APH101 Lecture Note

Kunyang He

April 2023

Great thanks to Jieyun Yin and Daiyun Huang.

Contents

Preface

This is the APH101 lecture note, the contents of the lecture note can be divided into 4 chapters. They are the **Hypothesis Test and parameter estimation**, **Parametric test**, **Nonparametric test**, **Regression analysis**. most of the contents come from the Powerpoint we have. But I add some mathematics methods to prove the ANOVA model and regression statistics' characteristics. you can read it or not. Moreover, I follow some universities' lecture notes to make the contents more interesting. They are Duke STA102, JHU PH140, USC BMTRY 711, Stanford STA 144, PSU STA462, Stanford STA 203. If there are some mistakes you find. please let me know, Great thanks.

Chapter 1

Hypothesis Test and parameter estimation

1.1 Hypothesis Test

1.1.1 Logic of hypothesis test

definition

The hypothesis test is also called significance testing. It tests a claim about a parameter using evidence/data from samples(s).

Small probability event

If the probability of some special event is very low (close to 0), then it is a rare event.
In one experiment, the event with very small probabilities will not occur.

Proof by contradiction

A conclusion drawn on the basis of an inductive method can never be fully proven. Proof by contradiction is relatively easy to form the negation of the proposition.

1) Females don't smoke.
2) All the swans are white.

Logic of hypothesis test

1) Aim ofthe study.

2) hypothesis. (H_0: Null hypothesis, H_1: Alternative hypothesis).

3) Calculate the test statistic.

4) Statistical decision: accept/reject H_0.

H_0: Null hypothesis

A null hypothesis is often concerned with a parameter or parameters of population(s). The hypothesis to be test. It is always stated in the null form, indicating no difference or no relationship between distributions or parameters.

H_1: Alternative hypothesis

A hypothesis that in some sense contradicts the null hypothesis H_0. Claims H_0 is false

$$P(H_0) + P(H_1) = 1$$

Level of significance

Definition: Probability $< \alpha$ is a rare event.(Usually $\alpha = 0.05$)
Determines the reject region.

1.1.2 Steps of hypothesis test

Here is an example. In previous large-scale survey, the birth length of male newborns was normally distributed, with an average of $49cm$. A gynecologist in a hospital randomly selected 25 male babies born between 2010 and 2013, and found that the average birth length of the 25 babies was $50cm$, and the standard deviation was $5cm$. The gynecologist wants to know whether the birth length of male babies born between 2010 and 2013 is different from $49cm$?

Step 1: State hypothesis and α

$$H_0 : \mu = \mu_0$$

The μ of birth length of babies born between 2010 and 2013 is identical to 49cm.

$$H_1 : \mu \neq \mu_0$$

The μ of birth length of babies born between 2010 and 2013 is not identical to 49cm

$$\alpha = 0.05$$

Choose method and calculate test statistics

Choose method: depends on data type, study design, study aims and conditions. The sampling distributions of a mean follows t distribution.

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}. \quad t = \frac{\bar{x} - \mu_0}{S_{\bar{x}}} = \frac{50 - 49}{5/\sqrt{25}} = 1.000$$

Make a statistical decision

Our example is a two-sided test

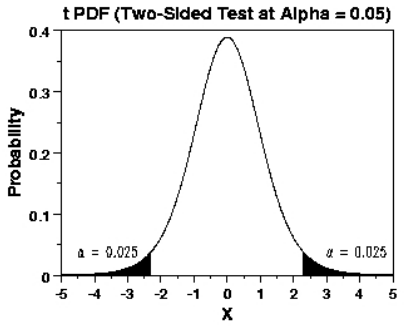


Figure 1.1: Our example is a two-sided test with $\alpha = 0.05$

$$t < t_{0.05/2.24} = 2.064. \quad P > 0.05$$

P value

The P value is the probability of getting values of the test statistic as extreme as, or more extreme than, that observed if the null hypothesis is true.

Errors

Type I and Type II error. Four possible outcomes

	Reject H_0, $accept\mathbf{H}_1$	Not reject H_0
True H_0	Type I error (α)	Correct decision. ($1 - \alpha$)
False H_0	Correct decision (statistic power, 1- β)	Type II error ((β)

A Type I error means rejecting the null hypothesis when it's actually true.
A Type II error means not rejecting the null hypothesis when it's actually false.
Statistics power is the probability of rejecting a false null hypothesis.

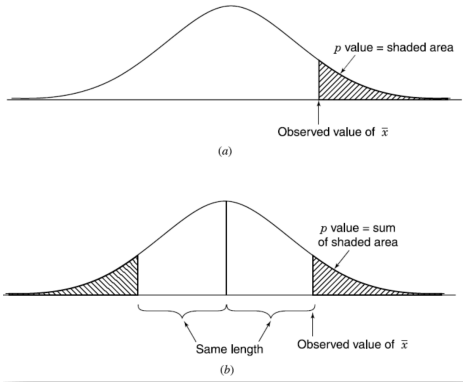


Figure 1.2: One-side and two-side

1.1.3 Relationship to confidence intervals

1) If μ_0 is not included in the 0.95 confidence interval for μ, H_0 should be rejected at the 0.05 level.

2) If μ_0 is included in the 0.95 confidence interval for μ_0, H_0 should not be rejected at the 0.05 level.

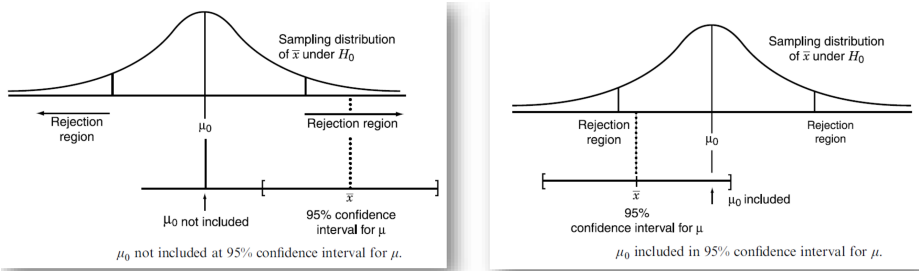


Figure 1.3: two conditions

1.2 Maximum likelihood estimation

Distribution	Parameters
$Bernoulli(p)$	$\theta = p$
$Poisson(\lambda)$	$\theta = \lambda$
$Uniform(a, b)$	$\theta = (a, b)$
$Normal\left(\mu, \sigma^2\right)$	$\theta = \left(\mu, \sigma^2\right)$
$Y = mX + b$	$\theta = (m, b)$

In the real world often you don't know the "true" parameters, but you get to observe data.

1.2.1 Point estimation Maximum likelihood estimation

- The central idea behind MLE is to select that parameters () that make the observed data the most likely.
- Raised by Gaussian (in 1821), development by Fisher (in 1912).

