M40005 term 2 revision

Teddy Wu

May 2025

1 Useful inequalities

Expectation minimizes expected square deviation:

$$\mathrm{E}[(X-a)^2] \geq \mathrm{E}[(X-\mathrm{E}[X])^2]$$

Proof:

$$\begin{aligned}
\mathrm{E}[(X-a)^2] &= \mathrm{E}[((X-\mathrm{E}[X]) + (\mathrm{E}[X]-a))^2] \\
&= \mathrm{E}[(X-\mathrm{E}[X])^2] + 2\,\mathrm{E}[(X-\mathrm{E}[X])(\mathrm{E}[X]-a)] + \mathrm{E}[(\mathrm{E}[X]-a)^2] \\
&= \mathrm{E}[(X-\mathrm{E}[X])^2] + 2(\mathrm{E}[X]-a)\mathrm{E}[X-\mathrm{E}[X]] + \mathrm{E}[(\mathrm{E}[X]-a)^2] \\
&= \mathrm{E}[(X-\mathrm{E}[X])^2] + \mathrm{E}[(\mathrm{E}[X]-a)^2] \\
&\geq \mathrm{E}[(X-\mathrm{E}[X])^2]
\end{aligned}$$

Sample mean minimizes sum of square deviation:

$$\sum_{i=1}^{n}(x_i-a)^2 \geq \sum_{i=1}^{n}(x_i-\bar{x})^2$$

Similar theorems hold for median and sample median, except they minimizes deviation, not square deviation.

Upper bound for variance: Suppose a random variable X only takes values in $[a,b]$, then

$$\mathrm{Var}(X) \leq \frac{(b-a)^2}{4}$$

Markov's inequality: For nonnegative random variable X we have

$$P(X \geq a) \leq \frac{E[X]}{a}$$

Proof: Define the random variable

$$Y_a = \begin{cases} 0 & X < a \\ a & X \geq a \end{cases}$$

Since X is nonnegative we have $Y_a \leq X$, hence $\mathrm{E}[Y_a] \leq \mathrm{E}[X]$ which yields

$$aP(X \geq a) \leq \mathrm{E}[X]$$

Chebyshev's inequality: If X is a random variable with mean μ and variance σ^2, then for all $c > 0$

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

Proof: This is a direct consequence of applying Markov's inequality to $(X - \mu)^2$.

2 Statistics and estimators

A **parameter** is a value which determines the distribution of a random variable and is often denoted by θ. For instance, the normal distribution has two parameters: μ and σ.

A **statistic** is a function of random variables. For instance, given some random variables $X_1, ..., X_n$ we can define the statistic $\frac{X_1 + ... + X_n}{n}$.

A special type of statistic is a **point estimator** $\hat{\Theta}(X_1, ..., X_n)$. A point estimator is used to estimate some parameter, say θ. Once the random variables are observed as $x_1, ..., x_n$, we calculate $\hat{\Theta}(x_1, ..., x_n)$ as an **estimate** of θ.

Another type of statistic is the **test statistic**, a statistic used in hypothesis testing.

Given some data $x_1, ..., x_n$, the sample mean is the (observed) statistic

$$\bar{x} = \frac{x_1 + ... + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

the sample variance is the (observed) statistic

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Suppose $X_1, ..., X_n$ are independent and identical distributions with mean μ and variance σ^2. Define the statistic

$$\bar{X} = \frac{X_1 + ... + X_n}{n}$$

and

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

then

$$\mathrm{E}[\bar{X}] = \mu, \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \mathrm{E}[S^2] = \sigma^2$$

Proof: The first two are quite obvious. For the third one rewrite S^2 as

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

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

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

Now

$$\mathrm{E}[S^2] = \frac{1}{n-1}\left(\sum_{i=1}^{n} \mathrm{E}[X_i^2] - n\sum_{i=1}^{n} \mathrm{E}[\bar{X}^2]\right)$$

$$= \frac{1}{n-1}\left(n(\mu^2+\sigma^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right)$$

note the subtle use of $\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2$ here. Expanding, we see this is just σ^2, completing the proof.

To determine whether an estimator estimates a parameter well we can calculate the **bias**, defined as

$$b_\theta(\hat{\Theta}) = \mathrm{E}[\hat{\Theta} - \theta] = \mathrm{E}[\hat{\Theta}] - \theta$$

When the bias $= 0$, or $\mathrm{E}(\hat{\Theta}) = \theta$ we say the estimator is unbiased. In the above example, we see that $\bar{X}$ and S^2 are unbiased estimators of μ and σ^2, respectively.

