Contents

1 Introduction

Population and sample groups. We sample from population using probability, and infer the population from the sample using statistics.

Aims to collect, organize, and summarize data from samples of populations, and make statistical inferences.

Parametric statistics assumes a statistical model about the population feature, which is a parameterized distribution with estimates from the sample. Non-parametric statistics works with ranked data and does not assume a specific distribution or parametric model.

1.1 Random Samples

Data collected from real population through random samples which we assume are representative. We model variables of interests from the population as a random variable with some distribution. Our random sample thus contains independent observations of the random variable.

A random sample is a set of IIDRV $X_1, \ldots, X_n$ of the population variable of interest X. Due to independence, its pdf is

$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = f_X(x_1) \ldots f_X(x_n)$$

Realizations are denoted by $x_1, \ldots, x_n$.

A statistic T is a function ψ of the random sample

$$T = \psi(X_1, \ldots, X_n)$$

which is another random variable. Its realization t is the function applied to the random sample realization

$$t = \psi(x_1, \ldots, x_n)$$

It is implicitly assumed that the statistic depends only on the random sample, and is not affected by any other unknown parameters.

More definitions

- Parameters are unknown constants of the population distribution
- Parameter space is the set of possible parameter values.
- Estimators (point estimators) are statistics that are used to estimate parameters, they are random variables
- Estimates (point estimates) are realizations of estimators
- Hat notation, if T is an estimator for θ, we refer to the estimator as $\hat{\Theta}$

1.2 Data Collection

Randomized controlled trials (designed study) have participants randomly assigned to treatment or control groups, receiving either treatments or placebos (or standard treatment). It requires randomization, and ethics approval is required. Primary research interest is causality. Confounding variables can complicate analysis, correlated factors hard to separate the influences. Gold standard to inferring causality.

Observational studies (cohort study) observe individuals and measure variables of interest without influencing or assigning treatments.

Case control studies identify individuals with a particular outcome (case) and individuals without the outcome (control), which are matched on factors considered to be important. It then looks back to compare past exposures to risk factors of the outcome.

Surveys collect data from a sample of individuals using standardized questions.

1.3 Sampling Methods

Simple random sampling lets every item in the population to have equal probability of being in the sample. Realistically we would sample unit by unit without replacement.

Stratified random sampling partitions the population into subsets (strata) based on variables of interests, then conduct simple random samples within each stratum. The population is heterogeneous, while each stratum is homogeneous, so measurements vary less within units of the same stratum. This improves precision of information within each stratum.

Cluster sampling partitions the population into clusters, each with similar characteristics to the population. Then randomly select a subset of clusters and conduct simple random sample within the selected groups. A two-stage sampling where we first sample clusters where units in clusters are grouped physically together, then sample from the selected few clusters.

Unrepresentative samples are also quite common

- Convenience sampling chooses members that are easy to reach or measure
- Voluntary response sampling, selection based on volunteers instead of researchers
- Systematic sampling, selecting items by starting at a random point, then selecting the other items at a regular interval

1.4 Data types

Categorical data, each unit can be assigned to a particular group. Nominal data means the groups has no ordering; ordinal data means that the groups have a natural or predetermined order.

Numeric data, each value is a numeric measurement. Discrete data has the data obtained by counting, the values can be finite or countably infinite. Continuous data can take values in a continuous range.

For numeric data, interval scale implies that differences between measurements are meaningful, but the scale has no absolute zero. Its data can have arbitrary zero points and can be negative. Ratio scale has a true zero point and the ratio (and also differences) between measurements are meaningful. Its values cant go below the zero point.

Different data types yield different representations and statistical techniques.

1.5 Probability Concepts

Probability underlies statistical models and methods.

- Random variables and realizations corresponds to estimators and estimates
- Discrete pmf and continuous pdf
- Pdf as a likelihood
- Cdf and the distribution function
- Pdf and cdf relationship
- Expectation, variance, covariance, correlation coefficient

For a continuous rv X, the qth quantile is a number c_q such that $P(X \leq c_q) = q$. Hence, $F_X(c_q) = q$ and $F_X^{-1}(q) = c_q$. This is also called the $(100q)$th percentile.

The 50th percentile is the median $m = c_{0.5}$, the 25th and 75th percentiles are the first and third quartiles, $Q_1 = c_{0.25}, Q_3 = c_{0.75}$.

Mgf to calculate distributions of sums of independent rv, and to compute moments.

Common discrete and continuous distributions, pdf, pmf, means and variances. Standard normal and transformations.

Approximation formulas for mean and variables of $\phi(X)$

$$\begin{aligned} E[\phi(X)] &\approx \phi(E[X]) + \frac{1}{2}\phi''(E[X])V[X] \\ V[\phi(X)] &= \phi'(E[X])^2 V[X] \end{aligned}$$

The law of large numbers, $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and $\mu = E[X], \sigma^2 = V[X]$.

$$\bar{X} \to \mu \quad \text{as} \quad n \to \infty$$

Defined in terms of convergence in probability.

The central limit theorem

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0,1) \quad \text{as} \quad n \to \infty$$

Defined in terms of convergence in distribution.

2 Data Visualization

Data graphics are paragraphs of data, the analogy between graphics and paragraphs are

- Graph intends to communicate a cohesive message about the data (paragraph)
- Component parts that the eye can decode separately (sentences)
- Small details can be examined for specific information (word)

It is very easy to create graphs that do not communicate well. Reasons are

- Designer too engaged with the data at graph creation
- Stupid software features
- Careless use of software
- Decorations and aesthetics instead of graphical clarity

2.1 Principles of Good Graphics

Five principles of good graphics

- Show the data clearly
- Use good alignment on a common scale for comparable quantities
- Use simplicity in design
- Use transparent visual encoding
- Prefer standard forms demonstrated to be effective

To show the data clearly, identify the source of the data, the choice of graphic should depend on its purpose, show the data on the graph, avoid distractions and distortions, and the labels on the title, axes, and data points should be informative and well-chosen.

Eyes are better at differentiating the size of things lined up on a common linear scale (bar chart over pie chart). This is empirically backed; all best standard forms follow this principle.

Design simplicity, avoid 3d effects, shapes, or additions that block effective communication. The lack of context of pictographs may communicate misleading information. The data to ink ratio (pixels used for data over the total non-background pixels) measures the density of information in the representation. Aim for high data-ink ratios.

Graphs involve encoding data, we aim to make the visual decoding as simple as possible. Transparent visual encoding implies that viewers barely notice the data encoding. However, good graphs can also contain substantial content.

2.2 Standard Forms

We should always prefer standard forms for data visualization.

Time series plot shows the value of a variable changing over time. A very common plot. Time on the horizontal axis, points joined by lines. Also known as line plots.

Bar chart shows the value of a numeric variable by levels of a categorical variable. Prefer horizontal bars, so labels are read naturally. Categories labeled on the category axis.

Dot plot shows the detail and distribution of a numeric variable. Use for small sample sizes (under 100). Can be extended to show more than one group.

Histogram shows the distribution of a numerical variable. Values are grouped into equal width intervals, the height of the bar represents the frequency or relative frequency in the interval. The number of intervals must be chosen sensibly. There are flexibility in the construction of a histogram due to interval sizes and grouping. Used to display the shape of distribution.

Boxplots are abbreviated distributions of a numerical variable. Divides data into 4 intervals: lower quartile, median, upper quartile. Min and max are shown as whiskers. Unusual points are shown separately as outliers. There can be categorical boxplots for comparison, even with color. The boxplots can also be horizontal.

Unusual values warrant investigation, if there are no problems, we must include them in your consideration for they might provide insights. It is an error to remove outliers from datasets for analytic convenience.

Extreme/unusual data points are flagged as outliers. A mild outlier lies more than 1.5 IQR below Q1 or above Q3. It is beyond the inner fence of $(Q_1 - 1.5IQR, Q_3 + 1.5IQR)$. An extreme outlier is more than 3 IQR below Q1 or above Q3. It is beyond the outer fence of $(Q_1 - 3IQR, Q_3 + 3IQR)$. This is a rule of thumb. Standard convention is for boxplots to highlight outliers. R defaults to highlighting mild outliers, its whiskers extend to only within the inner fence.

Scatter plot shows the relationship between two numeric variables. Conventionally, time is always on the horizontal axis, and response variables are on the vertical axis. We can use different symbols to indicate groupings on the same plot. A panel plot shows the groups separately on different panels and also with different symbols.

Standard forms summarized

- Time series for relationship of variable with time
- Bar chart for relationship of numerical variable against categorical
- Scatter plot for relationship between two numerical
- Dot plot for distribution of numerical variable, perhaps by a categorical variable, small sample
- Histogram for distribution of a numerical variable, large sample
- Box plot for main features of distribution of a numerical, by a categorical variable, not for showing distribution shape

3 Descriptive Statistics

The first step towards understanding the data, they are

- Location: mean and median
- Spread: std dev, IQR, range
- Shape: symmetry, skewness. Through pmf, pdf, df
- Other population characteristics: proportion of defects/outliers, quantiles

3.1 Location and Spread

On a sample of size n, we have

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_i x_i \\ s^2 &= \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 \\ s &= \sqrt{s^2}\end{aligned}$$

These are empirical moments of the random variable, or estimates of the population moments. They are based on observations rather than theory.

3.2 Shape

The empirical cdf or sample cdf is

$$\hat{F}(x) = \frac{1}{n} \sum_i I(x_i \leq x)$$

where I is the indicator function with value 1 if $x_i \leq x$ and 0 otherwise. It is a step function that jumps by $\frac{1}{n}$ for every passing of a data point.

If the underlying variable is discrete, we have a sample pmf corresponding to the sample cdf

$$\hat{p}(x) = \frac{1}{n} \sum_i I(x_i = x)$$

If the underlying is continuous, we estimate an empirical pdf by histograms or smoothed pdf. For histograms, $\hat{f}_h$ for bin length h, divides the range of values into bins, then counting the number of values within each interval. For each interval $[a, b)$, there is a rectangle of height

$$\hat{f}_h(x) = \frac{1}{hn} \sum_i I(a \leq x_i < b)$$

For a smoothed pdf, $\hat{f}_h$ for bandwidth parameter h

$$\hat{f}_h(x) = \frac{1}{hn} \sum_i K\left(\frac{x_i - x}{h}\right)$$

where K is a kernel that is non-negative, integrates to zero, and zero meaned. The bandwidth parameter h controls the level of smoothing.

3.3 Order Statistics

By arranging the samples in order of increasing magnitude, we have

$$x_{(1)} \leq x_{(2)} \cdots \leq x_{(n)}$$

then $x_{(k)}$ is the kth order statistic of the sample. It is a realization/observation of the kth Order Statistic $X_{(k)}$, a random variable. Over lots of samples of size n, the realization $x_{(k)}$ will be around $X_{(k)}$ according to its sampling distribution. The special ordered statistics are: $x_{(1)}$ as sample minimum, $x_{(n)}$ as sample maximum.

We can express the empirical/sample cdf using the order statistics

$$\hat{F}(x) = \begin{cases} 0 & x < x_{(1)} \\ \frac{k}{n} & x_{(k)} \leq x < x_{(k+1)} \\ 1 & x \geq x_{(n)} \end{cases}$$

3.4 Sample Quantiles

Define the pth sample quantile as the number $\hat{c}_p$ where $\hat{F}(\hat{c}_p) = p$. In R, there are 9 types of empirical quantiles. Type 1 assumes a probability of $1/n$ below $x_{(1)}$ and zero above $x_{(n)}$. Type 6 divides the line into $(n+1)$ intervals with equal probability. Type 7 is used by default on R summary and quantile, assigning zero probability both below $x_{(1)}$ and above $x_{(n)}$. They are mainly used for special estimations at smaller sample sizes, they all converge at larger samples.

Type 6 and 7 linearly interpolates the empirical cdf, while type 1 doesn't? Or is it discrete but assumed to be linear when computing quantiles.

Both the R summary and quantile outputs the five number summary, reporting the minimum, first quartile, median, third quartile, and maximum. We assume that the empirical cdf increases linearly within an interval between order statistics. Thus, the default R type 7 sample quantiles for $p \in [0, 1]$ are defined as

$$\hat{c}_p = x_{(k)} \quad \text{where} \quad k = 1 + (n-1)p$$

And if k is fractional, we linearly interpolate the closest order statistics.

Using R's default intuitive quantile definition, the sample median is

$$\hat{m} = \begin{cases} x_{((n+1)/2)} & \text{n is odd} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2)+1}) & \text{n is even} \end{cases}$$

this matches all but type 1's R definition of sample quantiles. The sample median is preferred as a measure of location for data that is highly skewed.

The sample IQR is the distance between the first and third sample quartiles

$$I\hat{Q}R = \hat{Q}_3 - \hat{Q}_1 = \hat{c}_{0.75} - \hat{c}_{0.25}$$

The IQR is the preferred measure of spread for highly skewed data. The sample IQR estimates the population IQR.

3.5 QQ plots

Tells if a sample is from a given distribution by finding a typical sample from said distribution and plotting it against our sample.

Convention for QQ plot plots the real data on the y-axis and typical population sample on the x-axis. The ordered data points are $(t_{(k)}, x_{(k)})$ where $t_{(k)}$ are the ordered typical population sample. If the distribution is correct, the plot should be a straight line. Switching the axes results in a probability plot.

The typical population sample of n samples assumes each of the $(n+1)$ interval to have equal probability, this is R type 6 quantiles. They are converted to the population quantiles using the population cdf F

$$t_{(k)} = c_{(\frac{k}{n+1})} = F^{-1}(\frac{k}{n+1})$$

Our sample are assumed to have type 6 estimates of the population quantile

$$x_{(k)} = \hat{c}_{(\frac{k}{n+1})}$$

The points in the QQ plot are

$$\{F^{-1}(\frac{k}{n+1}), x_{(k)}\}$$

note that $t_{(k)}$ and $x_{(k)}$ are both quantiles of values from the random variable.

Normality testing QQ plots are called Normal quantile plots. These plots does not assume a population mean or variable, and we can use the typical population sample of the standard normal. Intuitively, if $X \sim N(\mu, \sigma^2)$, and the normal model is correct, then

$$x_{(k)} \approx \mu + \sigma\Phi^{-1}(\frac{k}{n+1})$$

hence the points

$$\{\Phi^{-1}(\frac{k}{n+1}), x_{(k)}\}$$

should be a straight line with intercept μ and slope σ. The typical standard normal samples $\Phi^{-1}(k/(n+1))$ are called normal scores.

In R, qqnorm produces a normal qq plot, and qqline creates a non-regression line that passes the 0.25 and 0.75 population and sample quantile.

There are h tests on normality which utilize the distances between the empirical cdf and fitted population cdf.

4 Point Estimator

We often know the form/distribution of the population X, but is missing its parameters. Estimating the population distribution is then estimating those unknown parameters from a random sample.

Suppose the distribution of X depends on the parameter θ. Then its cdf is $F(x;\theta)$, pdf is $f(x;\theta)$ or pmf $p(x;\theta)$, mean $\mu(\theta)$ and variance $\sigma^2(\theta)$.

Given a random sample on X, a point estimator of θ is a statistic chosen so that it is likely to be close to θ. The realization of the estimator is a point estimate of θ.

The distribution of an estimator $\hat{\Theta}$ is its sampling distribution (for the statistic varies between each random sample). The sampling distribution depends on a specific population distribution and sampling scheme (assume always random sampling). The population standard deviation of the sampling distribution of $\hat{\Theta}$ ($sd(\hat{\Theta})$) likely depends on the parameter θ, so we define the standard error of the estimator ($se(\hat{\Theta})$) to be the estimated standard deviation of the sampling distribution using the realization of the estimator $\hat{\theta}$.

To select a good point estimator, its mean must be close to the parameter , and it should have a low standard deviation.

5 Properties of Point Estimators

Let T_n be an estimator for θ on a random sample of size n.

The desirable properties of point estimators are: unbiasedness, efficiency, and consistency.

- T_n is an unbiased estimator for θ if $E[T_n] = \theta$.
- T_n has a higher efficiency if it has a smaller variance. It is like T_n uses the sample data more efficiently to give a tighter estimate.
- T_n is consistent for θ if its sampling distribution approaches θ as the sample size $n \to \infty$. (Higher probability T_n is close to θ with large sample sizes)

5.1 Unbiasedness

Unbiasedness is a desirable feature, but a biased estimator does not rule it out to be a useful estimator if it also has high efficiency. To compare between estimators, we focus on how much they are from θ on average. The Mean Squared Error of an estimator for the parameter θ is

$$\text{MSE}(T_n) = E[(T_n - \theta)^2]$$

Note that MSE on an estimator is only defined relative to its use in estimating a parameter.

By definition, MSE for an unbiased estimator is simply its variance for $\theta = E[T_n]$. For biased estimators, consider

$$\begin{aligned}
\text{MSE}(T_n) &= E[(T_n - \theta)^2] \\
&= E[(T_n - E[T_n] + E[T_n] - \theta)^2] \\
&= E[(T_n - E[T_n])^2] + 2E[(T_n - E[T_n])(E[T_n] - \theta)] + E[(E[T_n] - \theta)^2] \\
&= V[T_n] + (E[T_n] - \theta)^2 \\
&= V[T_n] + bias(T_n)^2
\end{aligned}$$

because $E[T_n - E[T_n]] = 0$ and define $bias(T_n) = E[T_n] - \theta$.

In principle, we can choose the best estimator (biased or not) provided we have the components — by selecting the estimator with the lowest MSE. In practice, however, the bias term is unknown and is a function of the parameter θ. (We can try to remove the bias using another estimator B_n, but the combined estimator may have a greater MSE due to the additional variability of B_n).

Unbiasness is not usually preserved under transformations. Assume an unbiased estimator T_n for θ. Suppose that another parameter ν is a transformation of θ by $\nu = \phi(\theta)$. In general, $\phi(T_n)$ is not an unbiased estimator for ν because $E[\phi(T_n)] \neq \phi(E[T_n]) = \nu$ (unless the transformation is linear).

5.2 Efficiency

For the class of all unbiased estimators for parameter θ, there is a lower bound for the variance that no estimator can go below — the Rao-Cramer Lower Bound.

Define the efficiency of an unbiased estimator T_n as

$$\text{eff}(T_n) = \frac{RCLB}{V[T_n]}$$

and since $V[T_n] \leq RCLB$, the range of efficiencies are $0 \leq \text{eff}(T_n) \leq 1$.

If an estimator has an efficiency of one, it is an efficient estimator for θ as no other unbiased estimator could do better.

5.3 Consistency

We say that a sequence of random variables X_n converges stochastically to a constant c if and only if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - c| < \epsilon) = 1$$

This is also referred to as convergence in probability. We can denote this as

$$X_n \xrightarrow{P} c$$

Define the estimator T_n to be consistent for θ if T_n stochastically/probablistically converges to θ, namely $T_n \xrightarrow{P} \theta$. Note that the index of the estimator is always about the sample size, hence consistency is about the behavior of the statistic at larger samples.

Consistency is desirable for larger samples should provide more precise information abut θ.

For common statistics, this is pretty simple to establish. Refer to Chebysehv's inequality from probability theory.

6 Principles for finding good estimators

The main methods to find good estimators are: method of moments, method of maximum likelihood, and method of quantiles. The most important estimator is the MLEs, and other methods are backups.

6.1 Method of Moments

The MM estimator chooses the value of the parameter such that the theoretical moments equal the observed sample moments. For the simple case where $f_X(x;\theta)$ and $E[X] = \mu(\theta)$, the MM estimator is then $\hat{\Theta}_{MM} = \mu^{-1}(\bar{X})$.

In general

1. Given $X_1, \ldots, X_n$ with $f(x|\theta_1, \ldots, \theta_r)$
2. Let the kth moment be $\mu_k = E[X^k]$
3. The kth sample moment is $M_k = \frac{1}{n}\sum X_i^k$
4. Set $\mu_k = M_k$, and solve for $\theta_1, \ldots, \theta_r$

We can also use the variance instead of the second moment.

MM estimators are intuitive, and works in most situations. They are usually biased and non-optimal. We can use a bar to denote an MM estimator instead of a hat, $\bar{\Theta}$.

We can approximate the variance of an MM Estimator. Consider $\bar{T}$ with $\mu(\bar{T}) = \bar{X}$, applying the probability variance approximation formula

$$V[\mu(\bar{T})] \approx \mu'(\theta)^2 V[\bar{T}]$$

where we also assume that it is not too biased $E[\bar{T}] \approx \theta$. Then for $V[\bar{X}] = \sigma^2(\theta)/n$, there is

$$V[\bar{T}] \approx \frac{\sigma^2(\theta)}{n\mu'(\theta)^2}$$

therefore the standard error is

$$se(\bar{T}) \approx \frac{\sigma(\bar{\theta})}{\sqrt{n}|\mu'(\bar{\theta})|}$$

If finding $\sigma^2(\theta)$ is too difficult, we use s^2 as an estimate.

6.2 Maximum likelihood estimation

Aim to find the parameters that maximize the likelihood of the sample.

The likelihood function for a sample x_i and the parameters θ_i is

$$\begin{aligned} L(\theta_1, \dots, \theta_r) &= P(X_1 = x_1, X_2 = x_2, \dots | \theta_1, \dots, \theta_r) \\ &= \prod_{i=1}^{n} P(X = x_i | \theta_1, \dots, \theta_r) \\ &= \prod_{i=1}^{n} f(x_i | \theta_1, \dots, \theta_r) \end{aligned}$$

where we assume that the sample is fixed and only vary the parameters. It represents the likelihood of the sample given the parameters. For discrete rv, this is the product of pmfs; for continuous rv, this is the product of pdfs.

It is often better to maximize the log-likelihood instead

$$\ln L(\theta)$$

because log is one-to-one and increasing, the maximum parameter for the log-likelihood is the same as the parameter for the maximum likelihood.

The maximum likelihood estimates (MLEs) and maximum likelihood estimators (MLEs) $\hat{\theta}_1, \dots$ are values that maximize $L(\theta_1, \dots)$ (we use the hat notation of MLEs). We often find MLEs by taking logs of the likelihood and equating derivatives to zero. We can also maximize numerically given a sample, but then we lose the close-formed expression.

