins, height of each bin is the number of data points within the range
2. Boxplot, five number summary of quartiles. Outliers are 1.5IQR above Q_3 or below Q_1 and represented by crosses

Center

1. Mean, The statistical average
$$\mu = \bar{x} = \frac{1}{x}\sum x_i$$
2. Order statistic, The observations sorted by value. $x_{(i)}$ is the ith ordered statistics.
3. 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

1. Range, Difference between largest and smallest value. Sensitive/Include to outliers
2. IQR, Middle 50% of data:
$$IQR = Q_3 - Q_1$$
Used when dataset is skewed
3. Std Dev, Measures the consistency that observations are to the mean, the expected deviation of observation to mean. The sample standard deviation is
$$\hat{\sigma} = s = \sqrt{\frac{1}{n-1} \times \sum (x_i - \bar{x})^2}$$
4. 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: for most distributions, at least 75% of data points are within 2SD. Normal distributions with the 68-95-99.7 rule. Squaring std dev to get variance.

Use medians and IQR when dataset is skewed/non-normal.

2.5 Several numerical

Figures: scatterplots or dotplots on differences for paired data.

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

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

Correlation Attributes

1. From -1 to 1, with -1 and 1 indicating perfect correlation
2. Positive indicating positive correlation
3. 0 implies no linear association
4. Affected by outliers and unitless

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

We can also compare means (or medians) between two groups, or the ratio between their variances (or IQRs).

The covariance is unstandardized correlation:

$$\text{Cov}(X,Y) = s_x s_y r = E[XY] - E[X]E[Y]$$

3 Randomness models

Studying how sample statistics are generated by hypothetical parent populations.

Random Distributions

Empirical distribution Created by the samples we see, observed distribution

Hypothetical distribution Probability model that generates empirical distribution

Random process: an event where a single trial outcome is unpredictable.

Regression to the mean: averages of random processes becomes predictable over many trials.

Probability of an event is the relative frequency of its occurrence in an infinite sequence of trials. Denoted as $\Pr(E)$

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 (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.

Discrete probability distribution defined by a probability mass function. The pmf cannot be negative, and the sum of probabilities across the event space is 1.

Continuous: Any value in an interval. Values represent measurements.

Continuous probability distribution defined by probability density function. The area under the pdf between two values is the probability of a sample to be within the interval. The pdf cannot be negative, and the area under the graph is 1.

The cumulative mass (density) function is

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

and the interval probabilities are (disc. cont.)

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

3.2 Distribution properties

The mean (expected value) is

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

where

1. Does not have to be observable
2. Point of symmetry with symmetrical distribution
3. $E[aX + bY] = aE[X] + bE[Y]$

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 is

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

where

1. $\sigma^2 = E[x^2] - \mu^2$
2. The variance must be positive
3. $V[aX + b] = a^2 V[X]$
4. 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

The sum of a series of independent, identically distributed random variables (IIDRVS) on 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

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

The sampling distribution of sample means is

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

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

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 probability decreases.

4 Probability distributions

Uniform distribution has equal probabilities.

A normal distribution is

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

1. Bell shaped, extends indefinitely towards both directions
2. Symmetrical around mean, median equal to mean and mode
3. Inflection point at $\pm\sigma$
4. Sum of 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 outcome is either success or failure. The parameter p is the probability of success.

A binomial distribution is the sum of n Bernoulli trials. It requires:

1. Fixed number of trials
2. Independent trials
3. Constant probability of trial's success

Binomial properties

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

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

Adding independent binomials with same success prob ($X \sim Bi(n, p), Y \sim Bi(m, p)$)

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

Normal approximations for binomial distribution requires the expected successes and failures to equal or greater than 5.

Continuity corrections account for rounding in 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: empirical cdf plot of the samples with a 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 hill on the left; if it is left skewed, there is a hill on the right. A probability plot is a QQ plot with its axis flipped.

QQ plot (normal scores plot) plots the samples against the normal scores:

1. Compute the ordered statistics
2. Find the empirical cdf with the formula

$$F(x_{(i)}) = \frac{i}{n+1}$$

   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_i, x_i) with the ordered statistics

If the data is normally distributed, the QQ plot forms a line with equation

$$x_i = \sigma z_i + \mu$$

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.