The mean square error of an estimator is defined as

$$\mathrm{E}[(\hat{\Theta} - \theta)^2]$$

it measures how far, on average, the estimator is from the true value of the parameter it is trying to estimate.

A very useful formula is

$$\mathrm{E}[(\hat{\Theta} - \theta)^2] = b_\theta(\hat{\Theta})^2 + \mathrm{Var}(\hat{\Theta})$$

Proof: Using $\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2$ we see that

$$\mathrm{Var}(\hat{\Theta} - \theta) = \mathrm{E}[(\hat{\Theta} - \theta)^2] - \mathrm{E}[(\hat{\Theta} - \theta)]^2$$

$$\mathrm{Var}(\hat{\Theta}) = \mathrm{E}[(\hat{\Theta} - \theta)^2] - b_\theta(\Theta)^2$$

Let $x_1, ..., x_n$ be some data in ascending order. To find the **median** we need to calculate the index $i_{0.5}$, which is $(n+1)/2$. If this is an integer, the median $m = x_{i_{0.5}}$. Otherwise, $m = (x_{\lfloor i_{0.5} \rfloor} + x_{\lceil i_{0.5} \rceil})/2$. The index $i_{0.25} = (\lfloor i_{0.5} \rfloor + 1)/2 = (\lfloor \frac{n+1}{2} \rfloor + 1)/2$, and we can calculate the lower quartile (q_1) similar as median. To find the upper quartile (q_3) we find $i_{0.75} = n + 1 - i_{0.25}$.

The **interquartile range** (IQR) is defined as upper quartile $-$ lower quartile, and every data not in $(q_1 - 1.5 \cdot \mathrm{IQR}, q_3 + 1.5 \cdot \mathrm{IQR})$ is considered as an outlier (Tukey's criterion).

3 Normal random variables and confidence interval

We start the section by introducing an important theorem: Let $X_1, ..., X_n$ be i.i.d. random variables and distributed according to $N(\mu, \sigma^2)$. Then $\bar{X}$ and S^2 are independent random variables and

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

Another frequently used but often neglected theorem states that if we add independent normal variables we get another normal variable. Its mean and variance can be calculated using linearity of expectation and 'linearity' of variance.

Finally, the student t-distribution with dof $= v$ can be defined as

$$T = \frac{U}{\sqrt{V/v}}$$

where $U \sim N(0,1)$, $V \sim \chi^2_v$ and U, V independent.

Now we introduce the **confidence interval**. Let θ be a parameter of the random variable X. We say $[L(X), U(X)]$ is a $100(1-\alpha)\%$ confidence interval if

$$P(\theta \in [L(X), U(X)]) \geq 1 - \alpha$$

Suppose X is observed as x and we calculated the confidence interval $[L(x), U(x)]$, what can we say about θ? A common misconception is that θ has $100(1-\alpha)\%$ chance of being in the interval. Since θ and $[L(x), U(x)]$ are fixed, the probability is either 1 or 0. The correct interpretation is that if we continue to create confidence intervals, we would expect $100(1-\alpha)\%$ of them to contain θ.

3.1 Confidence interval for μ with σ known

Suppose $X_1, ..., X_n$ follow normal distribution with unknown mean θ, known variance σ^2. We first consider the sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\theta, \frac{\sigma^2}{n}\right)$$

Now let

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

$$P(z_{\alpha/2} < Z < z_{1-\alpha/2}) = 1 - \alpha$$

where z_α satisfying $\alpha = P(Z \leq z_\alpha)$ are the critical values of the normal distribution. Plugging in we have

$$P\left(-z_{1-\alpha/2} < \frac{\theta - \bar{X}}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

where we used the fact that $z_{\alpha/2} = -z_{1-\alpha/2}$ by symmetry. Rearranging yields

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Thus

$$\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\%$ confidence interval of θ.