Note when maximizing

- Set the gradient to zero then solve
- We cannot always use the derivative to find the maximum of the likelihood function
- Careful with the likelihood formula on the valid input parameter ranges (both theoretically and sample dependent)

MLE is invariant under transformations. If we know the MLE $\hat{\theta}$ for θ and $\phi = g(\theta)$, then the MLE for ϕ is $g(\hat{\theta})$. This implies that MLEs are usually biased since expectations are not invariant under non-linear transformations. We can easily prove this if g is invertible (one-to-one mapping between θ and ϕ), otherwise, this still applies but we will not prove it.

Strength of the MLE is its good asymptotic properties. In general, MLEs are: asymptotically unbiased, asymptotically efficient (variance approaches RCLB), and asymptotically normally distributed.

6.3 Method of Quantiles

Similar to the method of moments, it finds the parameter that matches the quantiles to the observed sample quantiles.

If $m(\theta) = c_{0.5}$, define the MQ estimator for θ, $\hat{\Theta}_{MQ}$ as

$$m(\hat{\Theta}_{MQ}) = \hat{M}$$

where $\hat{M}$ is the sample median. We can also use other quantiles for skewed distributions (but we will only use medians and only of one parameter).

7 Important Point Estimators and Sampling Distributions

Sampling distribution depends on population distribution, the statistic, and the sample size.

7.1 Sample Mean

For a random sample X_i, defined as

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

with

$$E[\bar{X}] = E[X]$$
$$V[\bar{X}] = \frac{V[X]}{n}$$

These properties are true irrespective of the underlying distribution of X and the sample size n.

The sample mean is always unbiased for μ. Also, because its variance tends to zero as n increases, we can establish convergence in probability of $\bar{X}$ to μ, so it is consistent for μ. (Use chebyshev's inequality and definition).

If we are sampling from a normal distribution, the sample mean is normal. Otherwise, the CLT will guarantee asymptotic normality of the sample mean (the standardized sample mean, specifically) as n increases. Realistically, we need at least 25 samples for skewed distributions. Symmetrical distributions need less samples to be approximately normal.

A statistical procedure or statistic is robust if it still works if the assumption is changed/incorrect, they are insensitive to deviations on their assumptions. So the sample mean and its sampling distribution is robust by being close to normally distributed even if it is not asymptotic (n is smallish). We can use normal approx if $n > 15$ when no outliers, and $n \geq 40$ even if there are outliers.

Note that while the sample mean may be consistent, the sample sum may not be. We could have the sd of the sample sum increase at a rate of $\sqrt{n}$ but since the sample mean divided it by n, it will collapse and converge.

7.2 Sample Variance

If we know the population mean μ, then an unbiased and consistent estimator for the variance is

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

We can prove the S_1^2 estimator's properties and find its distribution when sampling from a normal distribution using the properties chi-squared distribution.

A special version of the gamma distribution is the χ^2 distribution

$$f(t) = \frac{t^{k/2-1}}{2^{k/2}\Gamma(k/2)}e^{-t/2}, \quad t \geq 0$$

with the parameter $k > 0$ which is the degrees of freedom. The notation is $T \sim \chi_k^2 = \chi^2(k)$. The mean and variance are

$$E[T] = k \quad V[T] = 2k$$

The distribution is always positive and is right skewed.

The chi-squared relation to gamma is

$$\chi_k^2 \sim \Gamma(k/2, 1/2)$$

and that if $Z \sim N(0,1)$, then $Z^2 \sim \Gamma(1/2, 1/2) = \chi_1^2$. Generalizing, we have that for the sum of independent standard normals

$$Z_1^2 + Z_2^2 + \cdots + Z_k^2 \sim \chi_k^2$$

which can be proven using mgf.

The known mean sample variance is a scaled chi-squared if the population is a normal

$$\frac{nS_1^2}{V[X]} \sim \chi^2(n)$$

For unknown mean, we estimate the population mean by the sample mean and the population variance by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

The mean and variance is

$$E[S^2] = \sigma^2 \quad V[S^2] = \frac{\nu_4 - 3V[X]^2}{n} + \frac{2V[X]^2}{n-1}$$

with ν_4 being the fourth central moment (don't need to prove the variance). So the estimator is unbiased and consistent.

Intuitively, because we've estimated the sample mean, creating a constraint on the other n random samples. Our degrees of freedom is reduced by one to $(n-1)$ within the sum, so we divide by $(n-1)$. The variance estimate is simply a sum of squares divided by the degree of freedom.

When sampling from a normal, we can prove that

$$\frac{(n-1)S^2}{V[X]} \sim \chi^2_{n-1}$$

and that $V[S^2] = \frac{2V[X]^2}{n-1}$. This confirms the fact that S^2 has $(n-1)$ degrees of freedom (by the sum of $(n-1)$ independent standard normals).

The sample variance S^2 is only normal when sampling from normal distributions. It is NOT robust like the sample mean in that it will not be approximately normal if the population distribution is not normal. This has implications on the reliabilities of quantiles for CI and h tests.

7.3 Sample Proportions

We model the proportion parameter p in the population for a given characteristic as a Bernoulli distribution of success p. A random sample would consist of a series of $X_i \sim Bi(1,p)$ which is 1 if the ith person in the sample has the characteristic and 0 otherwise.

Suppose that N is the number in the sample of size n with the characteristic, then $N \sim Bi(n,p)$ with the sample proportions being

$$\hat{P} = \frac{N}{n} = \frac{\sum X_i}{n} = \bar{X}$$

so the sample proportion is just the sample mean when the population is bernoulli. Hence, it is unbiased, consistent, and approximately normal.

$$E[\hat{P}] = p$$
$$V[\hat{P}] = \frac{p(1-p)}{n}$$

and for large enough n

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

We also need a continuity correction if using this approximation. This does not work with small np or $n(1-p)$ (less than or equal to 5), for the binomial is very skewed. Under these conditions, we use a poisson approximation.

The sample proportion's unbiasness and consistency is the result underpining the frequentist view that $P(A)$ is the long term relative frequency of A as $n \to \infty$. (Sample proportion as estimator for $P(A)$, so $P(A)$ is the limiting distribution for sample proportions)

7.4 Sample/Empirical cdf, pmf

By definition, the empirical cdf is

$$\hat{F}(x) = \frac{1}{n}\sum I(X_i \leq x)$$

If we choose the event A to be $\{X \leq x\}$, and thus $F(x) = P(A)$, then our empirical cdf is a sample proportion on $Bi(1, P(A))$. This implies that $\hat{F}(x)$ is unbiased and consistent for all $x \in \mathbb{R}$, and is approximately normal.

For discrete X, the empirical pmf is

$$\hat{p}(x) = \frac{1}{n}\sum I(X_i = x)$$

and let event $A = \{X = x\}$, then $p(x) = P(A)$ and the sample pmf is also a sample proportion for all valid x. It is unbiased, consistent, and approximately normal.

This is the theory underlying simulating pmf by generating a table of frequencies for observations.

7.5 Order Statistics

Recall $X_{(k)}$ to be the kth order statistics in the sample of n. $X_{(1)}$ is the sample minimum, and $X_{(n)}$ is the sample maximum.

When sampling from a continuous distribution with cdf $F(x)$ and pdf $f(x) = F'(x)$, the cdf of $X_{(k)}$ is

$$\begin{aligned}
G_k(x) &= P(X_{(k)} \leq x) \\
&= P(\text{at least } k \text{ of } X_i \leq x) \\
&= \sum_{i=k}^{n} P(\text{exactly } k \text{ of } X_i \leq x) \\
&= \sum_{i=k}^{n} \binom{n}{i} F(x)^i (1 - F(x))^{n-i}
\end{aligned}$$

for the event $\{X_{(k)} \leq x\}$ is the same as at least k of X_i is below x. Because each X_i is independent and $P(X_i \leq x) = F(x)$, the probability that exactly k is below x is a binomial distribution.

The pdf is thus

$$\begin{aligned}
g_k(x) &= G_k'(x) \\
&= \sum_{i=k}^{n} \binom{n}{i} iF(x)^{i-1} f(x)(1 - F(x))^{n-i} \\
&+ \sum_{i=k}^{n-1} \binom{n}{i} F(x)^i (n-i)(1 - F(x))^{n-i-1}(-f(x)) \\
&= \sum_{i=k}^{n} \binom{n}{i} iF(x)^{i-1} f(x)(1 - F(x))^{n-i} \\
&- \sum_{j=k+1}^{n} \binom{n}{j-1} F(x)^{j-1}(n-j+1)(1 - F(x))^{n-j} f(x) \\
&= \sum_{i=k}^{n} \binom{n}{i} iF(x)^{i-1} f(x)(1 - F(x))^{n-i} \\
&- \sum_{j=k+1}^{n} j\binom{n}{j} F(x)^{j-1}(1 - F(x))^{n-j} f(x) \\
&= k\binom{n}{k} F(x)^{k-1}(1 - F(x))^{n-k} f(x)
\end{aligned}$$

where we use the chain rule, sub $j = i + 1$, recognize that

$$(n-j+1)\binom{n}{j-1} = (n-j+1)\frac{n!}{(n-j+1)!(j-1)!} = j\frac{n!}{(n-j)!j!} = j\binom{n}{j}$$

and notice that the sums all cancel except when $i = k$.

The special case minimum and maximum pdfs are

$$\begin{aligned} g_1(x) &= n(1 - F(x))^n f(x) \\ g_n(x) &= nF(x)^{n-1} f(x) \end{aligned}$$

which corresponds to the probability expressions of the tails of the distributions

$$\begin{aligned} P(X_{(1)} > x) &= (1 - F(x))^n \\ P(X_{(n)} \leq x) &= F(x)^n \end{aligned}$$

See appendix for an alternative derivation.

We can use this to find the pdf, expectation, and variances of MLE estimators that use order statistics.

Recall that if a continuous rv X has cdf $F(x)$, then $F(X) \sim U(0, 1)$. Because F is non-decreasing, we can transform our order statistics using F on a cdf scale

$$F(X_{(1)}) \leq \cdots \leq F(X_{(n)})$$

and because all of these transformed rv are from a uniform distribution, this is the order statistics from a standard $U(0, 1)$ distribution.

Let $W_k = F(X_{(k)})$ be the kth uniform order statistic with a sample of size n. Its pdf is

$$g_k(w) = k\binom{n}{k} w^{k-1}(1 - w)^{n-k}$$

for $0 \leq w \leq 1$. This is a beta distribution of parameter $W_k \sim Beta(k, n - k + 1)$. Its properties are

$$\begin{aligned} E[W_k] &= \frac{k}{n+1} \\ \text{mode}(W_k) &= \frac{k-1}{n-1} \quad \alpha, \beta > 2 \end{aligned}$$

Because applying the cdf on samples is to convert them into quantiles, this showcases the differences and objectives of type 6 and type 7 quantiles. Under type 6 quantiles $x_{(k)} = c_{k/(n+1)}$ which reflects its mean. Under type 7 quantiles $x_{(k)} = c_{(k-1)/(n-1)}$ which reflects its mode. Each of them assumes that the sample order statistics are exactly at the hypothetical mean or mode quantiles, and linearly interpolate the in-between.

7.6 Sample quantiles

For all definitions of sample quantiles converge when n tends to infinity, we will focus on the natural estimator of the sample quantiles using the empirical cdf.

We have that

$$\begin{aligned} F(c_q) &= q \\ \hat{F}(\hat{c}_q) &= q \\ \hat{F}(\hat{C}_q) &= q \end{aligned}$$

where c_q is the population quantile, $\hat{c}_q$ is an estimate of the quantile with an observed $\hat{F}$ given our sample, while $\hat{C}_q$ is the estimator for the population quantile with a random function $\hat{F}$ that reflects the variations of the sampling.

To approximate the mean and variance of $\hat{C}_q$ assuming a continuous distribution, we have

$$\begin{aligned} F(\hat{C}_q) &\approx F(c_q) + (\hat{C}_q - c_q)f(c_q) \\ \hat{F}(\hat{C}_q) &\approx \hat{F}(c_q) + (\hat{C}_q - c_q)f(c_q) \end{aligned}$$

for $\hat{F} \approx F$ and is the random function of the cdf, and we ignore substituting $f(c_q)$. Since $\hat{F}(\hat{C}_q) \approx q$, there is

$$\hat{C}_q \approx c_q - \frac{\hat{F}(c_q) - q}{f(c_q)}$$

Remembering that $\hat{F}(c_q)$ is a random variable (from the random function at point c_q) with $E[\hat{F}(c_q)] = q$ and $V[\hat{F}(c_q)] = \frac{q(1-q)}{n}$ from sample proportions. Then

$$\begin{aligned} E[\hat{C}_q] &\approx c_q \\ V[\hat{C}_q] &\approx \frac{q(1-q)}{nf(c_q)^2} \end{aligned}$$

These approximations are asymptotically valid and perform reasonably well if the sample size is large. We see that $\hat{C}_q$ is asymptotically unbiased and consistent for c_q. It is also the case that it is asymptotically normal

$$\hat{C}_q \approx N(c_q, \frac{q(1-q)}{nf(c_q)^2}) \quad n \to \infty$$

(we will not prove normality).

When sampling from a normal, the sample median $\hat{C}_{0.5}$ is not as efficient as the sample mean $\bar{X}$ for the parameter μ. This is because $f(m) = \frac{1}{\sigma\sqrt{2\pi}}$ and

$$V[\hat{M}] \approx \frac{1}{4nf(m)^2} \approx \frac{\pi\sigma^2}{2n} = 1.57V[\bar{X}]$$

In general, the efficiency/variance of the sample quantile depends on $f(c_q)$, the intensity of population density around c_q. If the intensity is high enough, the sample median can be more efficient than the sample mean.

8 Asymptotics and Sufficiency

8.1 Asymptotic properties of MLE

We need asymptotic properties because finding exact distributions of MLE is hard.

Fixed likelihood $l(\theta) = l(\theta; x_1, \ldots, x_n)$ is a function of the likelihood when considering an observation of a random sample $x_1, \ldots, x_n$. We aim to maximize the fixed likelihood on θ under a specific random sample to determine the max likelihood estimate. Random likelihood $L(\theta) = L(\theta; X_1, \ldots, X_n)$ is a random function that represents all the different likelihoods functions under all possible random samples (sampling variability). It is a function on θ and the random sample $X_1, \ldots, X_n$.

The assumptions for the asymptotic proof is

- X_i is a random sample
- Sampling from continuous population
- We can find the MLE by first order

$$U(\theta) = \frac{d}{d\theta} \ln L(\theta) = 0$$

- That θ is not a boundary parameter, that the support of the population distribution does not depend on θ.

To reduce ambiguity, let θ_0 be the population parameter, and θ be the parameter of the likelihood function. We denote the population pdf as $f_\theta(x)$ as θ varies, so $X \sim f_{\theta_0}(x)$.

8.1.1 Score, Observed information, and Fisher information function

The score function $U(\theta)$ is the first partial of the log likelihood

$$U(\theta) = \frac{\partial}{\partial\theta} \ln L(\theta)$$

It is a function that returns a random variable.

Some properties of the score function is

$$\begin{aligned} U(\theta) &= \frac{\partial}{\partial\theta} \ln \prod_i f_\theta(X_i) \\ &= \sum \frac{\partial}{\partial\theta} \ln f_\theta(X_i) \\ &= \sum_i U_i(\theta) \end{aligned}$$

with $U_i(\theta)$ as functions of X_i being also iidrv.

Also define the observed information function $V(\theta)$ as the negative second partial of the log likelihood

$$V(\theta) = -\frac{\partial^2}{\partial\theta^2} \ln L(\theta)$$

It is a function that returns a random variable.

Some properties of the observed information function

$$\begin{aligned} V(\theta) &= -\frac{\partial^2}{\partial\theta^2} \ln \prod_i f_\theta(X_i) \\ &= -\sum_u \frac{\partial^2}{\partial\theta^2} \ln f_\theta(X_i) \\ &= \sum_i V_i(\theta) \end{aligned}$$

where $V_i(\theta)$ is a function on X_i which are also iidrvs.

Define the fisher information (or expected information function) as

$$I_n(\theta) = E[V(\theta)]$$

which is also called the expected information function. It is a function that returns a number. This information function plays an important role in the asymptotical properties of the MLE.

Properties of the fisher information is

$$I_n(\theta) = E[V(\theta)] = nE(V_i(\theta)) = nI_1(\theta)$$

where $I_1(\theta)$ is the fisher information on a sample size of 1. This intuitively means that the information in a random sample of size n is just n times the information in a unit sample, for each observation carries the same amount of information. We often work with $I_1(\theta)$ instead of the full fisher information for that involves a whole likelihood compared to the one term $E[-\frac{\partial^2}{\partial\theta^2} \ln f_\theta(X_i)]$ that the unit fisher information has.

8.1.2 Asymptotic properties of MLE

To show the asymptotic properties of the MLE, we assume that the MLE $\hat{\Theta}_n$ is close to θ_0, so a taylor approximation is

$$\begin{aligned} U(\hat{\Theta}_n) &\approx U(\theta_0) + U'(\theta_0)(\hat{\Theta}_n - \theta_0) \\ \hat{\Theta}_n - \theta_0 &\approx \frac{U(\theta_0)}{-U'(\theta_0)} = \frac{U(\theta_0)}{V(\theta_0)} = \frac{U(\theta_0)/n}{V(\theta_0)/n} \end{aligned}$$

because by definition of the MLE, $U(\hat{\Theta}_n) = 0$. And due to both the score function and observed information function being sums of iidrvs, both the numerator and denominator are sample means from $U_i(\theta_0)$ and $V_i(\theta_0)$.

First consider the sample mean $U(\theta_0)/n$. By the CLT, it has

$$U(\theta_0)/n \approx N\left(E[U_i(\theta_0)], \frac{V[U_i(\theta_0)]}{n}\right)$$

And we can compute the mean and variance separately. First note that

$$\begin{aligned} E[U_i(\theta_0)] &= \int_{-\infty}^{\infty} [\frac{\partial}{\partial\theta} \ln f_{\theta_0}(x)] f_{\theta_0}(x)\, dx \\ &= \int \frac{\partial}{\partial\theta} f_{\theta_0}(x)\, dx \\ &= \frac{\partial}{\partial\theta} \int f_{\theta_0}(x)\, dx \\ &= 0 \end{aligned}$$

remembering that $\frac{\partial}{\partial\theta} f_{\theta_0}(x)$ implies that the partial derivative is taken on the $f_\theta(x)$ variant then the resulting θ is replaced with θ_0.

Secondly, consider taking the derivatives of the above integral

$$\begin{aligned} \int_{-\infty}^{\infty} [\frac{\partial}{\partial\theta} \ln f_{\theta_0}(x)] f_{\theta_0}(x)\, dx &= 0 \\ \int \left[[\frac{\partial^2}{\partial\theta^2} \ln f_{\theta_0}(x)] f_{\theta_0}(x) + \frac{\partial}{\partial\theta} \ln f_{\theta_0}(x) \frac{\partial}{\partial\theta} f_{\theta_0}(x) \right] dx &= 0 \\ \int \left[\frac{\partial}{\partial\theta} \ln f_{\theta_0}(x) \right]^2 f_{\theta_0}(x)\, dx &= -\int [\frac{\partial^2}{\partial\theta^2} \ln f_{\theta_0}(x)] f_{\theta_0}(x)\, dx \end{aligned}$$

because $\frac{\partial}{\partial\theta} f_{\theta_0}(x) = [\frac{\partial}{\partial\theta} \ln f_{\theta_0}(x)] f_{\theta_0}(x)$. The last expression is $E[U_i(\theta_0)^2]$ which is also its variance for it has zero mean, thus

$$V[U_i(\theta_0)] = E[U_i(\theta_0)^2] = E[V_i(\theta_0)] = I_1(\theta_0)$$

which is the unit fisher information when sampling from $f_{\theta_0}(x)$.

Altogether, the sample mean distribution is

$$U(\theta_0)/n \approx N(0, I_1(\theta)/n)$$

Now consider $V(\theta_0)/n$. Using the LLN (or equivalently the consistency of the sample mean)

$$V(\theta_0)/n \to E[V_i(\theta_0)] = I_1(\theta_0)$$

and we can replace the denominator with the limiting result.

Note that for the approximation, we use the CLT for the numerator for it has a zero mean, and use the LLT and consistency of the sample mean for the denominator.

Therefore the asymptotic distribution for the MLE is

$$\begin{aligned}\hat{\Theta}_n - \theta_0 &\approx \frac{1}{I_1(\theta_0)} N(0, \frac{I_1(\theta_0)}{n}) \\ &\approx N(0, \frac{1}{nI_1(\theta_0)}) \\ \hat{\Theta}_n &\approx N(\theta_0, \frac{1}{I_n(\theta_0)}) \quad n \to \infty\end{aligned}$$

which shows that MLE is asymptotically unbiased and normal. Because the RCLB is also the inverse of the fisher information, the MLE is asymptotically efficient as well.

The asymptotic results of the MLE also holds for discrete distributions.

Additionally, notice the two methods of calculating the fisher information

$$\begin{aligned}I_n(\theta) &= -nE[\frac{\partial^2}{\partial\theta^2} \ln f(X;\theta)] \\ &= nE[U_i(\theta)^2] = nE[(\frac{\partial}{\partial\theta} \ln f(X;\theta))^2]\end{aligned}$$

each can be useful in the right context.

8.2 Rao Cramer Lower Bound

We will prove the RCLB for the variance of the estimator T_n where $E[T_n] = \psi(\theta)$, which allows a bias term. We will ignore any regularity conditions.

Consider the covariance of T_n with $U(\theta)$

$$\begin{aligned}
Cov(T_n, U) &= \sum_i Cov(T_n, U_i) \\
&= nCov(T_n, U_i) \\
&= n[E[T_nU_i] - E[T_n]E[U_i]] \\
&= nE[T_nU_i]
\end{aligned}$$

for $E[U_i] = 0$.