1.2.2 Probability & Likelihood (function)

- $P(X;\theta)$: **Probability** refers to the chance that a particular outcome
- $L(\theta;X)$: **Likelihood** refers to how well a sample provides support for particular values of a parameter in a model, is the probability that we see the data we see if we set the parameter equal to θ.

Likelihood refers to how well a sample provides support for particular values of a parameter in a model, is the probability that we see the data we see if we set the parameter equal to of discrete distributions, likelihood is a synonym for the joint probability of your data. In the

1.2.3 Likelihood(function)

Suppose we have a countinous random sample $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \ldots \mathrm{X}_n$ for which the probability density (or mass) function of each X_i is $\mathrm{f}\left(x_\mathrm{i};\theta\right)$. Since we assumed each data point is independent, the likelihood of all our data is the product of the likelihood of each data point, then the joint probability mass (or density) function of $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \ldots \mathrm{X}_\mathrm{n}$, which we'll call $\mathrm{L}(\theta)$ is:

$$\begin{aligned}\mathrm{L}(\theta) &= \mathrm{P}\left(\mathrm{X}_1 = x_1, \mathrm{X}_2 = x_2, \ldots., \mathrm{X}_\mathrm{n} = x_\mathrm{n}\right) \\ &= \mathrm{f}\left(x_1;\theta\right) * \mathrm{f}\left(x_2;\theta\right) \ldots . \mathrm{f}\left(x_\mathrm{n};\theta\right) \\ &= \textstyle\prod_{i=1}^{n} f\left(x_\mathrm{i};\theta\right)\end{aligned}$$

To simplify notation, let the vector $\mathbf{X} = (x_1, x_2, \ldots . x_n)$ denote the observed sample. Then the joint pdf and likelihood function may be expressed as $f(\mathbf{X};\theta)$ and $\mathrm{L}(\theta;\mathbf{X})$, respectively

1.2.4 Maximum

In MLE our goal is to chose values of our parameters (θ) that maximizes the likelihood function. We are going to use the notation $\hat{\theta}$ to represent the best choice of values for our parameters. Formally, MLE assumes that:

$$\hat{\theta} = \theta \arg\max L(\theta)$$

If we find the arg max of the log of likelihood, it will be equal to the arg max of the likelihood. Therefore, for MLE, we first write the log likelihood function (LL)

$$LL(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f\left(X_i \mid \theta\right) = \sum_{i=1}^{n} \log f\left(X_i \mid \theta\right)$$

taking the derivative of the log-likelihood, and setting it to 0

$$\frac{\partial LL(\theta)}{\partial \theta} = 0$$

1.2.5 Bernoulli Case

If the $\mathrm{X_i}$ are independent Bernoulli random variables with unknown parameter p , then the probability mass function of each X_i is:

$$f\left(x_i; p\right) = p^{x_i}(1-p)^{1-x_i}$$
$$\mathrm{X_i} = 0 \text{ or } 1, 0 < \mathrm{p} < 1$$

Therefore, the likelihood function is

$$L(p) = \prod_{i=1}^{n} p^{x1}(1-p)^{1-x1} \times p^{x2}(1-p)^{1-x2} \times \cdots \times p^{xn}(1-p)^{1-xn}$$

Simplifying, by summing up the exponents, we get :

$$L(p) = p^{\sum_{i=1}^{n} xi}(1-p)^{n-\sum_{i=1}^{n} xi}$$

In this case, the natural logarithm of the likelihood function is:

$$LL(p) = \left(\sum_{i=1}^{n} x_i\right) \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

Now, taking the derivative of the log-likelihood, and setting it to 0 , we get:

$$\frac{\partial LL(p)}{\partial p} = \frac{\left(\sum_{i=1}^{n} x_i\right)}{p} - \frac{\left(n - \sum_{i=1}^{n} x_i\right)}{1-p} = 0$$

multiplying through by $p(1-p)$, we get:

$$\left(\sum_{i=1}^{n} x_i\right)(1-p) - \left(n - \sum_{i=1}^{n} x_i\right) p = 0$$
$$\sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i - np + p\sum_{i=1}^{n} x_i = 0$$
$$\sum_{i=1}^{n} x_i - np = 0$$

1.2.6 Normal Case

Let $X_1, X_2, X_3, \ldots . X_n$ be a random sample from a normal distribution with unknown mean μ and variance σ^2 . Find maximum likelihood estimators of mean μ and variance σ^2 .

In finding the estimators, the first thing we'll do is write the probability density function (to make it simpler, here $\hat{\mu}$ is wirte as μ , and $\widehat{\sigma^2}$ is write as σ^2).

$$f\left(x_i; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x_i - \mu\right)^2}{2\sigma^2}\right)$$

Now, that makes the likelihood function:

$$L\left(\mu, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x_i - \mu\right)^2}{2\sigma^2}\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}\left(x_i - \mu\right)^2}{2\sigma^2}\right)$$

and therefore the log of the likelihood function:

$$LL\left(\mu, \sigma^2\right) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$$

Now, upon taking the partial derivative of the log likelihood with respect to μ, and setting to 0 , we see that a few things cancel each other out, leaving us with:

$$\frac{\partial LL\left(\mu, \sigma^2\right)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 = 0$$

Denominator σ^2 can't be zero, then

$$\sum_{i=1}^{n}\left(x_i - \mu\right)^2 = 0, \sum_{i=1}^{n} x_i - n\mu = 0$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$$LL\left(\mu, \sigma^2\right) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$$

Now, for σ^2 . Taking the partial derivative of the log likelihood with respect to σ^2 , and setting to 0 , we get :

$$\frac{\partial LL\left(\mu, \sigma^2\right)}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 = 0$$

Consider σ^2 as a whole Multiplying through by $2\sigma^4$

$$\frac{\partial LL\left(\mu, \sigma^2\right)}{\partial \sigma^2} = \left[-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 = 0\right] \times 2\sigma^4$$

$$-n\sigma^2 + \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$
$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

And, solving for σ^2 , and putting on its hat, we have shown that the maximum likelihood estimate of σ^2 is (we already known that $\hat{\mu} = \bar{x}$):

Chapter 2

Parametric test

2.1 Related Distributions

2.1.1 Standard Normal distribution. Three major distributions

PDF:

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2}, -\infty < z < +\infty$$

CDF:

$$\Phi(z) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{z} e^{-z^2/2} dz, -\infty < z < +\infty$$

2.1.2 Chi-square distribution

Definition

(1) If $X \sim N\left(\mu, \sigma^2\right), (X_1, \ldots, X_n)$ are randomly selected from population X The distribution of $Y \sim \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \mu\right)^2$

$$X \sim N\left(\mu, \sigma^2\right) .so \frac{x_{\mathrm{i}} - \mu}{\sigma} \sim N(0, 1)$$