3.2 Confidence interval for μ with σ unknown

Suppose independent random variables $X_1, ..., X_n$ follow normal distribution with unknown mean θ, unknown variance σ^2. Since σ^2 is unknown we use s^2 to estimate variance, but now are statistic follows t-distribution with dof $v = n-1$:

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

Similarly

$$P\left(-t_{v,1-\alpha/2} < \frac{\theta - \bar{X}}{s/\sqrt{n}} < t_{v,1-\alpha/2}\right) = 1-\alpha$$

$$P\left(\bar{X} - t_{v,1-\alpha/2}\frac{s}{\sqrt{n}} < \theta < \bar{X} + t_{v,1-\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

Thus

$$\left(\bar{X} - t_{v,1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + t_{v,1-\alpha/2}\frac{s}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\%$ confidence interval of θ.

3.3 Confidence interval for σ^2

Suppose independent random variables $X_1, ..., X_n$ follow normal distribution with unknown mean θ, unknown variance σ^2. To obtain a confidence interval for σ^2 we use

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

$$P\left(\chi^2_{n-1,\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,1-\alpha/2}\right) = 1-\alpha$$

Note that since χ^2 is not a symmetric distribution we can't write $\chi^2_{n-1,\alpha} = -\chi^2_{n-1,1-\alpha/2}$.

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right) = 1-\alpha$$

Thus

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right)$$

is a $100(1-\alpha)\%$ confidence interval of σ^2.

3.4 Confidence interval using Chebyshev's inequality

This method applies when $X_1, ..., X_n$ follow some distribution F_X (not necessarily normal) with unknown mean θ but known variance σ^2.

Mostly the question will assume that a population has some unknown proportion $\hat{p}$ of people support A. After we randomly selected n people and calculated the sample proportion p, find the $100(1-\alpha)\%$ confidence interval of $\hat{p}$. To start with, defined the test statistic X as the proportion of supporters in n randomly chosen people.

Clearly, the total number of supports among n people follows $Bin(n, p)$ with mean np and variance $np(1-p)$. Thus, the proportion will have mean p and variance $\frac{p(1-p)}{n}$.

$$P(|X-\mu_X| \geq c) \leq \frac{\sigma_X^2}{c^2}$$

$$P(|X-p| \geq c) \leq \frac{p(1-p)}{nc^2}$$

$$P(|X-p| \leq c) \geq 1-\frac{p(1-p)}{nc^2} \geq 1-\frac{1}{4nc^2}$$

By choosing c that ensures $1-\frac{1}{4nc^2} > 1-\alpha$, the $100(1-\alpha)\%$ confidence interval is simply

$$(p-c, p+c)$$

3.5 Student's two-sample test

Suppose we have i.i.d. random variables $X_1, ..., X_n \sim N(\theta_1, \sigma_1^2)$ and i.i.d. random variables $Y_1, ..., Y_m \sim N(\theta_2, \sigma_2^2)$. We also assume every X_i is independent of Y_j and $\sigma_1 = \sigma_2 = \sigma$.

Clearly, $\bar{X} \sim N(\theta_1, \sigma^2/n)$ and $\bar{Y} \sim N(\theta_2, \sigma^2/m)$ so

$$\bar{X}-\bar{Y} \sim N\left(\theta_1-\theta_2, \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right)$$

Then we have

$$T = \frac{(\theta_1-\theta_2)-(X_1-X_2)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$$

where

$$S_p^2 = \frac{1}{n+m-2}((n-1)S_X^2+(m-1)S_Y^2)$$

4 Hypothesis testing

To conduct hypothesis testing we first set the **null hypothesis** H_0. This is often the hypothesis we are willing to reject. The **alternate hypothesis** H_1 is the complement event to H_0. We will also choose the **significance threshold** α in advance, which is a value between 0 and 1 and is often chosen as 0.05.

Next we define some test statistic and suppose we observed the test statistic as x. The **p-value** is the probability of obtaining a test statistic as extreme as x, under the assumption that the null hypothesis is true.

If $p < \alpha$, we reject the null hypothesis. Otherwise, there are insignificant evidence to reject the null hypothesis.

However, sometimes we might not be able to calculate p-value directly. This happens when we conduct t-tests, since we only know the values of $P(T < t)$ for some values of t. In this case we can calculate the absolute value of test statistic $|t|$ and compare it with the critical value $t_{v,1-\alpha/2}$. If $|t| >$ critical value then we reject null hypothesis.

An alternate way of conducting hypothesis test is to use confidence interval. If the observed test statistic is not in the $100(1-\alpha)\%$ confidence interval, we reject the null hypothesis.

If we view the p-value as a random variable, it is worth remembering that it follows $U(0, 1)$ **only when the null hypothesis is actually true**.