To compute $E[T_nU_i]$, consider the product rule

$$\frac{\partial}{\partial\theta}\prod_i f(x_i;\theta) = \sum_i \left[\frac{\partial}{\partial\theta} f(x_i;\theta) \prod_{j\neq i} f(x_j;\theta)\right] = \left[\sum_i \frac{1}{f(x_i;\theta)} \frac{\partial}{\partial\theta} f(x_i;\theta)\right] \prod_i f(x_i;\theta)$$

then

$$\begin{aligned}
E[T_nU_i] &= E[T_n \frac{1}{f(X_1;\theta)} \frac{\partial}{\partial\theta} f(X_1;\theta)] \\
&= \frac{1}{n}\sum_i E[T_n \frac{1}{f(X_i;\theta)} \frac{\partial}{\partial\theta} f(X_i;\theta)] \\
&= \frac{1}{n}\int \cdots \int t_n \left[\sum_i \frac{1}{f(x_i;\theta)} \frac{\partial}{\partial\theta} f(x_i;\theta)\right] \prod_i f(x_i;\theta)\, dx_1 dx_2 \ldots \\
&= \frac{1}{n}\int \cdots \int t_n \frac{\partial}{\partial\theta} \prod_i f(x_i;\theta)\, dx_1 dx_2 \ldots \\
&= \frac{1}{n}\frac{\partial}{\partial\theta} E[T_n] \\
&= \frac{1}{n}\psi'(\theta) \\
E[T_nU] &= \psi'(\theta)
\end{aligned}$$

where we used multidim expectation definition, the joint pdf as a product of individual pdf due to independent samples, the product rule, and the biased mean of T_n.

Now consider the correlation coefficient

$$\begin{aligned}|Cor(T_n, U)| &= |\frac{Cov(T_n, U)}{\sqrt{V[T_n]V[U]}}| \\ &= |\frac{\psi'(\theta)}{\sqrt{V[T_n]I_n(\theta)}}| \leq 1 \\ \frac{(\psi'(\theta))^2}{V[T_n]I_n(\theta)} &\leq 1 \\ V[T_n] &\geq \frac{(\psi'(\theta))^2}{I_n(\theta)}\end{aligned}$$

using the covariance result and $V[U] = V[\sum_i U_i] = nV[U_i] = nI_1 = I_n$. This is the RCLB for the lower bound variance of an estimator T_n with mean $E[T_n] = \psi(\theta)$. For an unbiased estimator where $\psi(\theta) = \theta$, the minimum variance is

$$V[T_n] \geq \frac{1}{I_n(\theta)}$$

For a biased estimator T_n with $E[T_n] = \theta + b(\theta)$, the RCLB is

$$V[T_n] \geq \frac{(1 + b'(\theta))^2}{I_n(\theta)}$$

Define the precision of a random variable as the reciprocal of its variance, the smaller the variance, the higher the precision. The fisher information determines the asymptotic precision of the MLE. Intuitively, the larger the sample size, the larger the fisher information, and the higher the asymptotic precision of the MLE.

We can rewrite the efficiency of an unbiased estimator as

$$\text{eff}(T_n) = \frac{1/I_n(\theta)}{V[T_n]} = \frac{1}{I_n(\theta)V[T_n]}$$

The MSE of a biased estimator $E[T_n] = \theta + b(\theta)$ is

$$MSE(T_n) \geq \frac{(1 + b'(\theta))^2}{I_n(\theta)} + b(\theta)^2$$

which can have a smaller bound than the MSE of an unbiased estimator if $(1 + b'(\theta))^2 + b(\theta)^2 I_n(\theta) < 1$. So it is possible for biased estimator (such as S_1^2) to have a smaller MSE than unbiased ones.

8.3 Sufficiency

Intuitively, a sufficient statistic is one where the conditional probabilities of the sample on the sufficient statistic does not depend on the population parameter, so the statistic is extracting all relevant information. We want such a statistic that extracts all information without taking on board excess variability. (Alternatively, the statistics captures all information about the parameter and ignores the information that the parameter doesn't affect)

The definition is: a statistic $T = \psi(X_1, \ldots, X_n)$ is sufficient for an underlying parameter θ if the conditional probability distribution of the random sample X_i given the statistic $f(x_1, \ldots, x_n|t)$, does not depend on the parameter θ.

If the conditional pd of the random sample conditioned on multiple statistic does not depend on θ (sometimes we need to condition on multiple statistics), then we say that the statistics are jointly sufficient for θ.

We can generate a complete sample that is stochastically indistinguishable from a real sample from a sufficient statistic.

The factorization theorem states that: let X_i have joint pdf $f(x_1, \ldots, x_n|\theta)$ (this is just the pdf with parameter θ), a statistic $T = \psi(x_1, \ldots, x_n)$ is sufficient for θ if and only if

$$f(x_1, \ldots, x_n|\theta) = g(\psi(x_1, \ldots, x_n)|\theta)h(x_1, \ldots, x_n)$$

where g only depends on the statistic (and implicitly on x_i through ψ) and θ, and h doesn't depend on θ. These restrictions apply on the function domain as well. (This Fisher's factorization theorem is not proved).

Moreover, a statistic $T_1 = \psi(X_1, \ldots, X_n)$ is sufficient for the parameter θ iff for all other statistic T_2, the conditional distribution

$$f_{T_2|T_1}(t_2|t_1)$$

is independent of θ (in both formula and domain). We don't prove it.

We can find a sufficient statistic by calculating the likelihood function (just the pdf of the random sample) and trying to factorize it. Or we can use intuition.

If a statistic is sufficient for a parameter θ, then every single-valued function $\phi(T_1)$, which doesn't have θ and has a single valued inverse, is also a sufficient statistic for θ. We can prove this by the factorization theorem and subbing $\theta = \phi^{-1}(\phi(\theta))$ into g.

Consider the exponential family of distributions, which have pdfs in the form

$$f(x|\theta) = \exp(K(x)p(\theta) + S(x) + q(\theta))$$

Suppose a random sample $X_1, \ldots, X_n$ from this family, then the statistic $\sum_i K(X_i)$ is sufficient for θ. This is because the joint pdf is

$$\begin{aligned} f(x_1, \ldots, x_n|\theta) &= \exp(p(\theta)\sum_i K(x_i) + \sum_i S(x_i) + nq(\theta)) \\ &= \exp(p(\theta)\sum_i K(x_i) + nq(\theta)) \exp(\sum_i S(x_i)) \end{aligned}$$

which by the factorization theorem, shows sufficiency.

The MLE for θ will be a function of sufficient statistics on θ, if there are any. This is because the likelihood function is of the form

$$L(\theta) = f(x_1, \ldots, x_n|\theta) = g(\phi(x_1, \ldots, x_n)|\theta)h(x_1, \ldots, x_n)$$

by the factorization theorem on a sufficient statistic $T_n = \phi$. As the MLE is found by maximizing the likelihood using θ, and thus maximizing $g(\phi|\theta)$ which is a function of θ and the statistic, the MLE must be a function on the sufficient statistic.

Sufficient statistics are important because

- After the sufficient statistics, there are no additional information on the parameter from the sample
- Samples with the same sufficient statistic as each other will yield the same estimates
- Most optimal estimators and tests are based on sufficient statistics
- It's easy to find sufficient statistics in special cases (exponential family)

But sufficiency requires knowing the population distribution, so it is mostly theoretical. In practice, we should look at all aspects of the data and perform sanity checks on our assumed sufficient statistics.

9 Interval Estimation

9.1 Quantify estimate uncertainty

Point estimates are often insufficient to conclusively answer real questions of interest. The lurking questions under point estimates are

- How confident are you in the estimate?
- How accurate is it?

We aim to quantify and communicate the uncertainties of our estimate.

Some populations are very variable, some are very stable. Initially, we should investigate such population variabilities for it may shape the analysis and study design. We can use: graphical summaries of all variables, numeric measures of location/scale/shape, association/correlation/relationships for bivariate populations.

Estimators contains both a point estimate and its variability. It is usually best to report the point estimate and its standard deviation for they have the same units. We often need the population parameter to get the sd of the sampling distribution, so we estimate the parameter with the point estimate which nets the standard error of the estimator.

More specifically, the standard deviation of an estimator is

$$sd(\hat{\Theta}) = \sqrt{V[\hat{\Theta}]}$$

but sd often depends on θ which is the parameter for the estimator. So we sub the parameter with the point estimate and approximate the sd, creating the standard error of the estimator. We can represent the standard error as se or $\hat{sd}$. The standard error of an estimator is its estimated standard deviation using its point estimate.

Notations

- Parameter θ
- Estimator $\hat{\Theta}$
- Estimate $\hat{\theta}$
- Sd of estimator $sd(\hat{\Theta})$
- Se of estimator $se(\hat{\Theta})$

note that in practice some might write $se(\hat{\theta})$ to denote the standard error as part of the observed estimate, this standard error of the estimate is incorrect.

We can report the standard error by

- $\hat{\theta} \quad (se(\hat{\Theta}))$
- $\hat{\theta} \pm se(\hat{\Theta})$
- $\hat{\theta} \pm 2se(\hat{\Theta})$

9.2 Random intervals

A random interval $[L, U]$ or (L, U) is an interval on the line with random variables as starting and ending points. We have the restriction that $L < U$.

We do not rule out (we allow) that one of the random endpoints being fixed (degenerate) or $-\infty$ or $+\infty$.

Random intervals can have

- Random location only and fixed length
- Random location and length
- Fixed location and random length

When random intervals fall (observed) on the line, they may or may not hit something of interest. This is their hitting probability.

9.3 Confidence Intervals

A confidence interval for a parameter θ is a random interval with a certain hitting probability for θ (probability of landing on θ).

The width of the CI can be adjusted using quantiles of the pivotal quantity to change the hitting probability of θ to a desired level.

The hitting probability of a CI is called its confidence level.

The CI simultaneously gives us an idea of where the parameter is and our uncertainty concerning its value.

The interval estimator (a confidence interval) is a random interval that can be calculated from the sample. Its target parameter is fixed and unknown.

Before the sample is taken, the probability that the random interval contains θ is the confidence level. After the sample is taken, we've realized and interval that either contains θ or not.

Most of the time, the CI is of the form

$$CI = \text{est} \pm \text{error}$$

If we take m samples of size n and calculate a 95% CI for the parameter for each sample, the distribution of CIs that hits/contains the parameter is a binomial with size m and success probability 95%.

9.4 Example Mean Normal with known variance

Here's the CI derivation for the normal population $X \sim N(\mu, \sigma^2)$. Assume that σ^2 is known. Let $1 - \alpha$ be the confidence level.

Using the sample mean, the sampling distribution is $\bar{X} \sim N(\mu, \sigma^2/n)$. Let $\Phi^{-1}(1-\alpha/2) = c$ be the selected quantile given the CL α, we have

$$P\left(-c < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < c\right) = 1-\alpha$$
$$P\left(\mu - c\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + c\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$
$$P\left(\bar{X} - c\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + c\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

with the first statement by definition, the second from rearranging creating the probability interval (the interval with probability $1-\alpha$ that our estimator lies in), and the third also from rearranging creating the CI with confidence level $1-\alpha$.

The resulting random interval (CI)

$$(\bar{X} - c\frac{\sigma}{\sqrt{n}}, \bar{X} + c\frac{\sigma}{\sqrt{n}})$$

contains μ with a probability $1-\alpha$. It is called a $100(1-\alpha)\%$ confidence interval for the population mean μ.

We can use our sample mean estimate $\bar{x}$ to calculate the observed $100(1-\alpha)\%$ confidence interval by replacing $\bar{X}$ with it.

Notice that we've chosen our quantiles evenly $(-c, c)$. The quantiles are selected by ensuring that the probability our pivotal quantity is within the quantiles is the confidence level. They split the miss probability evenly above and below, and importantly, if we had an asymmetric quantile, the resulting CI width would be larger — this is because the tails are thinner than the middle, so if we shift one side inwards, the other side will have to shift outwards relatively more thus increasing the overall width.

9.5 Example Mean Non-normal known variance

If X is not normal, we can resort to the CLT of the sample mean if n is large enough

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

Then the same normal CI formula holds.

9.6 General CI derivation

Point estimator is a statistic that is a good estimator for a parameter, it is a single number with no expression of uncertainties.

Interval estimator is a pair of statistics defining a random interval that has a hitting probability for a parameter. It provides a range of plausible parameter values and expresses uncertainty.

A probability interval is the (general) probability that a random variable lies between an interval. A probability interval is the probability that a statistic lies within an interval. A confidence interval is the probability that a parameter lies within a random interval.

A quantity $Q(X_1, \dots, X_n; \theta)$ of a random sample on a parameter is a random variable that is

- A function of the unknown parameter θ and no other unknown parameters.
- A function on the random sample and other known parameters
- Has a known distribution that does not depend on the random sample X_i nor θ.

It is a random variable but not a statistics (because statistics can't depend on θ).

To find the $\alpha\%$ CI for the parameter θ, first start with a pivotal quantity (or just pivot) for θ called $Q(X_1, \dots, X_n; \theta)$. Then write the central probability interval

$$P(c_{\alpha/2} < Q(X_1, \dots, X_n; \theta) < c_{1-\alpha/2}) = 1 - \alpha$$

where the cs are quantiles of the quantity distribution, which does not depend on X_i nor θ. We should then decompose the quantity and rewrite the probability interval into a probability interval

$$P(a(\theta) < T < b(\theta)) = 1 - \alpha$$

where T is a statistic depending only on the random sample, and $a(\theta), b(\theta)$ are endpoints that only depend on the parameter. Lastly, invert a and b to net

$$P(b^{-1}(T) < \theta < a^{-1}(T)) = 1 - \alpha$$

which is a confidence interval $(b^{-1}(T), a^{-1}(T))$ for the parameter θ with confidence level $1 - \alpha$. To create an observed CI, substitute the observed statistic into the random interval.

A statistic parameter diagram visualizes the derivation of the CI. It has θ on the x-axis and the observed statistic t on the y-axis. The two functions a and b are graphed on the axes, and we can see the translation of the plausible range of T values (probability interval) to the plausible range of θ values (confidence interval). In detail, the CI for a given sample mean contains population means where the probability interval contains the said sample mean, which is exactly from $b^{-1}(T)$ to $a^{-1}(T)$.

We can usually convert a CI for a parameter to another by applying the same transformation on both the endpoints, we might need to flip the order of the endpoints if the transformation changes the inequality.

CI are drawn with error bars graphically.

9.7 CI Interpretation

The width of the CI depends on

- Inherit variation of the data
- Choice of estimator and pivot
- Confidence level
- Sample size

Narrower widths usually indicates stronger evidence about the plausible true values of the estimated parameter

- If CI is wide, cannot conclude much except we have insufficient data
- If moderate, conclusion will depend on location of CI
- If narrow, more confidence about true value, can be conclusive

What constitutes as wide or narrow, and how conclusive the CI is depends on the context of the study question.

9.8 Important distributions

9.8.1 Chi-squared

We've seen $\chi^2(k), k > 0$ distribution with $E[T] = k$ and $V[T] = 2k$. It is positive and right-skewed. It is the sum of squared iid standard normals. It's important for the sample variance of a random sample from a normal distribution has a scaled Chi-squared distribution

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

9.8.2 T-distribution

Student's t-distribution t_k has a single parameter $k > 0$ which is the degrees of freedom. For $T \sim t_k$, its pdf is

$$f(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}\left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}, \quad -\infty < t < \infty$$

with the moments

$$E[T] = 0, \quad k > 1$$
$$V[T] = \frac{k}{k-2}, \quad k > 2$$

The t-distribution is similar to a normal but with heavier tails. It is asymptotically normal

$$t_k \to N(0,1) \quad \text{as } k \to \infty$$

If $Z \sim N(0,1)$ and $U \sim \chi^2(r)$, and they are independent, the random variable

$$T = \frac{Z}{\sqrt{U/r}} \sim t_r$$

has a t-distribution. It originates from the sampling distribution of standardized sample means from a normal population $X \sim N(\mu, \sigma^2)$ with unknown variance, namely

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = t_{n-1}$$

as $\bar{X}$ and S^2 are independent for a random sample of a normal.

9.8.3 F-distribution

The fisher distribution $F_{m,n}$ has parameters $m, n > 0$ which are the degrees of freedom. We say that $W \sim F_{m,n}$ if W is f-distributed.

If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ are independent, we have that

$$\frac{U/m}{V/n} \sim F_{m,n}$$

so the scaled ratio of independent chi-squares has a f-distribution. This often arises when taking the ratio between sample variances.

The random variable $1/W$ is also a f-distribution with the degrees of freedoms flipped.

9.9 CIs in common scenarios

We mainly focus on inference of means and variances in normal distributions and proportions.

We assume a confidence level of 95%.

9.9.1 Normal, Single mean, known sd

Given the random sample $X_i \sim N(\mu, \sigma^2)$ and we know σ. The pivot is

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Hence we just need the fixed quantiles a and b such

$$P(a < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < b) = 0.95$$

which are $a = \Phi^{-1}(0.025)$ and $b = \Phi^{-1}(0.975)$. Then rearrange for μ for the CI.

9.9.2 Normal, Single mean, unknown sd

Given the random sample $X_i \sim N(\mu, \sigma^2)$ and we don't know σ.

A pivot for μ is

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

is the standardized sample mean using the sample standard deviation.

Hence, for c chosen as the 0.975th quantile of t_{n-1}, the derivation for the CI is

$$P(-c < \frac{\bar{X} - \mu}{S/\sqrt{n}} < c) = 0.95$$

$$P(\bar{X} - c\frac{S}{\sqrt{n}} < \mu < \bar{X} + c\frac{S}{\sqrt{n}}) = 0.95$$

This nets the CI

$$(\bar{X} - c\frac{S}{\sqrt{n}}, \bar{X} + c\frac{S}{\sqrt{n}})$$

The CIs based on t-distributions has the form

$$\text{estimate} \pm c \times \text{standard error}$$

where c is the quantile determined by the sample size and confidence level.

This is appropriate only if the sample is from a normal distribution, which can be checked using a QQ plot. Otherwise

- If n is large, construct approximate CI using normal distribution (not t-distribution). This is usually fine if the dist is continuous, symmetric and unimodal
- If n is small, use distribution-free methods

9.9.3 Normal, Two means, known sd

Random samples from two populations, $X_1, \ldots, X_n \sim N(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma_Y^2)$. Assuming the samples are independent within sample and between sample, and σ_X^2 and σ_Y^2 are known.

We want a CI on $d = \mu_X - \mu_Y$ which is how much their means differ. The pivot is

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

due to $\bar{X}$ and $\bar{Y}$ being independent normal.

Then let c be the normal 0.975 quantile

$$P(-c < \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} < c) = 0.95$$

$$P(\bar{X} - \bar{Y} - c\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} < \mu_X - \mu_Y < \bar{X} - \bar{Y} + c\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}) = 0.95$$

which nets the CI for $\mu_X - \mu_Y$

$$\bar{X} - \bar{Y} \pm c\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

9.9.4 Normal, two means, unknown sd, large samples

If n and m are large, estimate the sd with the sample standard deviations (for they are good estimates if the sample size is large). Then our pivot (using the estimated sd) is approximately normal with $N(0,1)$, and we net an approximate CI for the mean difference.

9.9.5 Normal, two means, unknown sd, small sample, equal variance

If the sample sizes are small, but we assume a common variance $\sigma^2 = \sigma_X^2 = \sigma_Y^2$, we can find a pivot. First note that

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma^2/n + \sigma^2/m}} \sim N(0,1)$$

And because the samples are independent and the sum of chi-squared rv is also chi-squared

$$U = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2(n+m-2)$$

Moreover, as the sample means and sample variances are independent, Z and U are also indep, we can construct a pivot T

$$\begin{aligned} T &= \frac{Z}{\sqrt{U/(n+m-2)}} \sim t_{n+m-2} \\ &= \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_P\sqrt{1/n + 1/m}} \\ S_P &= \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}} \end{aligned}$$

where we can simplify T to remove the σ. Intuitively, S_P is the pooled estimate of the common variance. We can then find the CI by using the 0.975th quantile c for t_{n+m-2} and solving

$$P(-c < T < c) = 0.95$$

to net the CI for $\mu_X - \mu_Y$

$$\bar{X} - \bar{Y} \pm S_P\sqrt{1/n + 1/m}$$

9.9.6 Normal, two means, unknown sd, small samples, unequal variances

If the sample sizes are small and the variances are not equal. We use Welch's approximation of the pivot

$$\begin{aligned} W &= \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}} \approx t_r \\ r &= \frac{(S_X^2/n + S_Y^2/k)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}} \end{aligned}$$

where W is approximately t-distributed with degrees of freedom r (W is the Welsh statistic). The approximate CI is therefore

$$\bar{X} - \bar{Y} \pm c\sqrt{S_X^2/n + S_Y^2/m}$$

where c is the quantile from t_r.

This choice of pivot is often the default for constructing CI of differences, because there are no assumptions of large sample sizes, known sd, and equal variances. The result between Welsh's approximation and pooled variance CI are basically the same.

Usually, the pooled variance CI will net a tighter CI than the Welsh approximate CI. But there is likely not a big difference.

9.9.7 Normal, two means, paired samples

The measures are in the form of independent pairs of normal observations

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

and we want a CI for $\mu_D = \mu_X - \mu_Y$. We cannot use the non-paired CI because the two random variables within each pair is not independent. We can exploit the relationship within pairs to improve and simplify our estimates.

Let $D_i = X_i - Y_i$, then $D_i \sim N(\mu_D, \sigma_D^2)$ and they are independent normals. We can then apply the method for single means CI on D_i.

If we know the variance σ_d^2, then the CI for μ_D is

$$\bar{D} \pm c\frac{\sigma_D}{\sqrt{n}}$$

where c is the quantile from a normal distribution. If we don't know the variance, the CI is

$$\bar{D} \pm c\frac{S_D}{\sqrt{n}}$$

where c is the quantile from a t_{n-1} distribution.

The key skill here is recognizing that the dataset are paired.

9.9.8 Normal, single variance

For a random sample of normal $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, we wish to infer σ.

A pivot for σ is

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Then given the 0.025th quantile a and 0.975th quantile b from the $\chi^2(n-1)$ distribution, we have

$$P(a < \frac{(n-1)S^2}{\sigma^2} < b) = 0.95$$
$$P(\frac{(n-1)S^2}{b} < \sigma^2 < \frac{(n-1)S^2}{a}) = 0.95$$

where we've re-arranged the probability interval in terms of σ^2 to get the CI for the variance

$$\big((n-1)S^2/b, (n-1)S^2/a\big)$$

Because the variance contains squaring, we would need a large sample size to get a precise variance CI.