According to the definition of chi-square distribution

$$Y = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right)^2 \sim x_n^2$$

Properties of chi-square distribution

(2) If $Y \sim x_n^2$, then E(Y) = n , and D(Y) = 2n

Prove:

$$Y = x_1^2 + x_2^2 + x_3^2 + \ldots. + x_n^2$$
$$Ex_i^2 = Dx_{\mathrm{i}} + \left(Ex_{\mathrm{i}}\right)^2 = 1$$
$$EY = E\left(\sum_{i=1}^{n} x_{\mathrm{i}}^2\right) = \sum_{i=1}^{n} Ex_{\mathrm{i}}^2 = n$$

because:

$$Dx_i^2 = Ex_{\mathrm{i}}^4 - \left(Ex_{\mathrm{i}}^2\right)^2 = 3 - 1 = 2$$

$$DY = D\left(\sum_{i=1}^{n} x_i^2\right) = \sum_{i=1}^{n} Dx_i^2 = 2n$$

Properties of chi-square distribution

(3) $Y_1 \sim x_m^2, Y_2 \sim x_n^2$, and $Y1$ and $Y2$ are independent.

$$Y_1 + Y_2 \sim x_{m+n}^2$$

Properties of chi-square distribution

(4)

$$\begin{aligned} x_n^2 = \sum_{i=1}^{n} \left(\tfrac{x_i - \mu}{\sigma}\right)^2 &\approx \sum_{i=1}^{n} \tfrac{\left(x_i - \bar{X}\right)^2}{\sigma^2} \\ S^2 &= \sum_{i=1}^{n} \tfrac{\left(x_i - \bar{X}\right)^2}{n-1} \\ \sum_{i=1}^{n} \left(x_i - \bar{X}\right)^2 &= (n-1)^* S^2 \\ \tfrac{(n-1)S^2}{\sigma^2} &\sim \chi_{n-1}^2 \end{aligned}$$

2.1.3 F distribution

Definition

if $X \sim x_n^2, Y \sim x_m^2$. X and Y are independent.

$$F = \tfrac{X/n}{Y/m}, \text{ then}$$
$$F \sim F_{n,m}, F \sim F(n,m)$$

The PDF of an F random variable with $r1$ numerator degrees of freedom and r2 denominator degrees of freedom is:

$$f(w) = \frac{\left(r_1/r_2\right)^{r_1/2} \Gamma\left[\left(r_1 + r_2\right)/2\right] w^{(r_1/2)-1}}{\Gamma\left[r_1/2\right] \Gamma\left[r_2/2\right] \left[1 + \left(r_1 w/r_2\right)\right]^{(r_1+r_2)/2}}$$

over the support $w \geq 0$.

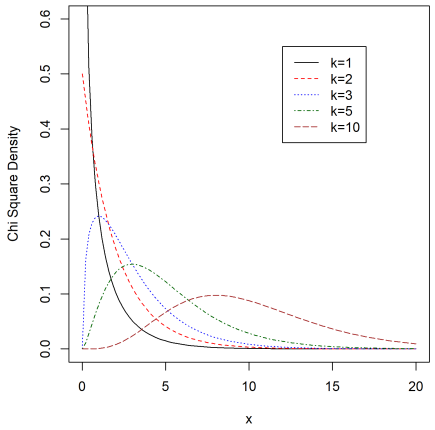


Figure 2.1: 1st argument is the value of x we want to compute the density and the second one is the degree of freedom.

Properties of an F-distribution

(1) If $T \sim t_n$, then $T^2 \sim F(1, n)$
Prove According to the definition of t distribution

$$T = \frac{x}{\sqrt{Y/n}}$$

Of which, $\mathrm{X} \sim \mathrm{N}(0,1), \mathrm{Y} \sim \mathrm{X}_{\mathrm{n}}^2$, and X and Y are independent

$$T^2 = \frac{x^2}{Y/n} = \frac{x^2/1}{Y/n}$$

Of which, $\mathrm{X}^2 \sim \mathrm{X}_1^2$

2.1.4 T distribution

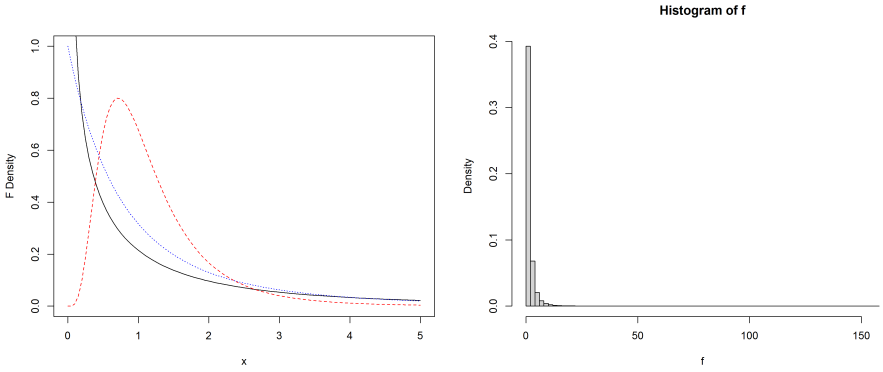
Definition

If $X \sim N(0,1), Y \sim X_n^2$, and X and Y are independent.

$$T = \frac{x}{\sqrt{Y/n}}$$

Then T follows a t distribution with df = n , we write

$$T \sim t_n$$



(a) different degree (b) f distribution using chisquare distributions

Figure 2.2: (a) plot f distributions with different degree. (b) construct f distribution using chisquare distributions

Probability density function Student's $t-distribution$ has the probability density function (PDF) given by

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

where ν is the number of degrees of freedom and Γ is the gamma function. This may also be written as

$$f(t) = \frac{1}{\sqrt{\nu}\mathrm{B}\left(\frac{1}{2},\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

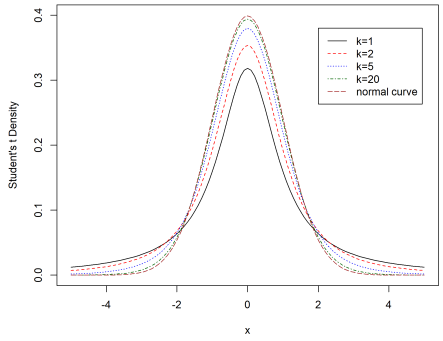


Figure 2.3: The following code chunk plots the t curves with various degrees of freedom.

Properties of t-distribution

1) If $n = 1$, then it follows a Cauchy distribution
2) If $\underline{n > 2}$, then $ET = 0, DT = \frac{n}{n-2}$
3) If n becomes large, the t distribution converges to an $\mathrm{N}(0, 1)$ distribution.
4) t distributions have heavier tails than $\mathrm{N}(0, 1)$

2.2 T test

2.2.1 One sample t test

In testing the null hypothesis that the population mean is equal to a specified value μ_0 , one uses the statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation and n is the sample size. The degrees of freedom used in this test are $n - 1$. Although the parent population does not need to be normally distributed, the distribution of the population of sample means $\bar{x}$ is assumed to be normal.
By the central limit theorem, if the observations are independent and the second moment exists, then t will be approximately normal $\mathrm{N}(0, 1)$.