Signals are relationship explaining variables

Noise are random variations

Dirt are mistakes

Statistical models have the form

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

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

5.1 Modeling process

Predicted values: $\hat{y}$ are values predicted from the deterministic equations. Observed values: y are sampled statistics. Estimated error is the residual between the observed and predicted values: $\hat{e} = y - \hat{y}$

Error: the underlying model error. Residuals: error within a specific sample.

Residual standard deviation to estimate the population error standard deviation $s_{\hat{e}} \approx s_e$.

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

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

Extrapolation is an act of faith. Interpolation is acceptable. Model performance is based on the residual standard deviation (smaller the better).

Model assumptions

1. Normal (probability plot), random, and independent (study design) errors/residuals. Zero mean and constant std dev (residual vs fitted)
2. Suitable deterministic part. Same residual distribution throughout fitted values.

Least squared model guarantees zero residual means.

Model Types

1. Null model, use the sample mean. The residual standard deviation is the sample standard deviation:

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

2. Numeric, use a polynomial fit

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

3. Categorical, Apply grouped fits, an index for each category fit

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

We wish to get estimates and predictions out of our models, and decide the simplest and most appropriate model.

6 Sampling Distribution

Models the distribution of inferred population parameters from all samples.

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

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

from CLT, $\overline{X}$ is normal when the sample size is large or X is normal.

The law of large numbers: the sample mean is a better estimate to the population mean as the sample size increases. This is due to the decreasing sampling distribution std dev.

Sampling Terms

1. Parameters are summary measures of the population (means)
2. Statistics are summary measures of the sample data (sample means)
3. Estimators are random variables containing all estimates of a parameter ($\overline{X}$)
4. Estimate a realization of an estimator ($\bar{x}$)

Sampling variability: a parameter that measures the spread of the sampling distribution.

Standard error: the estimated sampling variability using the sample std dev:

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

$\hat{P}$ is the estimator for proportions, a random variable of all sample proportions of size n. Its distribution is approximately normal with normal approximation conditions

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

6.1 Probability Interval

The interval with p% probability that the sample mean will fall within. (Knowing pop mean)

Knowing population standard deviation, this is

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

Not knowing population standard deviation, this is

$$\mu \pm t_{n-1}\,\text{SE}(\overline{X})$$

Proportions is identical after normal approximations, using a Z distribution instead.

6.2 Prediction Interval

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

Knowing population std dev, this is

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

due to the sampling variance of the sample mean.

Not knowing the population std dev, we need

$$\bar{x} \pm t_{n-1}s\sqrt{1 + \frac{1}{n}}$$

Proportions is identical after normal approximations, using a Z distribution instead.

7 Confidence Interval

Inference is the use of sample data to infer population parameters.

Against a population parameter, a point estimate is the best guess using a point, an interval estimate is the best interval guess.

CI provides a range of plausible parameters

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

1. Distribution refers to the confidence level and distribution of sampling distribution
2. Variability is the sampling variability
3. Margin of error is the half-width of the CI. It's the furthest a plausible parameter can be from the sample mean.

Increasing confidence level increases the CI range. Increasing sample size decreases the CI range.

CIs are realizations of random intervals with p% chance of capturing the mean.

Knowing population std dev, it is

$$\text{CI} = \bar{x} \pm z\sigma_{\overline{X}}$$

With unknown population std dev, it is

$$\text{CI} = \bar{x} \pm t_{n-1}SE$$

Same with proportions after normal approximations, with a z-distribution.

CI is exact when the sampling distribution is normal. Either way, if there are no strong pop skew, t-dist produce conservative estimates due to: CLT, law of large numbers. It fails with outliers or skewness when sample size is small.

T-distribution assumptions/rules

- Exact when population is normal
- 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 ($\geq$ 40), t is fine even for skewed populations

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

Confidence Interval assumptions

- Random and independent samples
- Normal population when sample size is small

7.1 Sample size

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

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

8 Hypothesis Testing

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

H-test Terminologies

1. Null hypothesis is the current best hyp
2. Alternative hyp is the hyp we are testing
3. One tail tests for one-sided inequalities; two tail tests for general inequalities
4. Test statistic is the standardized estimate under H_0
5. P-value is the level of extremeness observed under H_0
6. Critical value is the standardized significance level

H-test Process

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

The test statistic is

$$z = \frac{\bar{x} - \mu}{\sigma_{\overline{X}}} \qquad t_\nu = \frac{\bar{x} - \mu}{SE}$$

In one tail tests, the p-value is the probability of getting the sample mean or more extreme. In two tail tests, multiply the one-tail prob by 2.