9.9.9 Normal, two variances

We wish to compare the variances between two normally distributed population. We want a CI for σ_X^2/σ_Y^2.

The random samples are $X_1, \ldots, X_n \sim N(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma_Y^2)$.

Consider the pivot

$$\frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} = \frac{\left((m-1)S_Y^2/\sigma_Y^2\right)/(m-1)}{\left((n-1)S_X^2/\sigma_X^2\right)/(n-1)} \sim F_{m-1,n-1}$$

which is a ratio of indep Chi-squared random variables (for sample variances are indep) divided by their degrees of freedom, hence a f-distribution.

Let c and d be the 0.025th and 0.975th quantile of $F_{m-1,n-1}$, then

$$P(c < \frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} < d) = 0.95$$
$$P(c\frac{S_X^2}{S_Y^2} < \frac{\sigma_X^2}{\sigma_Y^2} < d\frac{S_X^2}{S_Y^2}) = 0.95$$

leading to the CI for σ_X^2/σ_Y^2 as

$$\left(c\frac{S_X^2}{S_Y^2}, d\frac{S_X^2}{S_Y^2}\right)$$

Be careful about the flipped order of degrees of freedom in the f-distribution.

9.9.10 Single proportions

Suppose we observed the Bernoulli trials $X_1, \ldots, X_n \sim B(p)$ with unknown probability p of success, and we want a CI for p.

The CLT for sample means shows that for large n

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \approx N(0,1)$$

By estimating the denominator p with $\hat{p}$ (Wald approximation), and rearranging

$$P(-c < \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} < c) = 0.95$$
$$P(\hat{p} - c\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + c\sqrt{\hat{p}(1-\hat{p})/n}) = 0.95$$

netting the approximate CI for p as

$$\hat{p} \pm c\sqrt{\hat{p}(1-\hat{p})/n}$$

Additionally, we can solve for p in the pivot which nets the quadratic approximation for the proportions CI.

9.9.11 Two proportions

Consider two samples of Bernoulli trials $X \sim B(n_1, p_1)$ and $Y \sim B(n_2, p_2)$. We want a CI for the difference in proportions.

When sample size is large, the sample proportions are approximately normal and the pivot is also approximately standard normal

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \approx N(0,1)$$

rearranging and estimating p_1 with $\hat{p}_1$ (and p_2) on the denominator using Wald's method, we have the approximate CI for $p_1 - p_2$

$$\hat{p}_1 - \hat{p}_2 \pm c\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$$

where c is the quantile from the standard normal

In general, a difference CI is better in inferring if a mean/proportion has changed compared to point estimates. This is because CI accounts for the sampling variability in both samples while point estimates don't.

9.10 Less common CIs

9.10.1 One-Sided CI

One-sided confidence intervals contains just a random variable lower or upper bound, (L, ∞) or $(-\infty, U)$.

To create a one-sided CI, start with finding a pivotal quantity, and write a one-sided probability interval about the pivot. Then by rearranging, we will receive the one-sided CI.

To get a upper bound CI with confidence level $1-\alpha$ with the pivot $T = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, use

$$P(T > c) = P(\bar{X} - c\frac{\sigma}{\sqrt{n}} > \mu) = 1 - \alpha$$

where $c = \Phi^{-1}(\alpha)$ is chosen so that this probability holds. For a lower bound CI, use

$$P(T < c) = P(\bar{X} - c\frac{\sigma}{\sqrt{n}} < \mu) = 1 - \alpha$$

where $c = \Phi^{-1}(1 - \alpha)$.

Notice that unlike 2-sided CI, we don't need to separate the α into two parts. For other scenarios (variances, unknown σ, or differences), we use the same approach but changing c such that it is a quantile of $(1 - \alpha)$ or α from the new quantity distribution.

In R, we can use the `alternative = less` for an upper bound CI and `alternative = greater` for a lower bound CI.

9.10.2 Exact CIs for discrete statistics

When we are working with continuous pivots and statistics, we can write down an exact central probability interval for the statistic as it is continuous.

If the distribution or the statistic is discrete, and the sample size is not large enough to justify an CLT, we would want an exact form of CI.

The limitation is that in the probability interval

$$P(a(\theta) \leq T \leq b(\theta)) = 0.95$$

both a and b can only take discrete values. So we may have to settle with an "at least" probability where

$$P(a \leq T \leq b) \geq 0.95$$

and for symmetry, we choose $a(\theta) = \arg\max_t P(T \leq t) \leq 0.025$ which is the largest value it could be to contain at most 0.025, and $b(\theta) = \arg\min_t P(T \geq t) \leq 0.025$ which is the smallest value it could be to contain at most 0.025.

However, when drawn on a statistic parameter diagram, we see that both a and b are step functions when the statistic is discrete, and it may be difficult to find inverses.

Clopper Person's exact CI for discrete binomial $B(n, \theta)$ selects the interval (θ_L, θ_R) not by finding the largest and smallest quantiles, but to set

$$P(X \geq x | \theta = \theta_L) = 0.025$$
$$P(X \leq x | \theta = \theta_R) = 0.025$$

where x is our sample successes and X is the population rv. The intuition is that the right probability decreases as θ decreases, so we use that as the lower bound θ at half the one minus confidence level, and vice versa.

These tail probabilities will vary continuously as θ varies, so we can have an exact CI for θ (proof required). We will only explore this through simulations.

9.10.3 Approximate CI from MLEs

If the parameter has an MLE point estimator, we can use the asymptotic properties (particularly normality) of the MLE to find an approximate CI.

Let the MLE for θ be $\hat{\Theta}$. Notice that

$$\hat{\Theta} \approx N(\theta, \frac{1}{I_n(\theta)})$$

if n is large. We can approximate $I_n(\theta)$ which is the expected information function by using the observed likelihood function $l(\theta|x_i)$. This leads to just the observed information function $V(\theta)$, which we can estimate by using the estimated $\hat{\theta}$. Notice that we've used the sample x_i and sample MLE $\hat{\theta}$ to estimate the fisher information, and thus creating the asymptotic (and thus approximate) standard error of the MLE

$$se(\hat{\Theta}) \approx \frac{1}{\sqrt{V(\hat{\theta})}}$$

Hence, if the sample size is large, using the approximate CI of a normal distribution with estimated variances/se, we can construct the approximate $(1-\alpha)$-CI on a MLE for θ by

$$\hat{\theta} \pm \frac{c}{\sqrt{V(\hat{\theta})}} \qquad \text{or} \qquad \hat{\theta} \pm c \times se(\hat{\theta})$$

where $c = \Phi^{-1}(1-\alpha/2)$.

To summarize all approximate CI based on normality approximation. We first approximate normality using CLT or MLE asymptotics, then substitute estimates into the estimator/statistic standard deviations to create the standard error, finally create the approximate CI by treating the se as the sd.

9.11 Coverage

Each CI has a badged CL, which is the target hitting probability. But sometimes we approximate the CI (Wald interval for a proportion approximating the Binomial with a normal and assume that the sample proportion variance is its estimates), and when the approximation breaks down, the actual CL may not be the badged CL.

We use the term coverage to mean the actual hitting probabilities of the CI and the displayed (targeted) confidence level as the nominal coverage.

For the same sample size and for different parameters, our inexact methods will produce different estimators/estimates for those CI with different actual hitting probability from the nominal coverage.

As the actual hitting probability achieved will vary across parameter values, the coverage is a function of the parameter. Generally we use simulations to find and plot the coverage function.

9.12 CI communication

For our Confidence Level

- If very high, more likely to capture parameter, but can be impractically wide and would not be a useful guide to show the plausible values based on the data.
- If very low, more useful roughly in the sense of being more selective about the plausible range, at the expense of a loss of confidence.

95% is the most comment CL, followed by 90%. 50% can be useful due to its easy interpretation, and can also be used if plotting many overlapping CIs to reduce clutter.

The choice of CL varies by application (difference scenarios), and we can use different CL for the same problem. But whatever we choose, the true value is never guaranteed to be inside the interval.

The confidence level on a CI relates to hypothetical repeated sampling. We can only describe it as: if we were to repeat this experiment, then 95% of the time the CI will capture the true value. Once a CI is calculated from observation, its CL CANNOT be interpreted as a probability and it is incorrect to say: CI has 95% chance of including the parameter, or we can be 95% confident that this CI includes the true value.

In practice, we can simply say “95% confidence interval” if we are reporting results to people who understand CI. An intuitive way to explain CI to laymen is “it is the set of plausible parameter values that are consistent with the data”, where plausible indicates uncertainty in the interval. But we should always aim to use the repeated sampling explanation when communicating CIs.

9.13 Prediction Interval

A prediction interval estimates the value of an independent future observation (another sample value), rather than a parameter of the distribution, using available data (a random sample) from the same population distribution.

To derive the prediction interval, let $\bar{X}$ be the sample mean and X^* be another sample, which are independent. We aim to find a pivot that is a scaled (k) version of $\bar{X} - X^*$ (it must be a function on the sample, future observation, and known parameters), of which we can do

$$P(-c/k < \bar{X} - X^* < c/k) = 0.95$$
$$P(\bar{X} - c/k < X^* < \bar{X} + c/k) = 0.95$$

by rearranging, to get a PI for X^*.

If $X_i \sim N(\mu, \sigma^2)$, then $\bar{X} - X^* \sim N(0, \sigma^2(1+1/n))$ and $\frac{\bar{X}-X^*}{\sigma\sqrt{1+1/n}} \sim N(0,1)$ so

$$P(\bar{X} - c\sigma\sqrt{1+1/n} < X^* < \bar{X} + c\sigma\sqrt{1+1/n}) = 1 - \alpha$$

where c is the quantile from the standard normal. Notice that the prediction interval is close to the CI albeit with an extra one term under the sqrt.

We can of course repeat this if we don't know the population variance by using the sample variance and the t-dist quantiles.

For all n, the PI for X will be much wider than the CI for μ, as it includes both the uncertainty for the parameter μ and the natural variation of X around μ.

When $n \to \infty$, the width of the CI will shrink to zero, and the width of the PI will tend to the width of the population probability interval with area CL.

While a CI is a random interval with a specific hitting probability for a parameter, the PI is a random interval with a specific probability of containing another random variable (a future observation X^*).

9.14 Sample size determination

The amount of data needed for a study is determined by how much precision is required, this is commonly measured by the desired width of a CI (smaller width is more precise).

Suppose we want a CI on the population mean. The CI is of the form $(\bar{x} - c\frac{\sigma}{\sqrt{n}}, \bar{x} + c\frac{\sigma}{\sqrt{n}})$, so we can equate its width to the desired width 2ε and solve for n

$$c\frac{\sigma}{\sqrt{n}} = \varepsilon$$
$$n = (\frac{c\sigma}{\varepsilon})^2$$

Note that we round up on n if it is fractional. Of course we use a different c if the population variance is unknown using t-distr.

Suppose we want a CI on the proportions. The approximated CI is of the form

$$\hat{p} \pm c\sqrt{\frac{p(1-p)}{n}}$$

where c is a normal quantile. In order to have a desired width of 2ε, we need

$$c\sqrt{\frac{p(1-p)}{n}} = \varepsilon$$
$$n = \frac{c^2 p(1-p)}{\varepsilon^2}$$

which depends on p. We can either use a preliminary estimate of the proportions if it is available, or we can use a conservative choice of $p = 0.5$ for it maximizes $p(1-p)$ which is the worst case upper bound of the CI width

$$p(1-p) \leq 1/4, \quad n \geq \frac{c^2}{4\varepsilon^2} \implies c\sqrt{\frac{p(1-p)}{n}} \leq \varepsilon$$

10 Hypothesis Testing

10.1 Classical hypothesis testing

Classical (Neyman Pearson) hypothesis testing frames statistical inferences as research questions, and then collects data and asks whether the data support the hypothesis.

10.1.1 Hypothesis

For hypotheses

- Hypothesis is a statement about the population distribution
- Parametric hypothesis are hypothesis about the parameters of the population distribution (the parameter is x or less than or greater etc)
- Null hypothesis is a hypothesis that specifics no effect, denoted as H_0
- Alternative hypothesis, hypothesis that specifies an effect of interest (that there is an effect), denoted as H_1

Null hypothesis is a default/conservative position, corresponding to no change. This is the default for we want the study that aims to demonstrate the effect to show sufficient evidence against the null. The H_0 depends on the study, and is usually a boundary of an alternative range.

The different types of parametric hypothesis are

- Simple hypothesis (sharp hypothesis), one value for parameter(s)
- Composite hypothesis, a range of values for parameter(s)

Null hypothesis are often simple; Alternative hypothesis are typically composite with either one-sided or two-sided.

10.1.2 Tests

For tests

- A statistical test (or hypothesis test) is a decision rule for deciding between H_0 or H_1
- A test statistic T is a statistic computed from the data that the test uses to make the decision
- The critical region (rejection region) is an interval where we reject H_0 if our T is in it, its boundaries are the critical values

There are two outcomes of a test: rejecting H_0 or failing to reject H_0. We don't say that we accept H_0 if we failed to reject it, we say that there is not enough evidence to reject.

10.1.3 Errors

Errors in the statistical tests are

- Type I error, rejecting H_0 when H_0 is actually true, a false positive. Statistical tests can set an acceptable level for this error
- Type II error, failing to reject H_0 when H_0 is false, a false negative. We often can't control this

Sensitivity is the probability of a true positive (one minus the beta), specificity is the probability of a true negative (one minus the size). Note that all of the true/false positive/negative probabilities are probabilities of the test conditional on the actual state of the hypothesis.

The positive predictive value is the probability that a positive result actually implies the hypothesis (Bayes Theorem). It flips the conditional probability and represents how much power does a positive result have in predicting the true value.

The size of the test is the probability of a type I error. It is the conditional probability of rejecting H_0 given that H_0 is true.

β is the probability of a type II error, a conditional probability of not rejecting H_0 given that H_0 is false. It is well-defined only when we condition on a simple alternative hypothesis (where we substitute the range alternative with an actual value)

The power of the test is the sensitivity, it is $1 - \beta$ which is also the probability of a true positive. It is the conditional probability to reject H_0 if H_0 is false. Strictly speaking, it is a power function on the parameter

$$K(\theta) = P(rejectH_0|\theta)$$

which gives us the probability of rejecting H_0 given the parameter. If θ_0 is the null hypothesis parameter, $K(\theta_0)$ is the size of the test (prob of Type I error).

When graphing the power function, there is a dip to the size of the test around the null parameter, and it approaches one as we move away from it. It is only defined for θ where the alternative hypothesis is true, so for one-sided tests, we have a one-sided power function.

10.1.4 Controlling Errors

We can construct a test such that its size is equal to a specified significance level α, this will control its type I error. Usually set $\alpha = 0.05$ so one type I error every 20 times. If the test statistic is discrete, we may not get the size to the significance level; we might get as close to α as possible while staying below it (conservative, even lower type I error). If the size equals the significance level, we may refer to the size as α.

We want to maximize power thus minimize type II error, while keeping the significance level constraint on the size. We can increase power by choosing optimal test statistics with lower variance and/or increasing sample sizes.

10.1.5 Alternative formulations of classical h-test

There are multiple equivalent ways to formulate a test. Some are more popular because they provide extra information.

For a test using the CI, consider a $100(1-\alpha)$ CI on the parameter. We reject H_0 if the CI does not contain θ_0. This gives a test with significance level α. If the CI is constructed from a statistic T, this is equivalent to using T as a test statistic in a h-test.

For a test using the p-value. We define the p-value as the probability of observing a result (realization of T) as extreme or more extreme than the observation assuming H_0 is true. This is a tail probability of T. We reject H_0 if the p-value is less than the significance level.

P-value is a shortcut to avoid calculating a critical value. Let c be the critical value, α is a significance level. For a one-sided p-value calculation, we have

$$p = P(T < t|H_0)$$

note that we define α for a c as $\alpha = P(T < c|H_0)$, so if $t = c$, the p-value is the same as the significance level α, otherwise, it is less than α, hence we skip calculating c and comparing it with t. The decision procedures are equivalent.

For a two-sided p-value calculation, we have the two-tailed probability

$$p = P(|T| > |t||H_0)$$

for the corresponding critical value decision rule of rejecting H_0 if $|T| > c$. If T is symmetric under H_0, this is double the one-sided p-value. Assuming that we define more extreme by the quantiles if the distribution is not symmetric, generally under any distribution, the two-sided p-value is double the one-sided one.

Of course, we can also do the test using the critical value, which is computed from the α significance level using the quantile of the null distribution.

10.2 Significance testing

Only uses a null hypothesis, no reference to an alternative. Use the p-value to assess the level of significance that the data is from the null.

A low p-value is informal evidence that the null hypothesis is unlikely to be true, otherwise, collect more data. This is not a decision procedure, there are no rejecting hypothesis.

Only use significance testing to draw provisional conclusions if the problem is unknown (we can't form any alternative hypothesis).

There are clashes between significance testing and NP classical. Today, we largely use the terminologies and formulations of the classical NP testing, but report the results using a p-value and talking about rejecting/not-rejecting the null.

10.3 Common h-test scenarios

10.3.1 Single proportions

Summarized by $Y \sim Bi(n, p)$, where $H_0 : p = p_0$, and $H_1 : p > p_0$ or two tail $H_1 : p \neq p_0$. Use α.

Exact binomial test will reject H_0 if $p = P(Y \geq y|p_0) < \alpha$ for one-tail, where c is set by $P(Y \geq c|p_0) = \alpha$ and reject if $y > c$. Two-sided is not possible.

For large n, under normal approximation

$$Z = \frac{Y/n - p_0}{\sqrt{p_0(1-p_0)/n}} \approx N(0, 1)$$

where Z is our standardized test-statistic. The standardized critical values are $\Phi^{-1}(1-\alpha)$ or $\Phi^{-1}(\alpha)$. We can also compute a p-value by $P(Z < z)$ or equivalent.

Uncommonly, we can also use the unstandardized critical value

$$c = p_0 n + \Phi^{-1}(\alpha)\sqrt{np_0(1-p_0)}$$

on the raw test statistic Y.

In summary, for normal approximations

H_0	H_1	T and c
$p = p_0$	$p > p_0$	$z = \dfrac{y/n - p_0}{\sqrt{p_0(1-p_0)/n}} > \Phi^{-1}(1-\alpha)$
$p = p_0$	$p < p_0$	$z < \Phi^{-1}(\alpha)$
$p = p_0$	$p \neq p_0$	$\|z\| > \Phi^{-1}(1-\alpha/2)$

where the p-values are calculated using z and standard normal.

Note that the CI uses the sample statistic $\hat{p}$, while h-test use the null p_0.

10.3.2 Two proportions

Comparing success probability in two different populations. $H_0 : p = p_1 = p_2$ vs $H_1 : p_1 > p_2$ or equivalent. Using independent samples and n_1 and n_2 trials with Y_1 and Y_2 successes.

Exact p-value and critical value using $Y_1 + Y_2 \sim Bi(n_1 + n_2, p)$, where p can be estimated by $p = \frac{Y_1+Y_2}{n_1+n_2}$.

Under normal approximation, assuming H_0, the test statistic is

$$\begin{aligned} Z &= \frac{Y_1/n_1 - Y_2/n_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \\ &= \frac{Y_1/n_1 - Y_2/n_2}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}} \approx N(0,1) \end{aligned}$$

If we don't know p, use $\hat{p} = (y_1 + y_2)/(n_1 + n_2)$ to approximate it

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$$

Note that R treats proportion testing the same as chi-squared goodness of fit. They produce equivalent results, with the z test statistic of the proportions being square-root of the chi-squared test statistic.

Summary

H_0	H_1	T and c
$p_1 = p_2$	$p_1 > p_2$	$z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}} > \Phi^{-1}(1-\alpha)$
$p_1 = p_2$	$p_1 < p_2$	$z < \Phi^{-1}(\alpha)$
$p_1 = p_2$	$p_1 \neq p_2$	$\|z\| > \Phi^{-1}(1-\alpha/2)$

and the p-values using the test statistic and normal dist.

10.3.3 Single normal, known variance

From population $X \sim N(\mu, \sigma^2)$. $H_0 : \mu = \mu_0$ and $H_1 : \mu > \mu_0$ or equivalent.

The standardized test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

In summary

H_0	H_1	T and c
$\mu = \mu_0$	$\mu > \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > \Phi^{-1}(1 - \alpha)$
$\mu = \mu_0$	$\mu < \mu_0$	$z < \Phi^{-1}(\alpha)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\|z\| > \Phi^{-1}(1 - \alpha/2)$

and we can also do the test using the raw test statistic $\bar{X}$. The p-value uses the normal distribution.

By rearranging on the conditions that the test statistic is not in the critical region, we can see that the raw critical region is equivalent in decision-making to the CI containing μ_0 (if CI has no μ_0, test statistic in the critical region, vice versa).

10.3.4 Single normal, unknown variance

If normal but sample size is large, we can use CLT to approximate the sample mean distribution and using the standard error. When sample size is small, so we can't use CLT nor raw critical values. Identical distribution and hypotheses. The standardized test-statistic is

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

and we will conduct the test using the standardized critical values and everything.

In summary

H_0	H_1	T and c
$\mu = \mu_0$	$\mu > \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{S/\sqrt{n}} > F^{-1}(1 - \alpha)$
$\mu = \mu_0$	$\mu < \mu_0$	$z < F^{-1}(\alpha)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\|z\| > F^{-1}(1 - \alpha/2)$

where $F(q)$ is the cdf of t_{n-1}. The critical region approach is equivalent to the t-dist CI containing μ_0 approach. The p-value uses the t-distribution.

If the population is not normal and the same size is small, we need the distribution to be continuous, symmetrical and unimodal.

10.3.5 Paired normal

Similar to CI, if we observe paired data (not independent) from two populations (technically two measurements on the same population), we can apply single sample z/t tests on their differences.

10.3.6 Normal, single variance

Assuming normal $N(\mu, \sigma^2)$ samples, null hypothesis is $H_0 : \sigma^2 = \sigma_0^2$ and $H_1 : \sigma^2 \neq \sigma_0^2$ or equivalent.

The standardized test statistic under the null is

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

and our critical regions and p-value calculations uses the χ^2 distribution. We can also use the raw test statistic S^2 and find the critical values by scaling the standard χ^2's quantiles.