2.2.2 Paired T test

Paired t-test is suitable for comparing two sample means of a paired design.
Testing condition: The testing condition of paired t-test includes the assumption of normal distribution of the differences.
H_0: population mean of difference = 0
The test statistic of a paired t test is

$$t = \frac{\bar{d} - 0}{S_d/\sqrt{n}} \quad v = n - 1$$

Where $\bar{d}$ is the sample mean of paired measurement differences, S_d is the standard deviation of the differences, $\underline{n}$ is the number of the pairs and v is the degree of freedom.

$$s_d = \sqrt{\left[\sum_{i=1}^{n} d_i^2 - \left(\sum_{i=1}^{n} d_i\right)^2 /n\right] /(n-1)}$$

n = number of matched pairs

2.2.3 t test for two independent samples

If X and Y are independent,

$$X \sim N\left(\mu_1, \sigma_1^2\right), Y \sim N\left(\mu_2, \sigma_2^2\right).$$
$$X - Y \sim \left(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2\right)$$

Independent Sampling from Two Populations.

$$\bar{X}_1 \sim N\left(\mu_1, \sigma_1^2/n_1\right), \bar{X}_2 \sim N\left(\mu_2, \sigma_2^2/n_2\right)$$
$$\bar{X}_1 - \bar{X}_2 \sim \left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$
$$z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Suppose $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\frac{\frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}{\sqrt{\left(\frac{(n_1-1)S_1^2}{\sigma_1^2} + \frac{(n_2-1)S_2^2}{\sigma_2^2}\right)/n_1+n_2-2}}$$
$$= \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}} \sim t\left(n_1 + n_2 - 2\right)$$

Equal Variance:
If $\sigma_1^2 = \sigma_2^2$

$$\bar{X}_1 - \bar{X}_2 \sim \left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \sim \left[\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]\Big)$$

Under H_0 , we know that $\mu_1 = \mu_2$.

$$\bar{X}_1 - \bar{X}_2 \sim N\left[0, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right], \quad \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

2.2.4 Two-Sample T Test for Independent Samples with Equal Variances

Suppose we wish to test the hypothesis $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$ with a significance level of α for two normally distributed populations, where σ^2 is assumed to be the same for each population. Compute the test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s = \sqrt{\left[(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2\right] / (n_1 + n_2 - 2)}$$

If $t > t_{n_1+n_2-2,1-\alpha/2}$ or $t < -t_{n_1+n_2-2,1-\alpha/2}$ then H_0 is rejected.
If $-t_{n_1+n_2-2,1-\alpha/2} \leq t \leq t_{n_1+n_2-2,1-\alpha/2}$ then H_0 is accepted.

2.2.5 Interval estimation(Equation variance case)

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

$$\left(\bar{X}_1 - \bar{X}_2 - t_{n_1+n_2-2,1-\alpha/2} S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X}_1 - \bar{X}_2 + t_{n_1+n_2-2,1-\alpha/2} S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

2.2.6 If the variances of the two populations are not equal

Behrens-Fisher problem

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Under $\mathrm{H}_0, \mu 1 = \mu 2$, If σ_1^2 and σ_2^2 were known

$$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \longrightarrow \mathrm{z} = (\bar{x}_1 - \bar{x}_2) \Big/ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Under $\mathrm{H}_0, \mu 1 = \mu 2$, If σ_1^2 and σ_2^2 were unknown.

$$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \longrightarrow \mathrm{t} = (\bar{x}_1 - \bar{x}_2) \Big/ \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

approximate t test is recommended.
Welch's t test

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Compute the approximate degrees of freedom

$$v' = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1+1} + \frac{\left(S_2^2/n_2\right)^2}{n_2+1}} - 2$$

Scatterthwaite's test

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Compute the approximate degrees of freedom

$$v' = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}} - 2$$

2.3 One-Way ANOVA

2.3.1 ANOVA Hypothesis Testing

Null: All the population means are equal, i.e.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$$

Alternative: Not all the μ_i 's are equal, i.e.

$$H_a : \mu_i \neq \mu_j$$

for some i, j

2.3.2 Calculation Method

Definition 1. Between Mean Square is defined by

$$\text{MSB} = \frac{\text{SSB}}{k-1}.$$

Definition 2. Within Mean Square is defined by

$$\text{MSW} = \frac{\text{SSW}}{n-k}.$$

$$\frac{MSB}{MSW} = \frac{SSB/k-1}{SSW/n-k} = \frac{\sum_{\text{i}=1}^{k}\sum_{j=1}^{ni}\left(\bar{X}_i - \overline{\bar{X}}\right)^2/k-1}{\sum_{\text{i}=1}^{k}\sum_{j=1}^{ni}\left(X_{ij} - \bar{X}_i\right)^2/n-k}$$

If divided by the same σ^2

$$\frac{MSB}{MSW} = \frac{\sum_{\text{i}=1}^{k}\sum_{j=1}^{ni}\left(\bar{X}_i - \overline{\bar{X}}\right)^2/(k-1)}{\sum_{\text{i}=1}^{k}\sum_{j=1}^{ni}\left(X_{ij} - \bar{X}_i\right)^2/(n-k)} = \frac{\chi^2_{k-1}/k-1}{\chi^2_{n-k}/n-k} \sim F_{(k-1,n-k)}$$

2.3.3 One-Way ANOVA Model

The statistic model for one-way ANOVA is

$$Y_{ij} = \mu + \theta_i + \epsilon_{ij}$$

1) Y_{ij} is the j^{th} observation of the i^{th} treatment (group)

2) μ is the overall mean level.

3) θ_i is the differential effect of the i^{th} treatment.

4) The θ_i are normalized: $\sum_{i=1}^{I} \theta_i = 0.\epsilon_{ij}$ is the random error and $\epsilon_{ij} \sim N\left(0, \sigma^2\right)$.