Proof: For simplicity we assume the inverse cumulative distribution function of the test statistic T exists.

The **Type I error** refer to the situation when the null hypothesis is true, but is rejected in a hypothesis test. That is, $P(\text{reject } H_0|H_0 \text{ true}) = P(p < \alpha|H_0 \text{ true})$. But if H_0 is actually true, then p follows uniform distribution and clearly this probability is just α. Thus, the probability that a Type I error occurs is α.

The **Type II error**, on the other hand, refer to the situation when the null hypothesis is false but we failed to reject it in a hypothesis test. Let $P(\text{fail to reject } H_0|H_0 \text{ false}) = \beta$. Now that p does not follow a uniform distribution, we are unable to express β in terms of α. But we do know that as α decrease, β increase. The value $1 - \beta$ is defined as the **power** of the hypothesis test.

5 Pitfalls

5.1 Bonferroni correction

Imagine we conduct 100 hypothesis tests in a row, however we know for each test, the null hypothesis is actually true. Let the significance threshold be α. Then for each test, we have a α chance of rejecting H_0. Hence, the probability of making at least one Type I error is

$$1 - (1 - \alpha)^{100}$$

If $\alpha = 0.05$, then the probability is about 99.4%. That is, we will almost always make a Type I error.

To prevent this from happening, we introduce the **Bonferroni correction**: when conducting n hypothesis tests in a row, use α/n as the significance threshold.

Let A_i denote the event that we make a Type I error in the i^{th} test. Then the probability that we make at least one Type I error is

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \bigcup_{i=1}^{n} P(A_i) = n \cdot \frac{\alpha}{n} = \alpha$$

5.2 Simpson's paradox

Simpson's paradox is a phenomenon that allows a statistical trend (for instance, proportion) in groups of data to be reversed if these data are aggregated together.

If the subgroups have been split unevenly between the treatments, or if the subgroups are of very different sizes, then this can provide conditions that allow Simpson's paradox to occur.

5.3 Correlation does not imply causation

That's all you need to write to get marks.

6 Covariance and correlation

We have already defined covariance and some of its basic properties. For completeness I include them here without proof:

$$\operatorname{Cov}(X, Y) = \operatorname{E}[(X - \mu_X)(Y - \mu_Y)] = \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y]$$

$$\operatorname{Cov}(X, Y) = 0 \text{ if } X \text{ and } Y \text{ are independent}$$

$$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab \operatorname{Cov}(X, Y)$$

Now we continue to derive some new properties. First we have

$$\operatorname{Cov}(X, a) = 0 \text{ for all } a \in \mathbb{R}$$

(here a denotes constant distribution: it always outputs a)

Proof:

$$\operatorname{Cov}(X, a) = \operatorname{E}[(X - \mu_X)(a - \operatorname{E}(a))] = 0$$

Linearity of covariance:

$$\operatorname{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \operatorname{Cov}(X_i, Y_j)$$

Proof: We only need to show $\operatorname{Cov}(aX + bY, Z) = a\operatorname{Cov}(X, Z) + b\operatorname{Cov}(Y, Z)$, and then use induction and symmetry of covariance.

$$\operatorname{Cov}(aX + bY, Z) = \operatorname{E}[(aX + bY - \operatorname{E}[aX + bY])(Z - \mu_Z)]$$

$$= \mathrm{E}[(a(X-\mu_X)+b(Y-\mu_Y))(Z-\mu_Z)]$$

$$= a\,\mathrm{E}[(X-\mu_X)(Z-\mu_Z)] + b\,\mathrm{Cov}[(Y-\mu_X)(Z-\mu_Z)]$$

$$= a\,\mathrm{Cov}(X,Z) + b\,\mathrm{Cov}(Y,Z)$$

Bounds for covariance: For any X and Y we have

$$|\,\mathrm{Cov}(X,Y)| \leq \sigma_X\sigma_Y$$

This theorem inspires the definition of **correlation**:

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

clearly, $|\,\mathrm{Cor}(X,Y)| \leq 1$

Suppose the random variables $X_1,...,X_n$ follows F_X and $Y_1,...,Y_n$ follows F_Y, then the **sample covariance** is defined as

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

Suppose $X_1,...,X_n$ are observed as $x_1,...,x_n$ and $Y_1,...,Y_n$ are observed as $y_1,...,y_n$, then the observed sample covariance is

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

The observed sample correlation is defined as

$$r_{XY} = \text{observed sample covariance}/(s_Xs_Y)$$

$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

7 Likelihood

Given an distribution X with unknown parameter $\theta \in \Theta$, to estimate θ we can take some observations $x_1,...,x_n$ and compute the confidence interval of θ. Another way is to compute the **maximum likelihood estimator**.