In summary

H_0	H_1	T and c
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} > F^{-1}(1-\alpha)$
$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\chi^2 < F^{-1}(\alpha)$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$F^{-1}(\alpha/2) > \chi^2 \vee \chi^2 > F^{-1}(1-\alpha/2)$

where F^{-1} is the inverse $\chi^2(n-1)$ cdf. To get the p-value, we check if our test statistic is above or below the median, than compute the tail probability (optionally double it if two-sided).

10.3.7 Normal, two means, equal and not equal variances

Consider two normal population with equal variances, $X \sim N(\mu_X, \sigma^2)$ and $Y \sim N(\mu_Y, \sigma^2)$, and hypotheses $H_0 : \mu_X = \mu_Y$, $H_1 : \mu_X < \mu_Y$ or equivalent.

Our test statistic using pooled variance under H_0 is

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{1/n + 1/m}} \sim t_{n+m-2}$$

which is the same as the two-sample pooled variance CI pivot without the $(\mu_X - \mu_Y)$ term due to H_0 assumption. S_P^2 is still the pooled variance estimate. All calculation done using t_{n+m-2} distribution.

Consider two normal population with different variances σ_X^2 and σ_Y^2, similar H_0 and H_1 on the means.

Our standardized test statistic has the Welch approximated t-distribution

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_X^2/n + S_Y^2/m}} \sim t_r$$

where r may be a fractional df.

If the two normal population's means are known, we use a normal test statistic.

The summary is identical to that of a single mean z/t test, with albeit different test-statistic distributions.

10.3.8 Normal, two variances

Two normal distributions with $N(\mu_X, \sigma_X^2)$ and $N(\mu_X, \sigma_Y^2)$ where $H_0 : \sigma_X^2 = \sigma_Y^2$ and $H_1 : \sigma_X^2 \neq \sigma_Y^2$.

Under H_0, the variances are equal, so the test statistic is

$$F = \frac{S_Y^2}{S_X^2} \sim F_{m-1,n-1} \quad \text{or} \quad F = \frac{S_X^2}{S_Y^2} \sim F_{n-1,m-1}$$

which is identical to the variance ratio CI pivot but canceling the variances for they are equal. Be careful of the ordering of the ratios, this test-statistic $F_{m-1,n-1}$ corresponds to the CI test for the parameter σ_Y^2/σ_X^2 with distribution $F_{n-1,m-1}$; while $1/F$ with $F_{n-1,m-1}$ corresponds to the CI test for σ_X^2/σ_Y^2 with $F_{m-1,n-1}$ which is what we've used.

Variance ratio CI always flips the order of the df on the f-dist, while variance ratio h-test doesn't.

In general for a f-distribution, we have the relationship

$$c_q(F_{n-1,m-1}) = c_{1-q}(F_{m-1,n-1})$$

which can be proven by writing out the probability and taking inverses on both sides.

Don't use the F-test in practice because it is not robust to non-normal populations. Simulations shows that the p-values are unreliable for small deviations from the normality. Instead, we could test that if

$$S_Y^2/2 < S_X^2 < 2S_Y^2$$

where we say that the variances are equal if their ratios are within $(1/2, 2)$.

10.4 Caveats of h-tests

The choice of the significance level is arbitrary. We usually follow $\alpha = 0.05$ due to convention. The appropriate balance between type I and type II error depends on the problem; different fields have different conventions for α.

The misinterpretations of p-values are

- The probability that the null hypothesis is true, or that the alternative hyp is false
- A significant p-value implies that the null hyp is false or that the alt is true
- A significant p-value implies that the effect detected is of large magnitude or of practical significant

The last point is incorrect because p-value doesn't tell you the magnitude of the effect (a large sample size with small actual differences will still net a significant p-value). Practical significant is also different from statistical significance, because we can have statistically significant p-values on an actual effect that is small and practically insignificant.

H-test with its binary decision doesn't carry much informative content; real scientific process is a process of a cumulative evidence collection, with increasing degrees of evidence instead of black and white truth claims.

H-test is popular because people want objective procedures to create conclusive truth statements, and h-test offers this with its p-values. But it is too good to be true, and p-values are easily misinterpreted, we cannot use it to draw strong conclusions.

The alternative is to formulate the problem in terms of estimation and predictions, where we estimate the size of the effect instead of a binary causation claim. We can answer these questions using CI.

Statistics is not magic, it quantifies uncertainty instead of removing it. Always think about if the results are plausible, and be conscious on how you describe your results and avoid biased reporting.

To improve the usage of h-tests, we use fisher's interpretation: h-test as an exploratory tool to inform further analyses. When reporting, report with context and avoid black and white conclusions.

It is still helpful in designing studies with the concept of error probabilities and statistical powers (balance between type I and type II error), and when we need decisions (quality control), we can use pure hypothesis to make the decisions. There are also more sophisticated procedures.

11 Regression

Regression is a simple model for a relationship between two continuous numeric variables.

We can visualize the relationship between two continuous numeric variables using paired data points on a scatter plot.

When we are interested in finding how Y depends on X, such as in regression, we assume that the X values are fixed and known (so we'll use x instead). We are then interested at the conditional probability distribution of Y given x.

Define the regression of Y on x is the conditional mean of Y given x

$$E[Y|X = x] = \mu(x)$$

where the function can be of any form.

Regression terminologies

- Y is the response variable, outcome variable, or target variable
- x is the predictor variable, explanatory variable, or covariate
- $\mu(x)$ is the linear predictor function, regression curve, or model equation
- The predictor function parameters (in this case α and β) are regression coefficients

11.1 Simple Linear Regression

A regression model is linear if its predictor function is a linear combination of the regression coefficients. It doesn't need to be a straight line in x, rather on the regression coefficients.

For instances, $\mu(x) = \alpha/x + \beta/x^2$ is linear but $\mu(x) = \alpha \sin(\beta x)$ is not. We may be able to transform models that are non-linear to linear using a different scale (taking log on both sides).

A simple linear regression model uses a straight line conditional mean, where

$$\begin{aligned} E[Y|x] &= \alpha + \beta x \\ V[Y|x] &= \sigma^2 \end{aligned}$$

where σ^2 the conditional variance of Y is independent of x.

The parameter α is the y intercept of the predictor function, but this is outside the range of the observed x values. Linear relationships are generally only for a specific range of x values, so we don't want to extrapolate the relationship outside the observed range.

We will shift our x axis so the intercept is at the x sample mean of the dataset. Let $\alpha_0 = \alpha + \beta\bar{x}$ so

$$E[Y|x] = \alpha_0 + \beta(x - \bar{x})$$

where we will perform analysis on α_0 and β instead. (This makes α_0 in the range of our data and simplifies calculations).

Note that we will implicitly assume that $E[Y_i] = \alpha_0 + \beta(x_i - \bar{x})$ for the corresponding x for Y_i is x_i.

Note that a simple linear regression of Y on x has a different model equation than a simple linear regression of X on y. But both lines pass the mean x and mean y of the dataset.

An orthogonal regression uses the sum of least squares from the points to the regression line to fit the parameters. This is commonly used to fit a bivariate distribution.

11.2 Goals

We want to

- estimate the slope, intercept, the variance of errors, and CI for all parameters
- use fitted model to make predictions on future observations, prediction intervals

note that we do not yet assume any distribution for Y, and that while Y_i are all independent, they are not identically distributed for they have different means for they depend on x_i.

11.3 Ordinary Least Squares Estimation

Define the sum of squared deviations for a given sample for the set of parameters as

$$H(\alpha_0, \beta) = \sum_i (y_i - \alpha_0 - \beta(x_i - \bar{x}))^2$$

The least squares estimation (or ordinary least squares) tells us to choose the regression coefficients that minimize the least squares. This leads to the normal equations (first order minima conditions)

$$0 = \frac{\partial}{\partial \alpha_0} H(\hat{\alpha}_0, \hat{\beta})$$
$$0 = \frac{\partial}{\partial \beta} H(\hat{\alpha}_0, \hat{\beta})$$

Some algebra yields the least square estimators

$$\hat{\alpha}_0 = \bar{Y}$$
$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_i (x_i - \bar{x})^2}$$

note that alternative $\hat{\beta}$ expression is due to $\sum_i (x_i - \bar{x})\bar{Y} = 0$.

The estimator for α is

$$\hat{\alpha} = \hat{\alpha}_0 - \hat{\beta}\bar{x} = \bar{Y} - \hat{\beta}\bar{x}$$

and the estimator for the predictor function

$$\hat{\mu}(x) = \hat{\alpha}_0 + \hat{\beta}(x_i - \bar{x}) = \bar{Y} + \hat{\beta}(x - \bar{x})$$

The fitted model equation (or predictor function) is just the estimated predictor function for the sample.

11.4 Coefficient estimator means

These coefficient estimators are all linear combinations of Y_i given x_i

$$\begin{aligned}\hat{\alpha}_0 &= \sum_i \frac{1}{n} Y_i \\ \hat{\beta} &= \sum_i \frac{x_i - \bar{x}}{K} Y_i\end{aligned}$$

where $K = \sum_i (x_i - \bar{x})^2$, always remembering that x_i are not random but rather fixed values.

Using this fact, the means of the coefficient estimators are

$$\begin{aligned}E[\hat{\alpha}_0] &= \frac{1}{n}\sum_i E[Y_i] \\ &= \frac{1}{n}\sum_i \alpha_0 + \beta(x_i - \bar{x}) \\ &= \alpha_0 \\ E[\hat{\beta}] &= \sum_i \frac{x_i - \bar{x}}{K} E[Y_i] \\ &= \frac{1}{K}\sum_i (x_i - \bar{x})(\alpha_0 + \beta(x_i - \bar{x})) \\ &= \frac{1}{K}\sum_i (x_i - \bar{x})\alpha_0 + \frac{K}{K}\beta \\ &= \beta\end{aligned}$$

and thus

$$\begin{aligned}
E[\hat{\alpha}] &= E[\alpha_0] - E[\beta\bar{x}] \\
&= \alpha \\
E[\hat{\mu}(x)] &= E[\hat{\alpha}_0 + \hat{\beta}(x - \bar{x})] \\
&= \mu(x)
\end{aligned}$$

Therefore, all estimators are unbiased.

11.5 Coefficient estimator variances

The variances are

$$\begin{aligned}
V[\hat{\alpha}_0] &= \frac{1}{n^2}\sum_i V[Y_i] \\
&= \frac{\sigma^2}{n} \\
V[\hat{\beta}] &= \sum_i \left(\frac{x_i - \bar{x}}{K}\right)^2 V[Y_i] \\
&= \frac{1}{K^2}\sum_i (x_i - \bar{x})^2\sigma^2 \\
&= \frac{\sigma^2}{K}
\end{aligned}$$

Their covariance is

$$\begin{aligned}
Cov(\hat{\alpha}_0, \hat{\beta}) &= Cov(\frac{1}{n}\sum_j Y_j, \sum_i \frac{x_i - \bar{x}}{K} Y_i) \\
&= \sum_j \sum_i \frac{x_i - \bar{x}}{nK} Cov(Y_j, Y_i) \\
&= \sum_i \frac{x_i - \bar{x}}{nK} V[Y_i] \\
&= 0
\end{aligned}$$

because $Cov(Y_j, Y_i) = V[Y_i]$ only when $i = j$ as the Y_is are independent.

Furthermore, the variances are

$$\begin{aligned}
V[\hat{\alpha}] &= V[\hat{\alpha}_0 - \hat{\beta}\bar{x})] \\
&= (\frac{1}{n} + \frac{(\bar{x})^2}{K})\sigma^2 \\
V[\hat{\mu}(x)] &= V[\hat{\alpha}_0 + \hat{\beta}(x - \bar{x})] \\
&= (\frac{1}{n} + \frac{(x - \bar{x})^2}{K})\sigma^2
\end{aligned}$$

Note that the smallest variance of the model equation (estimate of the $Y(x)$ mean) is near the centroid $\bar{x}$, with the variance increasing quadratically as we move further away from the x mean.

11.6 Model variance/error estimator

To find the standard error of the coefficient estimators, we need a model variance σ^2 estimator.

Remember that we've used the sample variance as an unbiased estimator for the single population variance. Given $X \sim N(\mu, \sigma^2)$

$$E[S^2] = E[\frac{\sum_i (X_i - \bar{X})^2}{n-1}] = \sigma^2$$

where we computed $\sum_i (X_i - \mu)^2 = \sum_i (X_i - \bar{X} + \bar{X} - \mu)^2$, simplified and took expectations on both sides, then rearranging for an unbiased estimator for the population variance.

We will apply that technique with the regression model, using $a = \hat{\alpha}_0$ and $b = \hat{\beta}$

$$\begin{aligned}
\sum_i (Y_i - E[Y_i])^2 &= \sum_i (Y_i - \alpha_0 - \beta(x_i - \bar{x}))^2 \\
&= \sum_i (Y_i - a - b(x_i - \bar{x}) + a + b(x_i - \bar{x}) - \alpha_0 - \beta(x_i - \bar{x}))^2 \\
&= \sum_i (Y_i - a - b(x_i - \bar{x}) + (a - \alpha_0) + (b - \beta)(x_i - \bar{x}))^2 \\
&= \sum_i (Y_i - a - b(x_i - \bar{x}))^2 + t_1 + \sum_i (a - \alpha_0)^2 + t_3 + \sum_i ((b - \beta)(x_i - \bar{x}))^2
\end{aligned}$$

where t_1 and t_3 are the cross terms of the two squarings. Noticing that $\sum x_i - \bar{x} = 0$ and $\sum Y_i - a = \sum Y_i - \bar{Y} = 0$ on t_1 and t_3, and expanding the remaining t_2, the cross terms are

$$\begin{aligned}
t_1 &= 2\sum_i (Y_i - a - b(x_i - \bar{x}))(a - \alpha_0) + (Y_i - a - b(x_i - \bar{x}))(b - \beta)(x_i - \bar{x}) \\
&= 2(0 + 0 + t_2) \\
t_2 &= 2\sum_i (Y_i - a - b(x_i - \bar{x}))(b - \beta)(x_i - \bar{x}) \\
\frac{t_2}{2(b-\beta)} &= \sum_i (Y_i - a)(x_i - \bar{x}) - \sum_i b(x_i - \bar{x})(x_i - \bar{x}) \\
&= \sum_i (Y_i - \bar{Y})(x_i - \bar{x}) - bK \\
&= \sum_i (Y_i - \bar{Y})(x_i - \bar{x}) - K\sum_i \frac{x_i - \bar{x}}{K} Y_i \\
&= -\sum_i \bar{Y}(x_i - \bar{x}) \\
&= 0 \\
\implies t_2 &= 0 \\
\implies t_1 &= 0 \\
t_3 &= 2\sum_i (a - \alpha_0)(b - \beta)(x_i - \bar{x}) \\
&= 0
\end{aligned}$$

Hence going back to the original regression model formula and taking expectations

$$\begin{aligned}
\sum_i (Y_i - E[Y_i])^2 &= \sum_i (Y_i - a - b(x_i - \bar{x}))^2 + n(a - \alpha_0)^2 + K(b - \beta)^2 \\
n\sigma^2 &= E[\sum_i (Y_i - \hat{\mu}(x_i))^2] + \sigma^2 + \sigma^2 \\
\implies E[\sum_i (Y_i - \hat{\mu}(x_i))^2] &= \sigma^2(n - 2)
\end{aligned}$$

And we end up with an unbiased estimator for σ^2

$$\hat{\sigma}^2 = \frac{\sum_i (Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x}))^2}{n - 2}$$

noticing that

- $\hat{Y}_i = \hat{\alpha}_0 + \hat{\beta}(x_i - \bar{x})$ are the fitted values for the means of Y_i at x_i using the fitted model

- $R_i = Y_i - \hat{Y}_i$ are the residuals, which are deviations of the observed Y values from its inferred means
- $D^2 = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i R_i^2$ is the sum of squared residuals, which we base our variance estimator on

Intuitively, we estimate the model variance by the sum of squared deviations of each observation divided by $n-2$ for the 2 degrees of freedom lost by estimating the coefficients.

11.7 Coefficient estimator standard errors

We can simply substitute the model variance estimator $\hat{\sigma}^2$ into the standard deviation formulas of the coefficient estimators to calculate their standard errors, like

$$SE[\hat{\alpha}_0] = \frac{\hat{\sigma}}{\sqrt{n}}, \quad SE[\hat{\beta}] = \frac{\hat{\sigma}}{\sqrt{K}}, \quad SE[\hat{\mu}(x)] = \hat{\sigma}\sqrt{1/n + (x-\bar{x})^2/K}$$

11.8 Confidence Intervals

To construct CIs, we need to assume that the response variables Y_i are normally distributed

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

or in an alternative notation

$$Y_i = \alpha + \beta x_i + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

11.8.1 MLE method

Consider the MLE on the regression coefficients α_0 and β, remembering that the Y_i are independent

$$\begin{aligned}
L(\alpha, \beta, \sigma^2) &= \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \alpha_0 - \beta(x_i - \bar{x}))^2}{2\sigma^2}} \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_i (y_i - \alpha_0 - \beta(x_i - \bar{x}))^2} \\
-\log L(\alpha, \beta, \sigma^2) &= \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_i (y_i - \alpha_0 - \beta(x_i - \bar{x}))^2 \\
&= \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2} H(\alpha_0, \beta)
\end{aligned}$$

Notice that the second term is the sum of squares and minimizing it is independent to σ^2. Hence, the α_0 and β that maximizes the likelihood are the same as the parameters that minimizes the sum of squares. This means that the OLS estimators are exactly the MLEs.

To find the MLE for σ^2, we can differentiate the likelihood against σ^2 to find

$$\hat{\sigma}^2 = \frac{\sum_i R_i^2}{n}$$

which is biased. We often prefer to use the unbiased estimator which is $\hat{\sigma}^2 = (\sum_i R_i^2)/(n-2)$.

11.8.2 Estimator Normality and Independence

We need the independence of the model variance estimator to compute pivots.

Under the Y_i independent normal assumption, because the coefficient estimators are both linear combinations of Y_i, they also have normal distributions with

$$\hat{\alpha}_0 \sim N(\alpha_0, \frac{\sigma^2}{n})$$
$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{K})$$

We characteristic a bivariate normal distribution as: two random variables U and V have a bivariate normal distribution iff all linear combinations of them have a normal distribution. Also note that if U and V are bivariate normal with zero covariance, then they are independent.

As our regression coefficients are weighted sums of independent Y_is, any linear combination of $\hat{\alpha}_0$ and $\hat{\beta}$ is also a linear combination of Y_i, which must be normal. Hence the regression coefficients form a bivariate normal distribution, and their zero covariance indicates that they are independent.

Similar to the sample variance, we can also establish independence of the variance decomposition terms for the model variance estimator to get

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

and find that it is also independent of both regression coefficients. Hence, the three estimators are pairwise independent. The proof of this uses the fact that when dividing everything by σ^2, the other three terms are χ^2 distributed and apply mgf on both sides.

11.8.3 Actual CIs

The important pivots are

$$\frac{\hat{\beta} - \beta}{\hat{\sigma}/\sqrt{K}} \sim t_{n-2}$$
$$\frac{\hat{\mu}(x) - \mu(x)}{\hat{\sigma}\sqrt{1/n + (x - \bar{x})^2/K}} \sim t_{n-2}$$

without the intercept for the intercept is rarely interesting.

Using the t-distribution CIs of form

$$\bar{x} \pm F^{-1}(1 - \alpha/2)SE$$

We can derive the CIs for the slope β and the means $\mu(x)$

$$CI(\beta) = \hat{\beta} \pm c\frac{\hat{\sigma}}{\sqrt{K}}$$
$$CI(\mu(x)) = \hat{\mu}(x) \pm c\hat{\sigma}\sqrt{1/n + (x - \bar{x})^2/K}$$

where c is the quantile from t_{n-2}.

We can similarly run t-distribution hypothesis tests on the slope or the means. Particularly, $H_0 : \beta = 0$ while $H_1 : \beta \neq 0$. This slope h-test allows us to test the statistical significance of the linear relationship between x and the mean response. Our test-statistic would be exactly the pivot with a t_{n-2} distribution.

R uses the uncentered α for its CIs and tests, but it gives the same results on estimators other than the intercept. We should always drop reference to the intercepts, centered or not, for it causes confusion in h-tests. The R command is `lm(response ~ predictor)`, of which the summary computes p-values and test-statistics for the null hypothesis of $H_0 : \alpha = 0$ and $H_0 : \beta = 0$.

11.9 Prediction Intervals

Using a similar trick as the one-sample model, we can derive the PI for $Y^*(x^*)$ by

$$Y^* \sim N(\mu(x^*), \sigma^2)$$
$$Y^* - \hat{\mu}(x^*) \sim N(0, (1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{K})\sigma^2)$$
$$PI(Y^*) = \hat{\mu}(x^*) \pm c\hat{\sigma}^2\sqrt{1 + 1/n + (x^* - \bar{x})^2/K}$$

where c is the quantile from the t_{n-2} distribution.

We can use `confint(model)` to get the parameter CIs. We can use `predict(model, newdata=data.frame(x=2), interval="")` for either the CIs or PIs on the mean of $Y(x)$.

We can use `abline(model)` to add the line of best fit.

The prediction intervals will always be wider than the CI, and they are both the narrowest around $\bar{x}$.

11.10 Linear regression assumptions

The assumptions we've made for linear regression is

- Linear model for the mean
- Equal variance for all observations (homoscedasticity)
- Normally distributed residuals

We can check these assumptions by

- Plot the data and fitted model curve
- Plot residuals vs fitted values
- QQ plot of residuals

in that order

The residuals vs fitted values plot should have equal width of residuals, random points above and below, and no trend or fans.

11.11 Multiple Regression

Suppose we observe more than one predictor, say $x_{i1}, \dots, x_{ik}$ for each Y_i. We can fit a multiple linear regression model

$$E[Y|x_1, \dots, x_k] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

and we can similarly fit the coefficients by method of minimizing least squares.

We will not go into these models in great detail.

We can phrase the two-sample problem (difference in population means) with equal variance into a multiple linear regression model. We need to merge the two observations and define two indicator variables where they are $(0, 1)$ for the first population and $(1, 0)$ for the second, so the β_i corresponds to the means of the two populations. A general linear model unifies different models together into a common framework.