The mean of the i^{th} treatment group is $\mu_i = \mu + \theta_i$. It follows that

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_I = 0$$

The analysis of variance is based on the following identity:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{..}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i.}\right)^2 + J\sum_{i=1}^{I}\left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2$$

and the identity may be symbolically expressed as

$$SST = SSW + SSB$$

SST is the total sum of squares (total variation)
SSW is the error sum of squares (variation within groups)
SSB is the sum of squares among groups (variation among groups)

Theorem:

$$E(SSW) = I(J-1)\sigma^2 \Longrightarrow E\left[\frac{SSW}{I(J-1)}\right] = \sigma^2$$

$$E(SSB) = J\sum_{i=1}^{I}\theta_i^2 + (I-1)\sigma^2$$

Under $H_0, E(SSB) = (I-1)\sigma^2$ and $E[SSB/(I-1)] = \sigma^2$.
Theorem:

$$SSB/\sigma^2 \sim \chi^2_{I-1}$$

$$SSW/\sigma^2 \sim \chi^2_{I(J-1)}$$

And SSW and SSB are independent. Consequently,

$$F = \frac{SSB/(I-1)}{SSW/[I(J-1)]} = \frac{MSB}{MSW} \sim F_{(I-1),I(J-1)}$$

2.4 Two-Way ANOVA Model

2.4.1 State Hypothesis

(1) H_0: There is no interaction between factors. H_1: There is a significant interaction between factors

(2) H_0: There is no effect of Factor A on the response variable. H_1: There is an effect of Factor A on the response variable

(3) H_0: There is no effect of Factor B on the response variable. H_1: There is an effect of Factor B on the response variable

2.4.2 Calculation Method

$$SS_{\text{Total}} = \sum_{i=1}^{i}\sum_{i=1}^{j}\sum_{i=1}^{k}\left(X_{ijk} - \bar{X}\right)^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{N}$$

$$SS_{\text{treatment}} = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\left(\bar{X}_{ij} - \bar{X}\right)^2$$

$$SSA = \sum_{i=1}^{a} \mathrm{n_i}\left(\left(\bar{X}_{i\cdot}\right) - \bar{X}\right)^2 \quad SSB = \sum_{j=1}^{b} \mathrm{n_j}\left(\left(\bar{X}_{\cdot j}\right) - \bar{X}\right)^2$$

$$SSAB = SS_{\text{treatment}} - SSA - SSB$$

$$SS_{error} = SST - SS_{\text{treatment}}$$

1) $\bar{X}_{ij}$: Cell mean

2) $\bar{X}_{i\cdot}$: mean of all observations in the ith row of the data matrix averaged across all columns.

3) $\bar{X}_{\cdot j}$: mean of all observations in the jth column of the data matrix averaged across all rows.

Table 3: Interaction Model ANOVA Table.

Source	**DF**	**SS**	**MS**	F-statistics
A	$\mathrm{I}-1$	SS_A	$\mathrm{MS}_A = \mathrm{SS}_A/\mathrm{df}_A$	$\mathrm{MS}_A/\mathrm{MS}_\mathrm{E}$
B	$\mathrm{J}-1$	SS_B	$\mathrm{MS}_\mathrm{B} = \mathrm{SS}_\mathrm{B}/\mathrm{df}_\mathrm{B}$	$\mathrm{MS}_\mathrm{B}/\mathrm{MS}_\mathrm{E}$
A:B (interaction)	$(\mathrm{I}-1)(\mathrm{J}-1)$	SS_{AB}	$\mathrm{MS}_{AB} = \mathrm{SS}_{AB}/\mathrm{df}_{AB}$	$\mathrm{MS}_{AB}/\mathrm{MS}_\mathrm{E}$
Error	$\mathrm{N}-\mathrm{IJ}$	SS_E	$\mathrm{MS}_\mathrm{E} = \mathrm{SS}_\mathrm{E}/\mathrm{df}_\mathrm{E}$	
Total	$\mathrm{N}-1$	$\mathrm{SS}_{\text{Total}}$		

2.4.3 Two-Way ANOVA Model

Two-way ANOVA

Two-way ANOVA model: observations: $\left(Y_{ijk}\right), 1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n_{ij}$: r groups in first grouping variable, m groups in second and n_{ij} samples in (i,j) -"cell":

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N\left(0, \sigma^2\right)$$

Constraints:

$$\sum_{i=1}^{r} \alpha_i = 0$$
$$\sum_{j=1}^{m} \beta_j = 0$$
$$\sum_{j=1}^{m} (\alpha\beta)_{ij} = 0, 1 \leq i \leq r$$
$$\sum_{i=1}^{r} (\alpha\beta)_{ij} = 0, 1 \leq j \leq m.$$

Two-way random effects model*

$$Y_{ijk} \sim \mu.. + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}, 1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n$$
$$\varepsilon_{ijk} \sim N\left(0, \sigma^2\right), 1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n$$
$$\alpha_i \sim N\left(0, \sigma_\alpha^2\right), 1 \leq i \leq r$$
$$\beta_j \sim N\left(0, \sigma_\beta^2\right), 1 \leq j \leq m$$
$$(\alpha\beta)_{ij} \sim N\left(0, \sigma_{\alpha\beta}^2\right), 1 \leq j \leq m, 1 \leq i \leq r$$
$$Cov\left(Y_{ijk}, Y_{i'j'k'}\right) = \delta_{ii'}\sigma_\alpha^2 + \delta_{jj'}\sigma_\beta^2 + \delta_{ii'}\delta_{jj'}\sigma_{\alpha\beta}^2 + \delta_{ii'}\delta_{jj'}\delta_{kk'}\sigma^2$$

ANOVA tables: Two-way (random)*

S S	d f	E(S S)
S S A=n m $\sum_{i=1}^{r}\left(\bar{Y}_{i..} - \bar{Y}\ldots\right)^2$	r-1	$\sigma^2 + nm\sigma_\alpha^2 + n\sigma_{\alpha\beta}^2$
S S B=n r $\sum_{j=1}^{m}\left(\bar{Y}_{.j} - \bar{Y}_{\ldots}\ldots\right)^2$	m-1	$\sigma^2 + nr\sigma_\beta^2 + n\sigma_{\alpha\beta}^2$
S S A B=n $\sum_{i=1}^{r}\sum_{j=1}^{m}\left(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j} + \bar{Y}_{\ldots}\ldots\right)^2$	(m-1)(r-1)	$\sigma^2 + n\sigma_{\alpha\beta}^2$
S S E=$\sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}\left(Y_{ijk} - \bar{Y}_{ij.}\right)^2$	(n-1) a b	σ^2

To test H_0: $\sigma_\alpha^2 = 0$ use SSA and $SSAB$.
To test H_0: $\sigma_{\alpha\beta}^2 = 0$ use $SSAB$ and SSE.