Let the pmf/pdf of X be $f(\cdot|\theta)$. Having observed the data $\vec{x}$ (write the independent observations $x_1,...,x_n$ as a vector), the **likelihood function** $L(\cdot|\vec{x}) : \Theta \to \mathbb{R}$ is

$$L(\theta|\vec{x}) = f(\vec{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

setting the derivative with respect to θ to be 0 and checking the second derivative is negative, we can find the $\hat{\theta}$ that maximizes $L(\theta|\vec{x})$, the maximum likelihood estimator (MLE). It is also important to check the endpoints of Θ to ensure the local maximum is indeed a global maximum.

Note: We can also find the $\hat{\theta}$ that maximizes the log likelihood $\ln(L(\theta|\vec{x}))$, which sometimes will simplify the calculations.

8 Simple linear regression

Given random variable X and Y, after we observed pairs of observations $(x_1, y_1), ..., (x_n, y_n)$, we attempt to establish a linear relationship between X and Y of the form $Y = \beta_0 + \beta_1 X$ (if it exists). Consider the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i \in \{1, 2, ..., n\}$$

where

x_i are fixed, known values;

β_0, β_1 are unknown, fixed values;

ϵ_i are independent random variables and follow $N(0, \sigma^2)$ for some unknown σ^2;

Suppose in the experiment, the Y_i are observed as $y_1, ..., y_n$.

Clearly $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and we can find the MLE estimate for β_0, β_1 and σ^2. Calculating shows that we need to minimize

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x))^2$$

That is, the MLE estimate minimizes the residues. Rewrite the equation as

$$\sum_{i=1}^{n}((y_i - \beta_1 x) - \beta_0)^2$$

This is minimized when

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i) = \bar{y} - \beta_1 \bar{x}$$

Now plug $\hat{\beta}_0$ back in we get

$$\sum_{i=1}^{n}((y_i - \beta_1 x) - (\bar{y} - \beta_1 \bar{x}))^2$$

$$= \sum_{i=1}^{n}((y_i - \bar{y}) - \beta_1(x_i - \bar{x}))^2$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\beta_1 \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) + \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= S_{yy} - 2\beta_1 S_{xy} + \beta_1^2 S_{xx}$$

which is minimized when

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

After we find β_0 and β_1, we need the plot the residues $\hat{\epsilon}_i = y_i - (\beta_0 + \beta_1 x_i)$ against x_i. If our model is correct, the residues should appear to be independent.

9 Permutation test and bootstrap

Suppose we have two samples of observations $x_1, ..., x_n$ and $y_1, ..., y_m$ and we wish to test if there is a statistically significant difference between these two samples.

Suppose the samples are independent observations of random variables $X_1, ..., X_n \sim F_X$ and $Y_1, ..., Y_m \sim F_Y$. Let the null hypothesis be

$$H_0 : F_X = F_Y$$

and the test statistic be the difference in two sample means:

$$T = \bar{X} - \bar{Y}$$

say we observed the test statistic as t.

Now consider the pooled observations $\mathbf{z} = \{x_1, ..., x_n, y_1, ..., y_m\}$. Randomly select n observations **without replacement** as the first sample and let the remaining m observations be the second variable. We can compute the test statistic t^* and repeat B times, where B is large. We call $\{t_1^*, ..., t_B^*\}$ the empirical **permutation distribution** of the statistic.

If our original test statistic t falls outside the middle $100(1-\alpha)\%$ of the empirical distribution, we reject the null hypothesis at a significant threshold of α.

To calculate the p-value, find $\tilde{p_{\geq}}$, the proportion of permuted t-statistics greater than t, and similarly $\tilde{p_{\leq}}$. Then the p-value is

$$2 \cdot \min\{\tilde{p_{\leq}}, \tilde{p_{\geq}}\}$$

The bootstrap is similar to the permutation test. However, when creating the empirical distribution of the data set $\{x_1, ..., x_n\}$ we random select n observations **with replacement**.