WE DO NOT NEED TO CARE ABOUT THIS.

12 Correlation

Define the correlation coefficient for two rv X and Y as

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

where $\sigma_{XY} = Cov(X, Y)$. We'd like to do inference on ρ given an iid sample pairs (X_i, Y_i).

To estimate the covariance, consider the sample covariance

$$S_{XY} = \frac{1}{n-1} \sum_i (X_i - \bar{X})(Y_i - \bar{Y})$$

which is unbiased, with $E[S_{XY}] = \sigma_{XY}$.

Define the sample correlation coefficient (Pearson's correlation coefficient) to estimate ρ

$$R_{XY} = \frac{S_{XY}}{S_X S_Y}$$

notice that the $n - 1$ in the numerator and denominator will cancel out when expanded. We also have that $|R| \leq 1$ just like $|\rho| \leq 1$. This sample correlation gives a point estimate to ρ.

Assume X and Y have a bivariate normal distribution with correlation ρ. The regression of Y on X (and vice versa) is linear

$$\begin{aligned}
E[Y|X = x] &= \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X) \\
&= \alpha_0 + \beta(x - \mu_X) \\
E[X|Y = y] &= \mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y - \mu_Y) \\
&= \alpha_0' + \beta'(y - \mu_Y)
\end{aligned}$$

This shows that we can always use linear regression on a bivariate normal population.

To see the connection between correlation and linear regression, we will analyze the variability of Y in the regression model a different way. Suppose we have independent pairs

(x_i, Y_i) and parameter estimates $\hat{\alpha}_0$ and $\hat{\beta}$, and the fitted values $\hat{Y}_i = \hat{\mu}(x_i)$. Consider

$$\begin{aligned}
\sum_i (Y_i - \bar{Y})^2 &= \sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\
&= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 + t_1 \\
t_1 &= \sum_i (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\
&= \sum_i (Y_i - \bar{Y} - \hat{\beta}(x_i - \bar{x}))\hat{\beta}(x_i - \bar{x}) \\
&= \hat{\beta}\left[\sum_i (Y_i - \bar{Y})(x_i - \bar{x}) - \hat{\beta}\sum_i (x_i - \bar{x})^2\right] \\
&= \hat{\beta}\left[\hat{\beta}K - \hat{\beta}K\right] \\
&= 0
\end{aligned}$$

Which gives us the linear regression variance decomposition

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2$$

which are the sum of squares

$$SS(Total) = SS(Error) + SS(Model)$$

where the model SS (also known as regression SS) is the explained variance by the model. The error SS is the unexplained variation around the mean line.

We can also show that

$$SS(Error) = (1 - R^2)SS(Total) \quad SS(Model) = R^2 SS(Total)$$

implying that R^2 (using the sample correlation estimator) is the proportion of Y variability explained by x. In this context, R^2 is the coefficient of determination. Intuitively, consider the conditional variance of the bivariate normal, and see that the term $(1 - \rho^2)$ appear signifying the unexplained variability after linear regression.

Some remarks about R^2

- Under simple linear regression, the coefficient of determination is the square of the sample correlation
- The proportion of Y explained by X is the same as the proportion of X explained by Y, both equaling R^2

- The R^2 for more complicated models needs to be calculated using all predictor variables, hence it will not be the sample correlation

To approximate the sampling distribution of R, define the fisher transformation

$$g(r) = \frac{1}{2}\log(\frac{1+r}{1-r}) = \text{artanh}(r), \quad g^{-1}(y) = \tanh(y)$$

This is used to approximate the sampling distribution of R under a bivariate normal distribution by

$$g(R) \approx N(g(\rho), \frac{1}{n-3})$$

and can be used to construct CIs. Note that because $E[g(R)] = g(\rho)$, then $\rho \neq E[g^{-1}(g(R))] = E[R]$, hence our sample correlation is not unbiased for ρ.

We call the fisher transformation $g()$ variance stabilizing because the limiting approximation variance of $g(R)$ does not depend on ρ.

The approximate CI for ρ is therefore

$$CI(\rho) = (g^{-1}\left(g(r) - \frac{c}{n-3}\right), g^{-1}\left(g(r) + \frac{c}{n-3}\right))$$

where c is the quantile from the standard normal.

We can use the R `cor.test` to construct this CI.

13 Analysis of Variance

The goal of anova is to compare the means of more than two populations. In general, however, anova explores the components of variances, and evaluates the fit of a linear model.

The procedure is a h-test, and it involves the variance decomposition formulas to compare different summaries of variation (namely the between group variance and within group variance).

13.1 One Way Anova

Random samples from k independent populations, each having a normal distribution. The ith group has n_i iid observations with mean μ_i. All group populations have the same variance σ^2. The h-test of ANOVA tests whether the population means are equal

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad H_1 : \text{not } H_0$$

This is called one-way anova or single-factor anova because group is the only factor.

The alternative hypothesis in ANOVA covers anything other than H_0, they could be

- All means are the same except one
- All means are distinct
- Population is clustered into subpopulations, so the groups within the clusters have the same mean, while groups between clusters have different means

The notations of one-way anova

- Population for the ith group is $N(\mu_i, \sigma^2)$
- The samples from the ith group are $X_{i,1}, X_{i,2} \dots X_{i,n_i}$
- The sample statistics of the ith group is $\bar{X}_i$ and S_i^2
- The overall sample mean and sample variance across all groups is $\bar{X}$ and S^2.

Be careful that the index order of the observation differs between texts (we use group first than within group second). The dot notation is representing $\bar{X}_i$ as $\bar{X}_{i,\cdot}$ and $\bar{X}$ as $\bar{X}_{\cdot,\cdot}$, where the dot shows the index that we've summed over (in this case second index first and all index second).

Some formulas for the sample statistics

$$\begin{aligned} n &= \sum_{i=1}^{k} n_i \\ \bar{X}_{i,\cdot} &= \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j} \\ \bar{X}_{\cdot,\cdot} &= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{i,j} = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{X}_{i,\cdot} \end{aligned}$$

where we have the: total sample size, group means, grand/overall mean.

13.1.1 ANOVA Decomposition

The SS is a proxy for variability.

We wish to define the sum of squares (SS) statistics for the sample. Define the total SS as

$$SS(total) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X})^2$$

which is the sum of squares of the observations against the grand mean. Define the treatment SS (or between group SS) as

$$SS(treatment) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

which is the sum of squares where we compare the sample group means against the grand mean. Define the error SS (or within group SS) as

$$SS(error) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_i)^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2$$

which are the sum of squares between observations and their sample group mean.

By decomposing the variances (or SS), we have that

$$SS(total) = SS(treatment) + SS(error)$$

which are similar to all the variance decomposition formulas derived earlier in single-mean model for the sample variance and regression model for the model variance estimator.

To prove this result, we use the add-and-subtract trick

$$\begin{aligned}
SS(total) &= \sum_i\sum_j(X_{i,j} - \bar{X})^2 \\
&= \sum_i\sum_j(X_{i,j} - \bar{X}_i + \bar{X}_i - \bar{X})^2 \\
&= \sum_i\sum_j(X_{i,j} - \bar{X}_i)^2 + \sum_i\sum_j(\bar{X}_i - \bar{X})^2 + t_1 \\
t_1 &= 2\sum_i\sum_j(X_{i,j} - \bar{X}_i)(\bar{X}_i - \bar{X}) \\
&= 2\sum_i(\bar{X}_i - \bar{X})\sum_j(X_{i,j} - \bar{X}_i) \\
&= 0 \\
SS(total) &= SS(error) + SS(treatment)
\end{aligned}$$

where t_1, the cross term, is zero because $\sum_j(X_{i,j} - \bar{X}_i)$ is zero.

13.2 SS statistics sampling distributions

Note that all the SS are statistics with a sampling distribution.

13.2.1 No H_0 assumption

Don't assume H_0 yet. To find the sampling distribution of $SS(error)$, we know that S_i^2 from each group is an unbiased estimator of σ^2 with $(n_i - 1)S_i^2/\sigma^2 \sim \chi^2(n_i - 1)$. Because the samples from each group are independent, the sum of the scaled sample variances are

$$\sum_i \frac{(n_i - 1)S_i^2}{\sigma^2} = \frac{SS(error)}{\sigma^2} \sim \chi^2(n-k)$$

This also nets an unbiased model variance estimator

$$\hat{\sigma}^2 = \frac{SS(error)}{n-k}$$

To show that $SS(error)$ and $SS(treatment)$ are independent, notice that each group has normally distributed samples, so we have that $\bar{X}_i$ is independent with S_i^2 for every i. Therefore

$$\begin{aligned} SS(treatment) &= \sum_i n_i(\bar{X}_i - \bar{X})^2 \\ &= \sum_i n_i(\bar{X}_i - \frac{1}{n}\sum_i n_i\bar{X}_i)^2 \\ SS(error) &= \sum_i (n_i - 1)S_i^2 \end{aligned}$$

where $SS(treatment)$ is a function of $\bar{X}_i$ and $SS(error)$ is a function of S_i^2. Hence $SS(treatment)$ and $SS(error)$ are independent.

13.2.2 H_0 assumption

Now we assume H_0. To find the distribution of $SS(total)$, notice that the combined data is a sample from $N(\mu, \sigma^2)$. Then

$$\begin{aligned} E[\frac{\sum_i\sum_j(X_{i,j} - \bar{X})^2}{n-1}] &= E[SS(total)/(n-1)] = \sigma^2 \\ \frac{SS(total)/(n-1)(n-1)}{\sigma^2} &= \frac{SS(total)}{\sigma^2} \\ &\sim \chi^2(n-1) \end{aligned}$$

by the sample variance properties.

To find the distribution of $SS(treatment)$, notice that $\bar{X}_i \sim N(\mu, \sigma^2/n_i)$, and treating $\bar{X}_i$ as a sample of the sample means, we have that

$$\begin{aligned}
E[\frac{\sum_i(\sqrt{n_i}\bar{X}_i - \sqrt{n_i}\bar{X})^2}{k-1}] &= \sigma^2 \\
\frac{(\sum_i n_i(\bar{X}_i - \bar{X})^2)/(k-1)(k-1)}{\sigma^2} &= \frac{SS(treatment)}{\sigma^2} \\
&\sim \chi^2(k-1)
\end{aligned}$$

where we've used the sample variance property on the scaled sample means (but in this case the sample mean changes in the sum). Alternatively, we could use the independence of $SS(treatment)$ and $SS(error)$ to derive the distribution of $SS(treatment)$ from the variance decomposition of $SS(total)$ (this is shown in the summary below).

To summarize, under H_0, we have that

$$\begin{aligned}
\frac{SS(total)}{\sigma^2} &= \frac{SS(treatment)}{\sigma^2} + \frac{SS(error)}{\sigma^2} \\
\frac{SS(total)}{\sigma^2} &\sim \chi^2(n-1) \\
\frac{SS(treatment)}{\sigma^2} &\sim \chi^2(k-1) \\
\frac{SS(error)}{\sigma^2} &\sim \chi^2(n-k)
\end{aligned}$$

with $SS(treatment)$ and $SS(error)$ being independent.

13.2.3 Under H_1

Intuitively, H_1 will make $SS(treatment)$ larger.

More formally, with no assumptions, let $\bar{\mu} = \frac{1}{n}\sum_i n_i \mu_i$ and consider

$$\begin{aligned}
E[SS(treatment)] &= E[\sum_i n_i(\bar{X}_i - \bar{X})^2] \\
&= E[\sum_i n_i \bar{X}_i^2 - n\bar{X}^2] \\
&= \sum_i n_i E[\bar{X}_i^2] - nE[\bar{X}^2] \\
&= \sum_i n_i \left(V[\bar{X}_i] + E[\bar{X}_i]^2\right) - n(V[\bar{X}] + E[\bar{X}]^2) \\
&= \sum_i^k n_i(\sigma^2/n_i + \mu_i^2) - n(\sigma^2/n + \bar{\mu}^2) \\
&= (k-1)\sigma^2 + \sum_i n_i \mu_i^2 - n\bar{\mu}^2 \\
&= (k-1)\sigma^2 + \sum_i n_i(\mu_i - \bar{\mu})^2
\end{aligned}$$

where we've used the identity that

$$\begin{aligned}
\bar{X} &= \frac{1}{n}\sum_i n_i \bar{X}_i \\
-2\sum_i n_i \bar{X}_i \bar{X} + \sum_i n_i \bar{X}^2 &= -2n\bar{X}^2 + n\bar{X}^2 \\
&= -n\bar{X}^2
\end{aligned}$$

for the initial cross term and the last cross term.

For the SS of the treatment, if H_0 is true, the second term is zero so

$$\frac{E[SS(treatment)]}{k-1} = \sigma^2$$

which is also evident from our H_0 sampling distribution. Otherwise, the second term is positive and

$$\frac{E[SS(treatment)]}{k-1} > \sigma^2$$

In contrast for the SS of error, we always have

$$\frac{E[SS(error)]}{n-k} = \sigma^2$$

13.2.4 F-statistic

By the χ^2 distributions and that both SS having the same expected values (except when H_0 is false), we are motivated to use the f test statistic

$$F = \frac{SS(treatment)/(k-1)}{SS(error)/(n-k)}$$

where the σ^2 cancels from the chi-squared distributions.

Under H_0, $F \sim F_{k-1,n-k}$ for it is a ratio of independent chi-squares. Under H_1 the numerator tends to be larger for its expected value is higher than that of H_0. Hence, we reject H_0 if $F > c$ for some critical value c. Note that this test is ALWAYS one-sided.

This is the f-test.

We can summarize the summary statistics used for the test in an ANOVA table

Source	df	SS	MS	F
Treatment	$k-1$	$\sum_i n_i(\bar{X}_i - \bar{X})^2$	$SS(treatment)/(k-1)$	$MS(treatment)/MS(error)$
Error	$n-k$	$\sum_i n_i S_i^2$	$SS(error)/(n-k)$	
Total	$n-1$	$SS(total)$		

Where we define $MS(A) = SS(A)/df$ as the mean squares by the df. Note that $\hat{\sigma}^2 = MS(error)$ is always an unbiased estimator for the model variance.

The MSE is kinda like the sample variance.

13.2.5 R

In R, we use `lm(response ~ factor(predictor))` to create the model, where `factor` denotes categorical variables. R also use residuals as the error row. To see the anova table, use `anova(model)`.

13.3 Two-Factor ANOVA

A two-way (or two-factor) ANOVA has observations partitioned by two factors (categorical variables).

13.3.1 Single Observation, Additive model

Suppose that we only have one observation per factor combination. The assumptions are

- a as the number of groups for factor 1, and b as the number of groups for factor 2

- X_{ij} as the observation with first factor at i and second factor at j
- A total of $n = ab$ observations
- $X_{ij} \sim N(\mu_{ij}, \sigma^2)$ and are independent

Then the two-way anova model is

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$
$$\sum_i^a \alpha_i = 0$$
$$\sum_j^b \beta_j = 0$$

with the last two conditions being regulatory constraints; if these two constraints are fulfilled, it is called an orthogonal parameterization. We can see that μ is the overall effect, α_i is the effect of the ith level of factor 1 and β_j is the effect of the jth level of factor 2. This model implies that each level of factor 1 and factor 2 adds the same effect regardless of the level of the other factor.

Under this model, we are interested in either

$$H_{0A} : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$
$$H_{0B} : \beta_1 = \beta_2 = \cdots = \beta_b = 0$$

that is, does factor 1 or factor 2 have an actual statistically significant effect.

Define

$$\bar{X}_{\cdot\cdot} = \frac{1}{ab} \sum_i \sum_j X_{ij}$$
$$\bar{X}_{i\cdot} = \frac{1}{b} \sum_j X_{ij}$$
$$\bar{X}_{\cdot j} = \frac{1}{a} \sum_i X_{ij}$$

which are the grand means, factor 1 means, and factor 2 means.

Similar to the one-way decomposition, we have

$$\begin{aligned}
SS(total) &= \sum_i \sum_j (X_{ij} - \bar{X})^2 \\
&= \sum_i \sum_j [(\bar{X}_{i\cdot} - \bar{X}) + (\bar{X}_{\cdot j} - \bar{X}) + (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})]^2 \\
&= b\sum_i (\bar{X}_{i\cdot} - \bar{X})^2 + a\sum_j (\bar{X}_{\cdot j} - \bar{X})^2 + \sum_i \sum_j (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2 \\
&= b\sum_i (\bar{X}_{i\cdot} - \bar{X})^2 + a\sum_j (\bar{X}_{\cdot j} - \bar{X})^2 + \sum_i \sum_j (X_{ij} - (\bar{X}_{i\cdot} - \bar{X}) - (\bar{X}_{\cdot j} - \bar{X}) - \bar{X})^2 \\
&= SS(factor1) + SS(factor2) + SS(error)
\end{aligned}$$

where we assume that all the cross terms are zero. Let factor 1 be A and factor 2 be B. Additionally, this implies that the estimators for α and β are

$$\begin{aligned}
\hat{\alpha}_i &= \bar{X}_{i\cdot} - \bar{X} \\
\hat{\beta}_j &= \bar{X}_{\cdot j} - \bar{X} \\
\hat{\mu} &= \bar{X}
\end{aligned}$$

If both H_{0A} and H_{0B} are true, then

$$\begin{aligned}
SS(A)/\sigma^2 &\sim \chi^2(a-1) \\
SS(B)/\sigma^2 &\sim \chi^2(b-1) \\
SS(error)/\sigma^2 &\sim \chi^2((a-1)(b-1))
\end{aligned}$$

and these statistics are also independent. We ignore the proofs here.

The f test statistic for H_{0A} is

$$F_A = \frac{SS(A)/(a-1)}{SS(error)/((a-1)(b-1))} \sim F_{a-1,(a-1)(b-1)}$$

and we reject H_0 if $F_A > c$. Similarly, the test-statistic for H_{0B} is that of H_{0A} but replacing A with B.

The two-way ANOVA table is

Source	df	SS	MS	F
Factor A	$a-1$	$SS(A)$	$SS(A)/(a-1)$	$MS(A)/MS(error)$
Factor B	$b-1$	$SS(B)$	$SS(B)/(b-1)$	$MS(B)/MS(error)$
Error	$(a-1)(b-1)$	$SS(error)$	$SS(error)/((a-1)(b-1))$	
Total	$ab-1$	$SS(total)$		

Once again, $\hat{\sigma}^2 = MS(error)$ is an unbiased estimator for the model variance always.

The df of the SS still adds up. First note that the factor A and B SS df has one less df because of the regulatory constraints. Then $df(error) = ab - 1 - df(A) - df(B)$ which is also just removing the last row and last column, and $df(total) = df(A) + df(B) + df(error)$.

Interaction plot is a grouped line plot showing the effects of two categorical factors on a response variable. We should expect parallel lines in the plot if the effects really are additive, because changing just one factor should have the same effect.

The R anova command outputs the results of both hypothesis tests as well as the full anova table.

If we have multiple samples and want to use the additive model, simply multiply everything (df, sample size, row column means, etc) by c which is the number of samples in each cell.

13.3.2 Interaction/General two-way anova

The additive model (for single observation) of two-way anova assumes that the relative effects of factor 1 is the same on all levels of factor 2. If this is not true, then there is a statistical interaction (or just interaction) between the two factors.

The general model that includes interactions is

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

where γ_{ij} is the interaction term. In addition to the additive constraints, we also impose

$$\sum_{i=1}^{a} \gamma_{ij} = 0 \quad \sum_{j=1}^{b} \gamma_{ij} = 0$$

We say that the terms α_i and β_j are the main effects (or average row and column effects if written as a table).

Now the hypothesis we can test are

- Are the row effects (α_i) zero
- Are the column effects (β_i) zero
- Are the interaction effects (γ_{ij}) zero

We should test for interaction effects first, and only use this model if the interactive effects are significant. Otherwise, use the additive model to test the row and column effects.

We need more than one observation per cell to conduct inference on the interactions. So define

- X_{ijk} to be the kth observation for factors (i, j)
- Assume that there are always c observations for each pair of factor combination (assuming that the design is balanced)

Then the means are

$$\begin{aligned}\bar{X}_{ij\cdot} &= \frac{1}{c}\sum_k X_{ijk} \\ \bar{X}_{i\cdot\cdot} &= \frac{1}{bc}\sum_{j,k} X_{ijk} \\ &\vdots\end{aligned}$$

and etc, just remembering that we need to divide by the size of the missing axes when taking the mean over it.

The anova decomposition is then

$$\begin{aligned}SS(total) &= \sum_{i,j,k}(X_{ijk} - \bar{X})^2 \\ &= bc\sum_i(\bar{X}_{i\cdot\cdot} - \bar{X})^2 + ac\sum_j(\bar{X}_{\cdot j\cdot} - \bar{X})^2 \\ &+ c\sum_{i,j}(\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X})^2 \\ &+ \sum_{i,j,k}(X_{ijk} - \bar{X}_{ij\cdot})^2 \\ &= SS(A) + SS(B) + SS(AB) + SS(error)\end{aligned}$$

once again, we ignore all cross terms. We state without proof that all SS statistics are scaled chi-squared distributions (under no factor influences) and are independent. The distributions of the new terms under no factor effects are

$$SS(AB)/\sigma^2 \sim \chi^2((a-1)(b-1)) \quad SS(error)/\sigma^2 \sim \chi^2(ab(c-1))$$

Notice that $SS(AB)$ is exactly the $SS(error)$ of the no-interaction model replacing X_{ijk} with $X_{ij\cdot}$. This is because now we have another parameter γ_{ij} to explain the variability of points in the same cell, which we will intuitively set using the mean difference between the A and B prediction and the cell observations, hence eliminating the biases of A and B predictions, while the variances of those predictions are now the new error terms. This explains why we need more than one observation, for if there are only one observation per cell, the γ_{ij} term will perfectly explain all variability (seen intuitively or by noticing that $SS(error)$ will be zero).