If the p-value is less than α, or if the test-statistic is more extreme than our critical value (one-tail), we reject H_0. Otherwise, retain H_0.

H-test decisions

1. True positive, correctly rejecting H_0, its probability is the statistical power
2. True negative, correctly retaining H_0
3. False positive, Type I error, incorrectly rejecting H_0, with probability of α
4. False negative, Type II error, incorrectly retaining H_0, with probability of β

Trade-off between Type I and Type II errors. Changing α to decrease one will increase another. Increasing sample size reduces both errors.

Statistical significance is not actual importance, as it depends on α 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.

One-tailed tests have higher power than equivalent two-tailed tests.

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

H-test Assumptions

1. Random and independent samples
2. Normal sampling distributions

H-test report

1. Study type, sample size
2. Significance, p-value, test stat and dist.
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 the same inference on the sample differences. Often, H_0 is that the difference is zero.

For proportions, find SE from H_0 proportions

and use H_0 proportions for normality assumptions. We can also use the exact binomial distribution to do h-test, for the normal approximation doesn't account for continuity corrections. This error is worse in approximating true p-values with more extreme p and smaller samples.

For skewed data, use a sign-test on the median. Convert the hypothesis on medians m to a proportions hypothesis:

$$H_0 : m = v, H_1 : m > v$$
$$\to H_0 : p = 0.5, H_1 : p > 0.5$$

where p is the sample proportions greater than v, and conduct h-test on the proportions.

9 Comparative Inference

For paired and dependent samples, conduct inference on the sample differences (to eliminate confounding variables).

For two independent populations, assume

1. Independent samples and populations (Study design), Similar sample sizes.
2. Normal sampling dist (Separate probability plots)
3. Equal/unequal variances

Equal variances

Assumption is $0.5 < (\frac{s_1}{s_2})^2 < 2$.

Inference uses the pooled SD (from s_r in modeling)

$$s_p = \sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}}$$
$$\text{SE}(\bar{X}_1 - \bar{X}_2) = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

The distribution is $t_{n_1+n_2-2}$ for both CI and h-test test statistics.

Unequal variances/Known population std dev

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

The distribution is $t_{\min(n_1-1,n_2-1)}$ or z for both CI and h-test. The t_ν distribution is a conservative approximation.

Use combined sample sizes in normal approximation assumptions identical to 1-sample t.

Two proportions, Assumptions

1. Independent, random samples
2. Binomial distributions, normal approximations (large samples)

The distribution and test-statistic is Z.

For confidence intervals,

$$\text{SE}(\hat{P}_1 - \hat{P}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

For H-Test, where $H_0 : p_1 = p_2 = p_0$

$$\hat{p}_0 = \frac{p_1 n_1 + p_2 n_2}{n_1+n_2}$$
$$\text{SE}(\hat{P}_1 - \hat{P}_2) = \sqrt{\hat{p_0}(1-\hat{p_0})(\frac{1}{n_1}+\frac{1}{n_2})}$$

10 ANOVA

One-way ANOVA

The comparison between single mean model and separate means model.

Assumptions

1. Normality of separate residuals (prob plot)
2. Independent samples within and between groups (study design)
3. Constant residual variances (res vs fitted)

Constant residual requires $0.5 < \frac{\max s}{\min s} < 2$.

Explained variability is between group variability. Unexplained variability is within group variability.

$H_0 : \mu_1 = \mu_2 = \ldots, H_1 :$ At least 1 μ different

Total row is H_0, Error row is H_1.

j are group indexes, i are element indexes. n is total samples, k is number of groups.

Source	DF	Sum of Squared
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	$n-1$	$SS_g + SS_e = \sum(y_{ij} - \bar{\bar{y}})^2$

$MS = SS/df$, $f = \frac{MS_g}{MS_e}$, $R^2 = \frac{SS_g}{SS_t}$

Model residual std dev are $s_r = \sqrt{MS}$. Error row for separate means, total for single mean.