Two-way mixed effects model*

$Y_{ijk} \sim \mu.. + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}, 1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n$
$\varepsilon_{ijk} \sim N\left(0, \sigma^2\right), 1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n$
$\alpha_i \sim N\left(0, \sigma_\alpha^2\right), 1 \leq i \leq r.$
$\beta_j, 1 \leq j \leq m$ are constants.
$(\alpha\beta)_{ij} \sim N\left(0, (m-1)\sigma_{\alpha\beta}^2/m\right), 1 \leq j \leq m, 1 \leq i \leq r$

Constraints:

$$\sum_{j=1}^{m} \beta_j = 0$$
$$\sum_{i=1}^{r} (\alpha\beta)_{ij} = 0, 1 \leq i \leq r.$$
$$Cov\left((\alpha\beta)_{ij}, (\alpha\beta)_{i'j'}\right) = -\sigma^2_{\alpha\beta}/m$$
$$Cov\left(Y_{ijk}, Y_{i'j'k'}\right) = \delta_{jj'}\left(\sigma^2_\beta + \delta_{ii'}\frac{m-1}{m}\sigma^2_{\alpha\beta} - (1-\delta_{ii'})\frac{1}{m}\sigma^2_{\alpha\beta} + \delta_{ii'}\delta_{kk'}\sigma^2\right)$$

ANOVA tables: Two-way (mixed)*

S S	d f	E(M S)
S S A	r-1	$\sigma^2 + nm\sigma^2_\alpha$
S S B	m-1	$\sigma^2 + nr\frac{\sum_{j=1}^{m}\beta_i^2}{m-1} + n\sigma^2_{\alpha\beta}$
S S A B	(m-1)(r-1)	$\sigma^2 + n\sigma^2_{\alpha\beta}$
S S E=$\sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}\left(Y_{ijk} - \bar{Y}_{ij\cdot}\right)^2$	(n-1) a b	σ^2

To test H_0: $\sigma^2_\alpha = 0$ use SSA and SSE .
To test H_0: $\beta_1 = \cdots = \beta_m = 0$ use SSB and $SSAB$.
To test H_0: $\sigma^2_{\alpha\beta}$ use $SSAB$ and SSE.

Chapter 3

Nonparametric test

3.1 Nonparametric Test

3.1.1 Introduction: Distribution-Free Tests

- **Parametric statistics** is a branch of statistics which assumes that sample data comes from a population that follows a probability distribution (i.e., normal distribution) based on a fixed set of parameters. Most well-known elementary statistical methods are parametric (z-test, t-test, F-test).
- **Nonparametric tests** also called distribution-free hypothesis tests, generally have fewer required conditions. In particular, nonparametric tests do not require the population to follow a particular distribution, such as the normal distribution .Nonparametric tests replace the actual data values with either signs (positive or negative) or ranks. Do not deal with specific population parameters, such as the mean or standard deviation.

Nonparametric test conditions

- Unknown distribution; Skewed distribution; Uncertain data
- Variances are not equal
- Ordinal data -example: good-better-best
- Nominal data -example: male-female
- Outlier

Advantages and disadvantage

Advantage:

- Used with all scales

- Easier to compute
- Developed originally before wide computer use
- Make fewer assumptions
- Need not involve population parameters
- Results may be as exact as parametric procedures

disadvantage:

- May waste information (If data permit using parametric procedures)
- Difficult to compute by hand for large samples
- Tables not widely available

3.1.2 Wilcoxon Signed-Rank Test: Paired comparisons

Wilcoxon Signed-Rank

- Tests probability distributions of two related populations
- Corresponds to t-test for dependent (paired) means
- Assumptions (1.Random samples 2.Both populations are continuous)
- Can use normal approximation if n¿25

Wilcoxon Signed-Rank Test procedure

- **Hypothesis Test**: Null hypothesis : the median score of the difference $d_i = X_{1i} - X_{2i}$ in order of $|d_i| = 0$
- Arrange the difference $d_i = X_{1i} - X_{2i}$ in order of $|d_i|$
- Discard differences with 0 value, and sample size reduce the number of zeros
- Assign ranks, R_i , from lowest to highest; for the same $|d_i|$ with different signs, give them the averaged rank
- Assign ranks the same signs as d_i.
- Sum '+' ranks (T_+) and '-' ranks (T-)
  - Test statistic is T_- or T_+ (one-tail test)
  - Test statistic is $T = smallerof T_-$ or $T_+ (two - tailtest)$

Under the null hypothesis, then the sum of the ranks, or rank sum (T), has the following properties

$$E(T) = \frac{n(n+1)}{4}, Var(T) = \frac{n(n+1)(2n+1)}{24}$$

where n is the number of nonzero differences If $n \geq 50$, and there are no ties, then a normal approximation can be used

$$z = \frac{T - E(T)}{\sigma_T} = \frac{|T - n(n+1)/4|}{\sqrt{n(n+1)(2n+1)/24}}$$

If n is large and there are ties, then the term $1/2$ in the computation of T serves as a continuity correction in the same manner as for the sign test

$$z = \frac{|T - n(n+1)/4 - 0.5|}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum_{i=1}^{g}(t_j^3 - t_j)}{48}}}$$

An alternative variance formula for $Var(T)$

$$Var(T) = \frac{\sum_{j=1}^{n} r_j^2}{4}$$

where t refers to the number of differences with the where r_j= rank of the absolute value same absolute value in the ith tied group and g is the of the jth observation number of tied groups Reject if

$$\mathrm{Z}_{\alpha/2} \leq \mathrm{Z}_0$$

Example

	Negative		Positive		Number of people with same absolute value	Range of ranks	Average rank
$\lvert d_i \rvert$	d_i	f_i	d_i	f_i			
10	−10	0	10	0		—	—
9	−9	0	9	0	0	—	—
8	−8	1	8	0	1	40	40.0
7	−7	3	7	0	3	37–39	38.0
6	−6	2	6	0	2	35–36	35.5
5	−5	2	5	0	2	33–34	33.5
4	−4	1	4	0	1	32	32.0
3	−3	5	3	2	7	25–31	28.0
2	−2	4	2	6	10	15–24	19.5
1	−1	4	1	10	14	1–14	7.5
		22		18			
0	0	5					

Figure 3.1: Wilcoxon Signed-Rank Test(Ties: no groups of differences with the same absolute value)(details in PPT)

Suppose we want to compare the effectiveness of two ointments (A, B) in reducing excessive redness in people who cannot otherwise be exposed to sunlight. Ointment A is randomly applied to either the left or right arm, and ointment B is applied to the corresponding area on the other arm. The person is then exposed to 1 hour of sunlight, and the two arms are compared for degrees of redness. Suppose instead the degree of burn can be quantified on a 10-point scale, with 10 being the worst burn and 1 being no burn at all.

$$E(T) = \frac{n(n+1)}{4} = \frac{40\times 41}{4} = 410$$
$$var(T) = \frac{n(n+1)(2n+1)}{24} - \frac{\sum_{i=1}^{g}(t_j^3 - t_j)}{48}$$

$$= \frac{40\times 41\times 81}{24} - \frac{(14^3-14)+(10^3-10)+\ldots+(1^3-1)}{48} = 5449.75$$

If the alternative variance formula is used, then

$$var(T) = \frac{\sum_{j=1}^{n} r_j^2}{4} = \frac{14\times 7.5^2 + 10\times 19.5^2 + \ldots + 40^2}{4} = 5449.75$$

$Thus \sqrt{var(T)} = \sqrt{5449.75} = 73.82$

$$z = \frac{|T - n(n+1)/4| - 0.5}{sd(T) = \sqrt{var(T)}} = \frac{|248 - 410| - 0.5}{73.82} = 2.19$$