The interaction test has the hypothesis

$$H_{0AB} : \gamma_{ij} = 0$$

Using the no factor chi-squared distributions of $SS(AB)$, the test-statistic for the interaction test under null is

$$F = \frac{SS(AB)/((a-1)(b-1))}{SS(error)/(ab(c-1))} \sim F_{(a-1)(b-1),ab(c-1)}$$

Similarly, we can test for the effects of A and B using the new $SS(error)$ term. For instance, for $H_{0A} : \alpha_i = 0$, the test statistic under null is

$$F = \frac{SS(A)/(a-1)}{SS(error)/(ab(c-1))} \sim F_{a-1,ab(c-1)}$$

The two-way model with interaction anova table is

Source	df	SS	MS	F
Factor A	$a-1$	$SS(A)$	$SS(A)/(a-1)$	$MS(A)/MS(error)$
Factor B	$b-1$	$SS(B)$	$SS(B)/(b-1)$	$MS(B)/MS(error)$
Factor AB	$(a-1)(b-1)$	$SS(AB)$	$SS(AB)/((a-1)(b-1))$	$MS(AB)/MS(error)$
Error	$ab(c-1)$	$SS(error)$	$SS(error)/(ab(c-1))$	
Total	$abc-1$	$SS(total)$		

The df are intuitive for

- Factor AB has ab values, but due to regulatory constraints, we remove one row and one column
- The total df of the factor A, factor B, factor AB adds to an entire grid ab, hence the error df is the overall observations subtract one grid creating $ab(c-1)$
- The total df is the sum of the dfs of factor A, factor B, factor AB, and one for the grand mean.

In R, to indicate interaction in the two-way anova, we write `factor(A) * factor(B)`.

13.4 Multiple comparisons

If H_0 is rejected, we need to start a detailed analysis of the groups to see how the null failed. The intuitive method is to conduct multiple pairwise comparisons of the group means, but this grows quickly with the number of groups k.

In general, we will not focus on calculations here.

13.4.1 Data snooping

Data snooping is analyzing interesting differences that appear after seeing the data.

If we pre-planned before the dataset, we can construct a $100(1-\alpha)$ CI for the difference between two group means by using the pooled anova model variance estimator (using mean squared error) and it would be accurate. But if we first rejected H_0 then calculated the same CI between the two most significantly different group means, then the CL would be misleading (misleadingly high).

A solution is to use a pre-planned multiple comparison procedure that performs all possible comparisons with an overall controlled error rate. Another solution is to formulate a hypothesis given this data (maybe on a pairwise mean difference) and test it on a new independent experiment. We will focus on the first one.

13.4.2 Error rates

Define the overall/family error rate of c comparison CIs at level α (so individual error rate is α) as the probability that one or more will indicate a significant difference given H_0. Assuming each comparison is approximately independent, the overall family error rate is

$$1-(1-\alpha)^c$$

The family error rate is always higher than individual error rates, so we cannot use the typical 95% CIs for the significant differences are not significant over the c comparisons.

13.4.3 LSD test

The fisher's least significant difference test preplans all pairwise comparisons (or a subset if theory suggests). Its parameters are the pairs to test and the individual α level, and we often select α by aiming to get the family error rate below a certain threshold.

It calculates the least significant difference (smallest difference that is significant) between means required for a significant CI, which is basically the margin of error of the CI. This LSD will vary between pairs if the experiment is not balanced, but it will be a single number if it is balanced.

It then compares all mean differences that are preplanned against the LSD, and clusters the groups with similar (difference below LSD) means. If the design is not balanced, check the CI for every planned pair instead.

We would need to decrease the individual α to keep the overall family error rate under control.

13.4.4 Bonferroni correction

The bonferroni correction for the LSD test makes a suggestion about the individual CI level needed to achieve a family error rate of α. It states that if there are m comparisons, the individual level should be at most

$$\alpha/m$$

if the family error rate is to be at most α.

To prove this, consider A_i to be that the ith CI is in error, then suppose we use the Bonferroni correction so that $P(A_i) = \alpha/m$, we have

$$\begin{aligned} P(\bigcup_i A_i) &\leq \sum_i P(A_i) \\ &= m\frac{\alpha}{m} = \alpha \end{aligned}$$

by Boole's inequality in probability (this inequality is intuitive for equality is only held when A_i are disjoint, and non-disjoint sets only increases the RHS). This shows that the probability that at least one CI is in error is at most α if we use Bonferroni's correction.

This correction is conservative and increases probability of type II errors. It therefore reduces statistical power of detecting actual differences.

13.4.5 Tukey's HSD

Tukey's honestly significant difference also controls the family error rate for all pairwise comparisons.

It treats the comparisons as h-test and models the test statistic on the difference of sample means as coming from a Studentized Range Distribution (the distribution of the sample range when sampling from a normal distribution, and also the distribution of the difference between the minimum and maximum sample means).

We declare each mean pair as significant if

$$\frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{MSE}{n}}} \geq W$$

where $\bar{X}_i > \bar{X}_j$ are sample means of the two groups, MSE is the mean squared error variance estimator, n is the sample size within each group, and W is a quantile from the Studentized Range distribution determined from the family error rate parameter α.

The theory states that if we perform this h-test on all mean pairs, we get a total family error rate of α.

13.4.6 Duncan's new multiple range test

Similar to Tukey's HSD, it also uses the Standardized Range Distribution. It doesn't use a family or individual error rate, instead, the sample means are ranked in order and the parameter α determines the error rate for each pairwise comparison in addition to the number of steps the means are apart.

If two groups are r steps apart after ranking, we set the significance parameter for the Standardized Range Distribution quantile as $(1-\alpha)^{r-1}$.

Duncan's test has higher power than Tukey's test.

13.5 ANOVA connection to Regression

As ANOVA breaks down the total variability into the explained model variability SS and the unexplained error variability SS, just like a regression model, we can analyze a regression model both using both regression and anova.

In the linear regression model, we are mostly interested in testing $H_0 : \beta = 0$, and rejecting it will suggest sufficient evidences of a linear relationship between the mean response and x. The test statistic we use is a t-distribution (of $n-2$ df) on the slope estimator.

For the ANOVA model, recall the variance decomposition

$$\begin{aligned}\sum_i (Y_i - \hat{Y})^2 &= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 \\ SS(total) &= SS(error) + SS(regression)\end{aligned}$$

but also define the mean squared using the (given) degrees of freedom

$$\begin{aligned}MS(error) &= SS(error)/(n-2) \\ &= \hat{\sigma}^2 \\ MS(regression) &= SS(regression)/1\end{aligned}$$

then by trusting that $SS(residual)$ and $SS(error)$ are both scaled chi-squared and independent under $H_0 : \beta = 0$, define the test-statistic

$$F = \frac{SS(regression)/1}{SS(error)/(n-2)} = \frac{MS(regression)}{MS(error)} \sim F_{1,n-2}$$

This can be summarized into an ANOVA table

Source	df	SS	MS	F
Model	1	$SS(regression)$	$SS(regression)/1$	$MS(regression)/MS(error)$
Error	$n-2$	$SS(error)$	$SS(error)/(n-2)$	
Total	$n-1$	$SS(total)$		

Then the ANOVA over the regression problem will test for the significance of the regression, namely, the hypothesis that $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$.

Specifically, the anova model uses the F test statistic to test if $H_0 : \beta = 0$ using $F_{1,n-2}$. The regression model uses the T test statistic to test if $H_0 : \beta = 0$ using t_{n-2}.

Both the regression model and the ANOVA model will yield the same p-value for the h-test. The relationship between the test statistics are

$$t_k^2 = F_{1,k}$$

and the test-statistics will be equivalent after standardizing. The reasons that the distributions are equivalent is because that

$$\begin{aligned} t_k^2 &= \frac{N(0,1)^2}{\chi^2(k)/k} \\ &= \frac{\chi^2(1)/1}{\chi^2(k)/k} \\ &= F_{1,k} \end{aligned}$$

We can also prove that the test statistics are equal using the raw definitions of F and T, substituting the expression $SS(error) = SS(treatment) - \dots$ from tutorial 8.

14 Distribution Free Methods

Classical parametric methods are limited by their various assumptions, and sometimes they will not be satisfied, requiring an alternative test that doesn't assume a distribution.

Non-parametric, or distribution free methods has generally a set of weak assumptions of the underlying distribution for their validity, sometimes not needing to assume a specific parent distribution (hence distribution-free test).

These methods work with statistics whose sampling distributions are derived from weak assumptions, thus requiring a simplified view of the data. We can then use these statistics as h-test. Interestingly, these test-statistics are often approximately normal due to the CLT.

14.1 Location Test

In general, specifying a distribution implies specifying a scale of measurement. There are two strategies to compare numbers without a scale: signs that record if a number is greater than a reference, or ranks that retain only the ordering information. Both forgoes some information.

Location test is a test about the mean or median of the population. If there are extreme skewness or outliers, parametric based mean tests are unreliable so we wish for a median test (for it is a more robust measure of central tendency). The distribution-free test for the median asks if the population median is a specific number with few assumptions.

14.1.1 Sign Test

Given a random sample on X with size n on a population with median m. Wish to test $H_0 : m = m_0$ against an alternative that is one or two sided.

First assume that X is continuous, as to remove ties between observations and m_0.

If H_0 is true, then the proportion of observations above or below the median are binomially distributed

$$N_{\text{below}} = \#X_i < m_0 \sim Bi(n, 0.5)$$

And for either type of alternative hypothesis, we can choose either N_{below} or N_{above} as the test-statistic. Be careful about the critical region.

Alternatively, we can define $Y = N_{\text{above}}$ as the number of positive numbers in $X_i - m_0$, which explicitly shows that we are transforming the dataset to $\text{sgn}(X_i - m_0)$. This transformation that reduces the sample to just their signs gives the sign test its name.

The sign test test-statistic has a binomial distribution with $p = 0.5$, so we use the exact binomial tests when defining the critical regions and the p-values. We can use randomized rejection region to create an exact type I error, by having a randomized chance to reject at specific critical values and have a certain chance to reject at other specific regions.

For paired samples where we are testing for $H_0 : m_{X-Y} = 0$ or equivalently $H_0 : P(X > Y) = 0.5$, simply replace the observations (x_i, y_i) with $x_i - y_i$ and set $m_0 = 0$ to net the signs $\text{sgn}(x_i - y_i)$.

The sign test has little assumptions (only needing continuous data) and thus is widely applicable. Its limitations are that it doesn't use information on the magnitude of the differences, so it has smaller power for it can be insensitive about big departures/differences from H_0.

14.1.2 Wilcoxon signed rank test

The signed rank test for one-sample accounts for the size of the differences. It assumes that the underlying distribution is symmetrical about the median and is continuous.

Intuitively, the positive differences and the negative differences should be symmetrically spaced out around the median. And if we rank the absolute differences and compare the rank sum of the between the positive and negative signs, their differences should be around zero.

Formally, consider $H_0 : m = m_0$ against a one or two sided alternative. Given a random sample X_i, the Wilcoxon signed rank test is

1. First rank the $|X_i - m_0|$ into $\text{rank}(|X_i - m_0|)$
2. Then transform the dataset into the signed ranks $\text{sgn}(X_i - m_0) \times \text{rank}(|X_i - m_0|)$
3. Define the wilcoxon signed rank statistic as the sum of all the signed ranks

$$W = \sum_i \text{sgn}(X_i - m_0)\,\text{rank}(|X_i - m_0|)$$

Under H_0, the conditional probability of the sign being positive or negative given the absolute difference (or given the rank) is 50-50, for the continuity and symmetry of X. These signs are also mutually independent for X_i are independent. As W is always the weighted sum of the ranks $1 \dots n$ with positive or negative sign weights (no ties due to continuous X), under H_0, it is the sum of independent rvs $\sum_i W_i$ where each W_i is 50-50 a positive or negative rank i

$$P(W_i = i) = P(\text{sgn}(X_i - m_0) = 1|\,\text{rank}(|X_i - m_0|) = i) = P(W_i = -i) = 0.5$$

This leads to an exact distribution of W under H_0, but without a nice looking expression. However, we can use computers to perform hypothesis testing with an exact significance level.

To approximate W's sampling distribution, note that under H_0

$$\begin{aligned} E[W_i] &= -i/2 + i/2 = 0 \\ V[W_i] &= E[W_i^2] = i^2 \\ E[W] &= 0 \\ V[W] &= \sum_i i^2 \\ &= \frac{n(n+1)(2n+1)}{6} \end{aligned}$$

and by the CLT that allows different but bounded variances for its summation terms, under a large n, the standardized W will follow a standard normal

$$Z = \frac{W - 0}{\sqrt{n(n+1)(2n+1)/6}} \sim N(0,1)$$

which can be used as a test-statistic or to compute approximate critical values and p-values.

Be careful of the direction of the critical region.

Note that $W = W^+ - W^-$, where W^+ are the summed positive difference ranks and W^- are the summed negative difference ranks. Note the relationship $n(n+1)/2 - W^+ = W^-$ due to the sum of ranks. R uses W^+ as its test-statistic which is just the sum of the positive difference ranks, calling it V, but there is a one-to-one transformation between W and V with similar sampling distributions, so the tests are identical. Also note that

$$E[V] = n(n+1)/4 \quad V[V] = n(n+1)(2n+1)/24$$

The R function `wilcox.test(x)` uses the exact sampling distribution of V for small samples and the normal approximation for large samples. We can also work with the exact sampling distribution for V using `psignrank` (or the other functions related to the signrank distribution). Note that the signrank distribution is discrete.

For paired samples with observations (x_i, y_i), we use the dataset $x_i - y_i$ and treat this as a sample from the distribution $X - Y$ with $m_0 = 0$. This will test for $H_0 : m_{X-Y} = 0$. Note that if we assume that X and Y has the same distribution under H_0, then we meet the symmetrical assumption of the Wilcoxon signed rank test for $X - Y \sim Y - X$ and is a symmetric distribution around the H_0 median 0. Often, this test is used under this paired samples setting for the plausibility of the equal distribution assumption.

If the distribution of X is discrete due to finite precision, we may have ties in the ranking. In such cases, we assign rank values for the tied entries as the average of the ranks they span (a tied rank 4 and 5 results in both differences assigned a rank 4.5), then use the continuous method with its ed test statistic sampling distribution. The sampling distribution derivation of W or V under ties is more complicated because it may be fractional, but R will still get the right answer. We can also just keep the unique differences and throw the equal differences away.

14.1.3 Wilcoxon rank sum test

In general to compare two populations, we have independent random samples X_i and Y_i from two continuous populations with sample size n_X and n_Y.

Suppose that X and Y are identically distributed under H_0, then if we combine both samples into a single combined sample and rank the observations, the subset of ranks associated with samples from either population will be a subset uniformly chosen from the $\binom{n_X+n_Y}{n_X}$ or $\binom{n_X+n_Y}{n_Y}$ options, and thus we can set the test-statistic as the sum of the ranks for each population sample and get its sampling distribution.

There are many ways to rank the combined sample, each corresponds to the alternative we want to test

- If the alternative is to test a shifted location, we rank from the smallest to the largest and anticipate that the rank-sum in one group is too large or too small (Wilcoxon rank sum test)

- If the alternative is to test the variability of one population over another, we rank the smallest rank 1, then the largest rank 2, then the second smallest rank 3, etc. This will net a large rank sum for the less variant population (Seigel Tukey test)
- Note that the test statistic $D = \sup_z |\hat{F}_X(z) - \hat{F}_Y(z)|$ is also useful for a lot of different h-tests, specifically the largest cdf difference for a normality test. It is also a function of the ranks and is thus a rank test.

We only care about the first location test.

For the wilcoxon rank sum test, consider independent samples X_i and Y_i, where we assume that $Y \sim X + \Delta$ for some fixed real number Δ. The null is $H_0 : X \sim Y$ or $H_0 : \Delta = 0$ against a one or two sided alternative where Δ is not zero or greater or less. This null hypothesis is equivalent to the median test that $H_0 : m_X = m_Y$ and the alternative is an inequality of the medians.

The process is

1. Order the combined sample, rank them from smallest to largest
2. Let W_X be the rank sum of the X observations and W_Y as the rank sum of the Y observations
3. Either group's rank sum can be used to test H_0. The test statistic W_X and W_Y are called the Wilcoxon rank sum test statistic

To get the critical region, think about what happens to the test-statistic if H_1 is true.

The exact sampling distribution of W_Y under H_0 for small sample sizes can be found by enumerating the possible outcomes (summation of a subset) from all the subsets. To approximate the sampling distribution for a large sample size, we are given

$$\begin{aligned} E[W_Y] &= \frac{n_Y(n_X + n_Y + 1)}{2} \\ V[W_Y] &= \frac{n_X n_Y(n_X + n_Y + 1)}{12} \end{aligned}$$

and that W_Y is normally distributed when both n_X and n_Y are large. To derive the formula for the mean, we intuitively take the total sum of the ranks, and assign a proportion of them weighted by $n_X/(n_X + n_Y)$ to W_X's mean.

14.1.4 Mann-Whitney test

The mann-whitney test is an equivalent form of the rank sum test. Its test statistic U has two forms dependent on the group that we focus on (like the Wilcoxon's test), with W_{XY} counting the total number of pairs (X_i, Y_j) where $X_i < Y_j$ and W_{YX} the other way.

The mann whitney test statistics are equivalent to the wilcoxon test statistic by

$$W_{XY} = W_Y - \frac{1}{2}n_Y(n_Y + 1)$$
$$W_{YX} = W_X - \frac{1}{2}n_X(n_X + 1)$$

and note that these mann whitney test statistics are just shifted W_Y and W_X, namely subtracting their minimum values so the smallest value is zero.

R's `wilcox.test(x,y)` actually computes the mann-whitney $U = W_{YX}$ test statistic instead of W_X, but is reported as the symbol W. For small sample sizes, it will use the exact sampling distribution. Otherwise, it uses a normal approximation. We can also get the exact sampling distribution of U by `pwilcox` (and use `wilcox` for the general wilcox distribution). Under H_0, we are also given

$$E[W_{YX}] = \frac{n_X n_Y}{2}$$
$$V[W_{YX}] = V[W_X] = \frac{n_X n_Y(n_X + n_Y + 1)}{12}$$

14.2 Goodness of fit Tests

In general, a goodness of fit tests how well a model fits a data sample. We will talk about Pearson's chi-squared test.

The pearson's chi-squared test operates on categorical or discrete data, but we can also apply it on continuous data by partitioning it into separate classes.

14.2.1 Binomial intuition special case

The binomial model to test for proportions is a specialized pearson's chi-squared test.

Consider $H_0 : p = p_1$ against $H_1 : p \neq p_1$ where $Y_1 \sim Bi(n, p_1)$ under H_0 which are the number of successes and success probability. The normal approximated test statistic under H_0 can be written as

$$Z = \frac{Y_1 - np_1}{\sqrt{np_1(1 - p_1)}} \sim N(0, 1)$$
$$Q_1 = Z^2 \sim \chi^2(1)$$

and we reject H_0 if $|Z| \geq c$ or $Q_1 \geq c^2$.

Notice that

$$\begin{aligned}Q_1 &= \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} \\ &= \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_1 - np_1)^2}{n(1-p_1)} \\ &= \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2}\end{aligned}$$

where $Y_2 = n - Y_1$ and $p_2 = 1 - p_1$. Intuitively, Y_1 is the observed number of successes, np_1 is the expected number of successes, ditto with Y_2 with the number of failures. So

$$Q_1 = \sum_{i=1}^{2} \frac{(Y_i - np_i)^2}{np_i} = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(1)$$

where O_i is the observed number for the category i, and E_i is its expected number. Note that we lose one degree of freedom in the chi-squared due to the constraint that $Y_1 + Y_2 = n$.

14.2.2 Multinomial model

To generalize to k categories, let p_i be the probability of the ith category. Let n be the number of samples, and Y_i be the number of samples with class i. Then $E[Y_i] = np_i$.

In the derivation of the chi-squared distribution, we assume the approximation of binomial with normal. This is only valid if all $E_i = np_i \geq 5$.

The test-statistic is therefore

$$Q_{k-1} = \sum_{i} \frac{(Y_i - np_i)^2}{np_i} = \sum_{i} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(k-1)$$

where we lose one df due to the constraint $\sum_i Y_i = n$. We will not need to know how the test-statistic sampling distribution is derived.

The process for the gof test is

1. Specify a categorical distribution $p_1 \ldots p_k$
2. Let null be that the p_is are all correct, and the alternative is that a different set of p_i defines the distribution
3. Compute test statistic Q_{k-1} with critical region $Q_{k-1} > c$ to see if data is consistent with the distribution
4. Under null, test statistic will be small and it measures the badness of fit. Hence, reject null if $Q_{k-1} > c$ where c is the $1 - \alpha$ quantile from $\chi^2(k-1)$.

The test-statistic Q_{k-1} is often called the Pearson's χ^2 statistic.

Note that R treat the residuals of the test as

$$R_i = \sqrt{\frac{(O_i - E_i)^2}{E_i}}$$

If we've rejected H_0, we can use the size of the residuals to find the cells that are abnormal.

14.2.3 Fitting distributions

Sometimes we are not given an exact set of p_i, but might specify a family of distributions with the parameters estimated from the sample.

Our H_0 would then be if the data has this distribution family (don't say the parameter), and we need to adjust the df for we've used the data to define H_0 using the parameter estimator (for otherwise, our distribution will seem closer to the data). We subtract 1 df for each estimated parameter, so the final df is $k-p-1$ where p is the number of estimated parameters.

Then to perform the test

1. Partition the distribution and partition the observations using the same partitions. We can use 0.5s if the distribution is discrete. The partition should cover the entire domain of the pdf/pmf
2. The larger the k, the more powerful the test is for the counts are more fine-grained, but E_i should be large enough for each range
3. Compute p_i for the partition ranges using the estimated distribution
4. Perform GoF test with the expected counts using p_i and the observed counts from the partition, make sure to subtract df for the estimated parameters!

To test GoF in R, use `chisq.test(x)`. We must manually adjust the df for the parameters estimated.

14.3 Contingency tables

For multiple categorical variables (or partitions of continuous variables), a contingency table records the number of observations for each cross-classification of these categorical variables. We assume that there are only two categorical variables, so the contingency table is 2D.

Let A_i be the classes for variable A, and B_i be the classes for variable B. Assume each sample is assigned to a single combination (A_i, B_j), then the n samples can be summarized in a contingency table of counts. A model for these cross-classified data is $p_{ij} = P(A_i \cap B_j)$ for the probability of the sample being in the ith class of A and the jth class of B.