Test-statistic dist is $f \sim F_{k-1,n-k}$. $f = t_\nu^2$ from an equal-variance 2-sample t test-stat.

Fisher Intervals

The pooled std dev is $s_p = \sqrt{MS_e}$. Use the Error degrees of freedom.

The separate group population mean CI is

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

The group pairwise mean CI is

$$CI(\overline{X}_j - \overline{X}_k) = \bar{x}_j - \bar{x}_k \pm t_{df_e} s_p\sqrt{1/n_j + 1/n_k}$$

Fisher Individual Test

Fisher intervals sets the individual confidence level, the CL for each pairwise CI. Higher Type I errors. Lower SCL.

Tukey intervals sets the simultaneous confidence level, the CL for all simultaneous pairwise CI. Higher Type II errors. Higher individual CI.

Tukey intervals have the benefits of

1. Guaranteed SCL, accounting all pairs and invariant to group numbers
2. More conservative intervals
3. More appropriate with > 2 populations
4. Consistent with one-way ANOVA

Tukey same as Fisher with two groups.

LSD Test

If the pairwise CI doesn't contain 0, the groups are significantly different.

For balanced groups, the least significant different is $t_\nu SE$. Two means are significantly different if their sample means differ by more than LSD.

Summary diagram

Sort the group means ascending, draw underlines connecting none significantly different groups.

11 Linear Regression

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

Simple linear regression tests the model

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

Assumptions

1. Normal, equal variant residuals around line (res vs fit)
2. Independent residuals (study design)
3. Linear relationship (scatterplot)

The regression line is

$$E(\hat{Y}|x) = \hat{\alpha} + \hat{\beta}x$$
$$\hat{\beta} = r\frac{s_y}{s_x}$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Line always crosses $(\bar{x}, \bar{y})$. SS_e is minimized.

ANOVA table has $df_g = 1$ and $df_e = n-2$. The model residual $s_r = \sqrt{MS_e}$ is the error spread around the line. $R^2 = r^2$.

Parameter/Response Inference

Both $\alpha, \beta \sim t_{n-2}$. Used for h-test and CI.

The null model is no linear relationship. Alternative is that H_0 is false. F-ratio tests for significant linear trends.

$$H_0 : \beta = \rho = 0, H_1 : H_0 \text{ false}$$

The CI and PI on responses are ($s_r \gg \text{SE}(f)$)

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

Correctness

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

Outliers have high standardized residuals when their y is far from line, they have leverage when their x is away from $\bar{x}$. They influence both model correctness and regression coefficients. Report both the analysis with and without the outliers.

Equal variant 2-sample 2-tail t, ANOVA, and regression produce identical: s_r, p-value and conclusion, error df, group mean CI and PI.

12 χ^2 Test

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

H_0 is that the two variables are independent. H_1 is that H_0 is false

Assuming H_0, the expected probability of each cell is the product of its row and column proportions:

$$X_{ij} \sim B(n,p), p = \frac{cr}{n^2}, E[X_{ij}] = np = \frac{cr}{n}$$

The test statistic is

$$U = \sum\frac{(x_{ij} - E[x_{ij}])^2}{E[x_{ij}]} \quad U \sim \chi^2_{(r-1)(c-1)}$$

The summation fraction is each cell's contribution to chi-squared. The expected counts should add to the row and column counts, with identical proportions.

Associations

Cells or lines of cells with large contributions have large differences between expected and observed counts. The direction of difference can identify trends and associations. Ignore ≤ 1 differences.

The chi-squared test produces a normally approximated chi-squared test-statistic.

2-sample 2-tail proportion h-test produces same results as 2x2 chi-squared tests. Same p-value, for sample statistics, $z^2 = U$.

Assumptions

1. Expected counts all ≥ 1
2. At least 80% of cells expected counts ≥ 5
3. Independent samples within cells

If assumption fails, merge similar rows or columns til it succeeds.

Goodness of Fit

Tests if observed data fit expected proportions.

H_0 is all expected proportion. H_1 is that H_0 is false.

The test-statistic U is computed with the squared residuals against expected proportion counts. The distribution is χ^2_{n-1}.

Association is deduced by the direction of cell differences.