The p -value of the test is given by

$$P = 2\times[1 - \Phi(2.19)] = 2\times(1 - 0.9857) = 0.029$$

3.1.3 Mann-Whitney U Test: Comparing two independent populations

Wilcoxon Rank Sum Test

- Tests two independent population probability distributions.
- Corresponds to t-test for two independent means.
- Can use normal approximation if $n_i > 10$.
- The original data from two independent samples are transformed into their ranks. It tests whether the two population medians are equal or not.
- The two samples are temporarily combined, and the ranks of the combined data values are calculated. Then the smaller sum of the ranks is used to calculate the test statistic.
- The null and alternative hypotheses:
  - H_0 : The distributions of two populations are identical
  - H_1 : The two population distributions are not identical

Favorable outcome		Poor outcome	
NIHSS score	Rank	NIHSS score	Rank
1	1	2	4.5
1	2	3	8
2	4.5	4	10
2	4.5	5	11
2	4.5	9	14
3	8	10	16
3	8	10	16
7	12	11	18
8	13	11	19
10	16	11	20
		12	21
		12	22
n_1=10	T_1=73.5	n_2=12	T_2=179.5

Figure 3.2: the NIHSS scores of 10 AIS patients with good outcome and 12 with poor outcome

Preliminary Steps of the Test

- Scores are ranked in ascending order, irrespective of which experimental group they come from (combine the data from the two groups);
- Tied scores take the mean of the ranks they occupy;
- Sum the ranks in the two groups;
- T=T with smaller sample size.

Steps of the Test

(1) State the hypotheses.

(2) Find the Test Statistic.

(3) Find the critical value.

(4) Conclusion

The Mann and Whitney U-test

1.Compare x_i with y_i.
u_1 = number of pairs $x_i > y_i$
u_2 = number of pairs $x_i < y_i$
and $u_1 + u_2 = n_1 n_2$.
2. Reject H_0 if u_1 is large or u_2 is small. Two Test Statistics are related as follows:

$$u_1 = w_1 - \frac{n_1(n_1+1)}{2},$$
$$u_2 = w_2 - \frac{n_2(n_2+1)}{2}$$

Advantage of the Mann-Whitney test:
Same distribution (whether u_1 or u_2) & Distribution range : $[0, n_1 n_2]$

P - value $= P\{U \geq u_1\} = P\{U \leq u_2\}$

At significant level α , we reject H_0 if P -value $\leq \alpha$ or $u_1 \geq u_{n_1,n_2,\alpha}$. Denote: $u_{n_1,n_2,\alpha}$- the upper α critical point
For large $n_1 and n_2$, the null distribution of U is Normal distributed. $U = \min(U_1, U_2)$

$$E(U) = \frac{n_1 n_2}{2}, Var(U) = \frac{n_1 n_2 (N+1)}{12}$$

Z-test(Large Sample)

$$Z = \frac{u_1 - \frac{n_1 n_2}{2} - \frac{1}{2}}{\sqrt{\frac{n_1 n_2 (N+1)}{12}}} = \frac{u_1 - E(U) - \frac{1}{2}}{\sqrt{Var(U)}}$$

We reject H_0 at significant level α , if $z \geq z_\alpha$

$$u_1 \geq \frac{n_1 n_2}{2} + \frac{1}{2} + z_\alpha \sqrt{\frac{n_1 n_2 (N+1)}{12}} \approx u_{n_1,,n_2,\alpha}$$

Two-sided test, Test Statistics:

$$u_{max} = \max(u_1, u_2)$$
$$u_{min} = \min(u_1, u_2)$$

P-value $= 2P\{U \geq u_{max}\} = 2P\{U \leq u_{min}\}$

Example

Failure Times of Capacitors (Wilcoxon-Mann-Whitney Test) 18 capacitors, 8 under control group and 10 under stressed group Perform the Wilcoxon-Mann-Whitney test to determine if thermal stress significantly reduces the time to failure of capacitors. $\alpha = 0.05$.

$$n_1 = 8 \quad n_2 = 10$$

The rank sums are

$$w_1 = 4 + 8 + 10 + 11 + 13 + 14 + 17 + 18 = 95$$
$$w_2 = 1 + 2 + 3 + 5 + 6 + 7 + 9 + 12 + 15 + 16 = 76$$
$$u_1 = w_1 - \frac{n_1(n_1+1)}{2} = 95 - \frac{(8)(9)}{2} = 59$$

Times to Failure for Two Capacitor Groups			
Control Group		Stressed Group	
5.2	17.1	1.1	7.2
8.5	17.9	2.3	9.1
9.8	23.7	3.2	15.2
12.3	29.8	6.3	18.3
		7.0	21.1

Ranks of Times to Failure			
Control Group		Stressed Group	
4	13	1	7
8	14	2	9
10	17	3	12
11	18	5	15
		6	16

Figure 3.3: Example data

$$u_2 = w_2 - \frac{n_2\,(n_2+1)}{2} = 76 - \frac{(10)(11)}{2} = 21$$

n_1	n_2	w_1	u_1	$P(W \geq w_1)=P(U \geq u_1)$
8	8	84	48	0.052
	8	87	51	0.025
	8	90	54	0.010
	8	92	56	0.005
	9	89	53	0.057
	9	93	57	0.023
	9	96	60	0.010
	9	98	62	0.006
	10	95	59	0.051
	10	98	62	0.027
	10	102	66	0.010
	10	104	68	0.006

Figure 3.4: Upper-Tail Probabilities of the Null Distribution of the Wilcoxon-MannWhitney Statistic

$$H_0 : F_1 = F_2 vs. H_1 : F_1 < F_2$$

Let F_1 be c.d.f of the control group. F_2 be c.d.f of the stressed group. Check that $u_1 + u_2 = n_1 n_2 = 80$. From Table P-Value =0.051 Large sample Z-test:

$$Z = \frac{u_1 - n_1 n_2/2 - 1/2}{\sqrt{\frac{n_1 n_2 (N+1)}{12}}}$$
$$= \frac{59 - (8)(10)/2 - 1/2}{\sqrt{\frac{(8)(10)(19)}{12}}} = 1.643$$

Conclusion:yields the P-Value= 1- $\Phi(1.643)$ =0.0502

3.1.4 Kruskal-Wallis test: Completely Randomized Design

Kruskal-wallis h-test

- Tests the equality of more than two (p) population probability distributions.
- Corresponds to ANOVA for more than two means.
- Used to analyze completely randomized experimental designs.
- χ^2 distribution with p – 1 df — if sample size $n_j \geq 5$

Analysis procedure

- State Hypothesis
  - H_0 All of the population medians are all equal.
  - H_1 Not all of the population medians are equal.
- Temporarily combine the three samples and arrange them in increasing order.
- Rank the data values from smallest to largest. Resolve ties using the mean rank.
- Calculate the sum of the ranks for each group, R1, R2, and R3.
- Calculate H (the test statistic).