To test if the two variables are independent, we use the null $H_0 : p_{ij} = P(A_i)P(B_j)$ for all p_{ij} against the alternative where at least one $p_{ij} \neq P(A_i)P(B_j)$. This has the same structure as the GoF test, so we reuse the Pearson's chi-squared statistic.

Model the population as a discrete bivariate distribution, then $P(A_i) = p_{i\cdot}$ and $P(B_j) = p_{\cdot j}$ are just the marginal pmf for the categorical variables, so the independence null is $H_0 : p_{ij} = p_{i\cdot} \times p_{\cdot j}$. But we don't have the marginal pmfs, so we must estimate them from the sample.

The steps for the test are

1. Put the observed counts into a contingency table
2. Estimate the marginal pmfs $\hat{p}_{i\cdot} = \sum_j y_{ij}/n$ and $\hat{p}_{\cdot j} = \sum_i y_{ij}/n$ from the sample, compute the joint pmf under H_0 by $\hat{p}_{ij} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = \frac{Y_{i\cdot}Y_{\cdot j}}{n^2}$
3. If we have all the p_{ij}, the Pearson's chi-squared statistic is

   $$Q = \sum_i \sum_j \frac{(Y_{ij} - np_{ij})^2}{np_{ij}}$$

   but we estimated the p_{ij} by $\hat{p}_{ij}$, so the test statistic is

   $$Q = \sum_i \sum_j \frac{(Y_{ij} - Y_{i\cdot}Y_{\cdot j}/n)^2}{Y_{i\cdot}Y_{\cdot j}/n} \sim \chi^2((r-1)(c-1))$$

4. Use test-statistic to reject or accept H_0 based on $Q > c$ where c is the quantile from χ^2.

The df is because we've estimated $r - 1$ marginal probabilities for A and $c - 1$ marginal probabilities for B to get the joint pmf under H_0, so we must subtract those. Here, $k = rc$ for the total number of classes, $p = r - 1 + c - 1$ for the estimated marginals, so $df = k - p - 1 = (r - 1)(c - 1)$.

Note that we've also approximated binomial distributions as normal, so we need the estimated expected counts for each cell to be greater than or equal to five. Otherwise, merge the rows or columns.

To test this in R, use `chisq.test(table, correct=FALSE)`.

15 Bayesian Methods

15.1 Probability review

Recall that for events A and B (which could be in terms of rv taking a value), we have concepts of

- Joint probability $P(A, B) = P(A \cap B)$.
- Marginal probability, $P(A) = P(A, B) + P(A, \bar{B})$.
- Conditional probability, $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Remember bayes' theorem

$$\begin{aligned} P(B|A) &= \frac{P(A|B)P(B)}{P(A)} \\ &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})} \end{aligned}$$

Recall that a partition B_i of the sample space is a set of events which are mutually exclusive but cover the sample space. The law of total probability relates marginal probability with conditional probability

$$P(A) = \sum_i P(A, B_i) = \sum_i P(A|B_i)P(B_i)$$

Hence, the expanded bayes' theorem is

$$\begin{aligned} P(B_i|A) &= \frac{P(A|B_i)P(B_i)}{\sum_i P(A|B_i)P(B_i)} \\ &\propto P(A|B_i)P(B_i) \end{aligned}$$

because the denom $P(A)$ is just a constant relative to i in the B_i.

Now consider continuous events through continuous density functions, using pdf on X and Y. The analogous definitions are

- Joint pdf, $f(x, y)$
- Marginal pdf and law of total probability

$$f(x) = \int f(x, y)\, dy = \int f(x|y)f(y)\, dy$$

- Conditional pdf, $f(x|y) = f(x, y)/f(y)$.
- Bayes' theorem, $f(x|y) = \frac{f(y|x)f(x)}{f(y)}$

15.2 Interpretations of probability

The frequentist method for inferences assumes the parameter for the hypothetical distribution exists. Bayesian statistics updates a prior given the sample information.

Probability has two interpretations

- Frequentist probability, models variation
- Bayesian probability, models uncertainty

Classical inference (everything we've learnt so far) only uses frequentist probability of modeling variations. Bayesian inference uses both frequentist and bayesian probability.

Under frequentist probability (aleatory, physical, frequency probability), probability is defined as the relative frequency of occurrence of an event in the long run, under hypothetical repetitions of an experiment. A statistical model implicitly assumes this interpretation of probability, for it contains a probability distribution that specifies a model of variation across multiple samples of the data.

The frequentist probability needs a well-defined experiment that can be repeated. Its interpretations of one-off events, or those already occurred, is problematic.

Bayesian probability (epistemic, evidential probability) defines probability as the degree of plausibility, or strength of belief, of a statement based on existing knowledge and evidence. This allows bayesian probability to be assigned to any statement, even those with no random process, and those where the event has already occurred. Intuitively, bayesian probability is the odds that people will accept if they are forced to bet on the outcome.

Note that interpretations of probability are different to its axiomatic definitions. Under mathematical probability, it is not obvious that the frequentist long run frequency exists and is equal to the underlying probability: this needs to be proven by the law of large numbers.

Reasons for bayesian probability

- More natural in interpretation
- Directly answer questions of interest
- Beyond a true/false answer, an extension of formal logic that allows reasons under uncertainties

15.3 Bayesian inference

For bayesian inference, we take our existing statistical models and model the parameters and hypotheses as random variables and probabilities.

This implies that parameters will have probability distributions, and hypotheses will have probabilities. These are Bayesian probabilities. Both the parameters and hypotheses probabilities quantify our uncertainty before and after seeing the data.

The posterior (distribution) is the probability of our parameters or hypotheses given the data. It quantifies our knowledge in light of the observed data. In Bayesian inference, the posterior distribution summarizes all the information about the parameters of interest. In formula, it is $P(\theta|X)$.

To compute the posterior, we need to specify

- The likelihood $P(X|\theta)$ which is the probability of observing the data under the parameter value
- The prior distribution $P(\theta)$ of the parameter

The prior is needed to compute the posterior. It represents the initial probability distribution of the parameter before seeing the data. We often need be careful in selecting an appropriate prior.

For discrete distributions of the parameter, say a prior of $P(\theta)$ and a likelihood $P(X|\theta)$, we use the discrete bayes' theorem to compute the posterior

$$\begin{aligned} P(\theta|X=x) &= \frac{P(X=x|\theta)P(\theta)}{P(X=x)} \\ &\propto P(X=x|\theta)P(\theta) \end{aligned}$$

where we ignore the denominator for it is constant relative to θ. Additionally, we can also ignore any factor that doesn't involve the parameter θ, for we can simply normalize the conditional/posterior distribution at the end.

For continuous priors $f(\theta)$ and the continuous likelihood $f(x|\theta)$, we use the continuous bayes' theorem

$$f(\theta|X=x) \propto f(X=x|\theta)f(\theta)$$

and we once against ignore any constant factors and aim to normalize all at the end by integrating against θ

$$K = \int f(X=x|\theta)f(\theta)\,d\theta$$

Note that in the continuous case, $f(x|\theta)$ is simply the likelihood function $L(\theta)$.

To make inferences using the posterior distribution, we can use its mean, mode, or median as a point estimate.

Define conjugacy by: a class of prior distributions on a parameter for a family of likelihood distributions using the parameter defines a conjugate family of distributions if the posterior distribution of the parameter is the same family as the prior distribution.

We can interpret priors for the parameters as unobserved (already observed) pseudo-data, with the same influence on the posterior as an actual sample with some sample size and particular pseudo observations. This provides an intuitive interpretation for the prior, particular if it is a conjugate prior. For example, a uniform prior B$(1,1)$ for the proportion is equivalent to observing one success and one failure (pseudocounts of success and failures), and a B(a, b) prior is equivalent to observing $a+b$ samples with the pseudocounts of a successes and b failures. Note that pseudocounts can be non-integers.

Remarks about Bayesian inference

- Often, we call the likelihood as the model or the model for the data. Note that we can also refer to the whole setup of prior and likelihood as the model.
- Classical inference only requires a likelihood, but is more restrictive about doing inference
- While parameters are modeled as random variables to express their uncertainties, we don't actually think of them as being random quantities. We still think of them as fixed underlying quantities but ones we can never observe for certain

15.4 Example Prior Posteriors

For a standard uniform prior $P(\theta) = 1$ and a binomial likelihood $X \sim Bi(n, \theta)$, the posterior is beta distributed with

$$\theta | X = x \sim \text{Beta}(x+1, n-x+1)$$

And if the prior is replaced with a generalized beta distribution $\theta \sim \text{Beta}(a, b)$, the posterior is

$$\begin{aligned} f(\theta | X = x) &\propto f(X = x | \theta) f(\theta) \\ &\propto \theta^x (1-\theta)^{n-x} \theta^{a-1} (1-\theta)^{b-1} \\ \theta | X = x &\sim \text{Beta}(x+a, n-x+b) \end{aligned}$$

Intuitively, a beta prior plus a binomial likelihood results in the same family beta posterior, therefore the beta distribution is the conjugate prior for the binomial distribution.

Reminder that the beta distribution is a distribution over the unit interval $[0, 1]$, with two parameters $\alpha, \beta > 0$, and for $X \sim \text{Beta}(\alpha, \beta)$, the pdf is

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)}$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the beta function. The properties of the beta distribution include

$$\begin{aligned} E[X] &= \frac{a}{a+b} \\ \text{mode}(X) &= \frac{a-1}{a+b-2} \quad a, b > 2 \\ V[X] &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

15.5 Summarizing Posteriors

To interpret posteriors, we can calculate point summaries/estimates such as

- Posterior mean, $E[\theta|X = x]$
- Posterior median, $m_{\theta|X=x}$
- Posterior mode, $\text{mode}(\theta|X = x)$

If we have a uniform prior, the posterior mode is the MLE for the parameter. (This is because the mode of the likelihood is just the MLE)

The posterior standard deviation $\text{sd}(\theta|X = x)$ gives a measure of the uncertainty in the parameter like the standard error. The estimate that is mean of the posterior distribution minimizes the average squared error (MSE). Similarly, to minimize the absolute error, we report the median of the posterior distribution.

Interval estimates on the posterior refers to simple probability intervals on the posterior, also known as credible intervals to indicate uncertainty. A 95% credible interval (a, b) is defined as

$$0.95 = P(a < \theta < b|X = x)$$

and a central credible interval has symmetrical tails.

Credible intervals are analogous to confidence intervals, but are easier to interpret and explain. They can also be one or two-sided.

Visual summaries are the plotting of the posterior distribution.

Posterior probabilities of events (which depends on the parameter) can also be calculated given the distribution. More generally, we can calculate a posterior distribution of any function of the parameters using the cdf method and the transformation of the parameter posterior distribution.

Note that we often cannot derive or write an analytic expression for the posterior, for all modern applications of Bayesian analysis. We instead rely on computational techniques that work with simulations/samples from the posterior. The most common method to compute posteriors computationally is Markov Chain Monte Carlo methods.

15.6 Further Bayesian Inference

For this subject, we only consider single parameter models, and conjugate priors. The examples would mimic the classical inference we've done before.

In general, we can factor a likelihood function into a function on the parameter and sufficient statistic $g(y|\theta)$ multiplied by a function on the sample. Because we will drop all functions not including θ, we can simply just use $g(y|\theta)$ which is the conditional pdf of the sufficient statistic Y given the parameter θ.

This essentially implies that we can represent observing a sample as observing its sufficient statistic.

15.6.1 Normal, single mean, known variance

For a random sample where $X_i \sim N(\theta, \sigma^2)$, summarize the data by $Y = \bar{X} \sim N(\theta, \sigma^2/n)$.

Define the prior as $\theta \sim N(\mu_0, \sigma_0^2)$, the posterior is

$$\begin{aligned} f(\theta|y) &\propto f(y|\theta)f(\theta) \\ &\propto \exp(-\frac{1}{2\sigma^2/n}(y-\theta)^2)\exp(-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2) \\ &= \exp(-\frac{(y-\theta)^2}{2\sigma^2/n} - \frac{(\theta-\mu_0)^2}{2\sigma_0^2}) \\ &= \exp(-\frac{(\theta-\mu_1)^2}{2\sigma_1^2}) \end{aligned}$$

where we've defined

$$\begin{aligned} \mu_1 &= \frac{\mu_0/\sigma_0^2 + y/(\sigma^2/n)}{1/\sigma_0^2 + 1/(\sigma^2/n)} \\ \frac{1}{\sigma_1^2} &= \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n} \end{aligned}$$

the steps are in the appendix.

The posterior is therefore a normal pdf so we can substitute the normalization constant and get the posterior distribution

$$\theta|y \sim N(\mu_1, \sigma_1^2)$$

Define the precision of a random variable X as $1/V[X]$. Then

- The posterior distribution of the population mean has a precision that is the sum of the prior and data precision.
- The posterior distribution mean is a weighted average between the sample mean y and the prior mean μ_0 by their precisions.

This implies that more data leads to higher data precision and higher precision for the posterior.

The credible intervals on the population mean is a probability interval on the normal posterior. The central credible interval on θ will have the form of

$$\mu_1 \pm c\sigma_1$$

where c is a quantile from the standard normal.

15.6.2 Improper prior

For the single mean, known variable example. We can make the prior less informative by reducing its precision $\sigma_0 \to \infty$ (this can be done with a normal prior or any other distribution).

At the limit, we will get a flat prior across the whole real number line and μ_0 essentially disappears. While this is not a valid probability distribution as it cannot integrate to one, it does indeed work and gives a valid posterior where

$$\mu_1 = y \qquad \sigma_1^2 = \sigma^2/n$$

and we note that the credible interval is exactly the confidence interval.

This type of prior where it is not a valid probability distribution of θ at the limit is called an improper prior. If the prior can integrate to one, it is called a proper prior.

Intuitively, improper priors are approximations to very uninformative priors.

15.6.3 Single proportions

Suppose that the random sample of bernoulli trials have $X_i \sim Bi(1, \theta)$. Using the prior $\theta \sim \text{Beta}(a, b)$ and the data model $X \sim Bi(n, \theta)$, the posterior is

$$\theta | x \sim \text{Beta}(a + x, b + n - x)$$

The posterior mean is

$$\begin{aligned} E[\theta|x] &= \frac{a+x}{a+b+n} \\ &= \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{x}{n} \end{aligned}$$

which is a sample size weighted average between the prior mean and the data MLE.

We can also find the posterior probability of a majority by

$$P(\theta > 0.5|x)$$

If we have an initial survey with a suggested probability p_0 and the knowledge is worth an equivalent pseudo sample size of n_0, we can include this as a prior by setting a, b such that

$$a + b = n_0 \qquad E[\theta] = \frac{a}{a+b} = p_0$$

15.6.4 Exponential distribution

For a random sample $X_i \sim \exp(\lambda)$ of size n, with the data summarized as $Y = \sum X_i \sim \gamma(n, \lambda)$, we have the conjugate prior which is a gamma distribution

$$\lambda \sim \gamma(r, \lambda_0)$$

and the posterior

$$\lambda|y \sim \gamma(n + r, \lambda + \lambda_0)$$

The posterior mean is

$$E[\lambda|y] = \frac{n + r}{\lambda + \lambda_0}$$

Note that we also could've used the likelihood

$$f(\lambda|x_i) \propto (\sum_i f(x_i|\lambda))f(\lambda)$$

but we when can factorize the likelihood and see that the sufficient statistic for λ is Y, so the results are the same.

15.6.5 Boundary problem

Given random samples from the shifted exponential with pdf

$$f(x) = e^{-(x-\theta)}1_{x>\theta}$$

or that

$$X \sim \theta + \exp(1)$$

Consider a flat improper prior for θ

$$\theta \sim N(0, \sigma^2)$$

where $\sigma^2 \to \infty$. The posterior is

$$\begin{aligned}
f(\theta|x_i) &\propto (\prod_i e^{-(x_i-\theta)} 1_{x_i>\theta}) \exp(-\frac{1}{\sigma^2}x^2) \\
&= \prod_i e^{-(x_i-\theta)} 1_{x_i>\theta} \\
&= e^{n\theta - \sum_i x_i} 1_{x_{(1)}>\theta} \\
&\propto e^{n\theta} 1_{x_{(1)}>\theta}
\end{aligned}$$

by normalizing with k, we have

$$\begin{aligned}
I &= \int_{-\infty}^{x_{(1)}} ke^{n\theta}\, d\theta \\
&= \frac{k}{n} e^{nx_{(1)}} \\
k &= ne^{-nx_{(1)}} \\
f(\theta|x_i) &= ne^{n(\theta - x_{(1)})} 1_{\theta<x_{(1)}}
\end{aligned}$$

To find a one-sided credible interval with a lowerbound $(a, x_{(1)})$, we find the df and solve for a where $F(a) = 0.05$.

15.7 Prior Distributions (NOT EXAMINABLE)

To choose an appropriate prior, consider

- Existing knowledge, trying to quantify it
- Plausibility of values
- Ability for the data to overwhelm the prior, we should allow sufficient data to overwhelm the diffuse prior

Usually, the prior will be much less precise than the data. If the prior vastly conflicts with the data, recheck your assumptions for someone has likely gone wrong.

However, since we expect the data to dominate the prior, we don't need to be too worried with the exact prior shape.

A noninformative prior ideally has no influence on the posterior. This can sometimes be achieved by an improper prior like $N(\theta, \sigma^2), \sigma^2 \to \infty$. However, noninformative usually depends on the specific parameterization, and a noninformative prior on one scale can be informative on another scale/transformation of the same parameter. Therefore, we generally talk about diffuse priors (which lets the data dominate) rather than noninformative priors.

Sensitivity analysis focuses on trying a range of different priors to see how the posterior is influenced by the prior. It is useful to try typical diffuse priors and some reasonable set of extreme priors.

A sensitivity to the prior is a key feature of Bayesian inference, not a problem, as it alerts you the relative amount of information in your data. If the prior is influential, and we don't have good reasons to believe it, then we don't have sufficient data. We would then need to either collect more data or to use a more reliable prior.

15.8 Bayesian vs Classical inference (NOT EXAMINABLE)

Similarities are

- Learning about population
- Estimating parameters
- Making decisions
- Predictions

Differences

- Use of probabilities
- Method of inference
- Result interpretation

The assumptions of Bayesian inference is a prior. While this may seem like a weakness, it is usually overplayed as being overly subjective. Classical inference requires potentially more subjective choices like the estimators to use, which can be just as subjective as choosing a prior. Bayesian openly lets you make the prior assumptions, while classical hides them.

The choice of likelihood in Bayesian is also important, and involves similar consideration and problems as choosing a prior, but people often overlook this.

Complex models often blur the boundary between the two inference methods.

Advantages of Bayesian

- Forces you to consider your assumptions
- Provides a direct way of including useful prior knowledge
- Inference after knowing the prior is simple: calculate the posterior
- Easier interpretation for we answer the question using the parameter distribution directly

Disadvantages

- Need to write a full probability model for parameters, which is difficult for complex problems
- More computation
- Harder to use, requires more experience
- Focuses mostly on parametric models

Though the two methods are similar in that posterior summaries are equivalent to estimators. We can evaluate the characteristic of the posterior summary under repeated sampling, or to take an estimator and ask what prior it corresponds to. This is however not always possible in high dimensional models.

We should

- Learn both methods
- Use whatever works for the problem
- Classical techniques are generally used more often for convenience, familiarity, or convention, not necessarily because they are the best
- In simple settings, both approaches lead to similar procedures
- More important when the problems are more complex

We should avoid using Bayesian if

- A specific method is expected and a strong convention
- More familiar and proficient with non-Bayesian methods
- More computationally demanding for complex problems

The inference process under a Bayesian perspective is

- Try bayesian inference
- If impractical, use non-Bayesian methods
- Always consider the implicit assumptions that exists and that is made by classical methods, and the Bayesian model it is equivalent to

A Order Statistic pdf

Alternatively, consider that $P(X_{(k)} \approx x) = g_k(x)dx$. This is also the probability that we observe $k-1$ in $(-\infty, x - dx/2]$, 1 in $(x - dx/2, x + dx/2]$, and $n-k$ in $(x + dx/2, \infty)$. This is a trinomial distribution with the overall probability

$$\begin{aligned} g_k(x)dx &= \frac{n!}{(k-1)!1!(n-k)!} F(x)^{k-1} f(x)dx(1-F(x))^{n-k} \\ &= k\binom{n}{k} F(x)^{k-1} f(x)dx(1-F(x))^{n-k} \end{aligned}$$

and dividing both sides by dx yields the same formula.

By the same logic, the joint pdf of two order statistics is

$$f_{X(i),X(j)}(u,v) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} F(u)^{i-1} f_X(u)(F(v)-F(u))^{j-i-1}(1-F(v))^{n-j} f_X(v)$$

B Normal Prior

To show that

$$\exp(-\frac{(y-\theta)^2}{2\sigma^2/n} - \frac{(\theta-\mu_0)^2}{2\sigma_0^2}) \propto \exp(-\frac{(\theta-\mu_1)^2}{2\sigma_1^2})$$

We consider

$$\begin{aligned} \frac{(y-\theta)^2}{\sigma^2/n} + \frac{(\theta-\mu_0)^2}{\sigma_0^2} &= \frac{y^2}{\sigma^2/n} - 2\frac{y\theta}{\sigma^2/n} + \frac{\theta^2}{\sigma^2/n} + \frac{\theta^2}{\sigma_0^2} - 2\frac{\theta\mu_0}{\sigma_0^2} + \frac{\mu_0^2}{\sigma_0^2} \\ &= \frac{\theta^2}{\sigma_1^2} - 2\theta\frac{(y/(\sigma^2/n) + \mu_0/\sigma_0^2)\sigma_1^2}{\sigma_1^2} + K \\ &= \frac{\theta^2}{\sigma_1^2} - 2\frac{\theta\mu_1}{\sigma_1^2} + K \\ &= \frac{(\theta-\mu_1)^2}{\sigma_1^2} + K_2 \\ \exp(-\frac{(y-\theta)^2}{2\sigma^2/n} - \frac{(\theta-\mu_0)^2}{2\sigma_0^2}) &= \exp(-\frac{(\theta-\mu_1)^2}{2\sigma_1^2} + K_3) \\ &\propto \exp(-\frac{(\theta-\mu_1)^2}{2\sigma_1^2}) \end{aligned}$$