Total variability of ranks (TV), Variability between group (BV), Variability within group (WV).
average rank $= \frac{1+2+\cdots+n}{n} = \frac{n+1}{2}$ The average of the total sum of squares of

$$rank(\text{aTV}) = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(R_{ij} - \frac{n+1}{2}\right)^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n_i}\left(i - \frac{n+1}{2}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n_i} i^2 - \frac{n(n+1)^2}{4}\right)$$

$$= \frac{1}{n-1}\left(\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4}\right) = \frac{n(n+1)}{12}$$

$$BV = \sum_{i=1}^{k} n_i\left(\frac{R_{i.}}{n_i} - \frac{n+1}{2}\right)^2$$

$$KW = \frac{BV}{aTV} = \frac{12}{n(n+1)}\sum_{i=1}^{k} n_i\left(\frac{R_i}{n_i} - \frac{n+1}{2}\right)^2 = \frac{12}{n(n+1)}\sum_{i=1}^{k} R_i\frac{R_i}{n_i} - 3(n+1)$$

R_i and n_i is the rank sum and sample size of the ith group, respectively

Example

To study the effect of arginine on lymphocyte transformation function in mice after amputation, 21 mice were divided into 3 groups: control group A, amputation group B, amputation +arginine treatment group C. Then 3H absorb value (CPM) was measured to show the spleen lymphocyte proliferation stimulated by heparin enzyme (HPA) . The measurement is shown in Table 4.1, try to analyze whether the spleen lymphocyte proliferation in these 3 groups are different. (The population variance are not equal)

Group A (1)	Rank (2)	Group B (3)	Rank (4)	Group C (5)	Rank (6)
3012	11	2532	8	8138	15
9458	18	4682	12	2073	6
8419	16	2025	5	1867	4
9580	19	2268	7	885	2
13590	21	2775	9	6490	13
12787	20	2884	10	9003	17
6600	14	1717	3	690	1
R_i	$R_1 = 119$		$R_2 = 54$		$R_3 = 58$
n_i	7		7		7

Figure 3.5: The spleen lymphocyte proliferation stimulated by heparin enzyme (HPA) measured as 3H absorb value

$$H(KW) = \frac{12}{21(21+1)}\left(\frac{119^2}{7} + \frac{54^2}{7} + \frac{58^2}{7}\right) - 3(21+1) = 9.848$$

Find the critical and state conclusion

- If group $k = 3$, and each group has $n_{\mathrm{i}} \leq 5$, check H table;
- If each $\mathrm{n_i} > 5$, H or H_c is approximately distributed as a χ^2 with $\mathrm{v = k-1}$, check χ^2 table

In this case, $n_{\mathrm{i}} = 7 > 5$, then check χ^2 table, $\chi^2_{0.05,2} = 5.99$, $H = 9.848 > \chi^2_{0.05,2} = 5.99$, therefore, P¡0.05 , reject H_0 , not all of the population medians are identical.

3.1.5 Friedman test: Randomized Block Design repeated measures

The Friedman test is a nonparametric version of the randomized block design ANOVA or repeated measured ANOVA. Sometimes this design is referred to as a two-way ANOVA with one item per cell because it is possible to view the blocks as one factor and the treatment levels as the other factor. The test is based on ranks.

Logic and procedure

The basic idea of Friedman's rank sum test is that the observed values in each block group are ranked in ascending order; If the effects of all treatments are the same, the rank 1, 2,... k (k is the number of treatment groups) should appear in each treatment group (column) with equal probability, and the rank sum of each treatment group should be roughly equal, and it is unlikely to be significantly different. If the sample rank of each treatment group and R_1, R_2,... vary greatly, it is reasonable to doubt whether the overall distribution is the same among treatment groups.
Rank order the scores SEPARATELY FOR EACH SUBJECT'S DATA with the smallest score getting a value of

(1) If there are ties (within the scores for a subject) each receives the average rank they would have received;

(2) Compute the sum of the ranks for each condition;

(3) Compute Friedman's M

$$M = \sum_{i=1}^{k} \left(R_i - \bar{R}\right)^2$$

k is the number of groups or treatment levels
FOR LARGE SAMPLES (K¿5 OR B¿13) Approximate chi-square method

$$x^2 = \frac{12}{bk(k+1)} \sum_{i=1}^{k} R_i^2 - 3b(k+1)$$

Look at the χ^2 table with $k-1$ degrees of freedom
If there are too many ties

$$\chi_c^2 = \frac{\chi^2}{c} \quad C = 1 - \frac{\sum \left(t_j^3 - t_j\right)}{bk\left(k^2-1\right)}$$

- where R_i is the sum of the ranks for sample i.
- where t_j is the number of j_{th} observation with the same rank in each block group.
- b is the number of independent blocks.
- k is the number of groups or treatment levels

Example

A university uses students' comprehensive scores to evaluate the teaching effect of courses. Now 10 medical students are randomly selected to evaluate the teaching effect of three basic medical courses, as shown in Table 5.1, and try to compare whether the teaching effect of these three basic medical courses is the same.

student	Anatomy		Physiology		Histoembryology	
	score	rank	score	rank	score	rank
1	4.0	1.5	4.0	1.5	5.0	3
2	2.5	1	4.0	2.5	4.0	2.5
3	4.0	2	3.5	1	4.5	3
4	3.5	1	4.0	2	5.0	3
5	3.5	2	3.0	1	4.0	3
6	2.5	1	3.5	2.5	3.5	2.5
7	4.0	3	3.5	1.5	3.5	1.5
8	3.5	1.5	3.5	1.5	4.5	3
9	3.0	1	4.0	2.5	4.0	2.5
10	2.5	1	3.0	2	4.0	3
R_i		15		18		27

Figure 3.6: Comparison of the comprehensive scores of 10 medical students on the teaching effects of three basic medical courses

$$\bar{R} = \frac{15+18+27}{3} = 20$$
$$M = \sum_{i=1}^{k} \left(R_i - \bar{R}\right)^2 = (15-20)^2 + (18-20)^2 + (27-20)^2 = 78$$

Look at the M table, determine the critical value, $M_{0.05} = 62$ (when k=3, b=10)
For this example: $b = 10, k = 3, R1 = 15, R2 = 18, R3 = 27$

$$\chi^2 = \frac{12}{10\times3(3+1)} \sum_{i=1}^{k} \left(15^2 + 18^2 + 27^2\right) - 3 \times 10(3+1) = 7.8$$
$$C = 1 - \frac{\sum\left(t_j^3 - t_j\right)}{bk(k^2-1)} = 1 - \frac{1}{10\times3(3^2-1)} \left[\left(2^3-2\right) + \left(2^3-2\right) + \ldots. + \left(2^3-2\right)\right] = 0.85$$
$$\chi_c^2 = \frac{\chi^2}{c} = 9.18, \chi_{0.05,2}^2 = 5.99$$

3.1.6 Rank correlation(Spearman's Rank Correlation Kendall rank correlation)

The Spearman's rank-order correlation is the nonparametric version of the Pearson product-moment correlation. Spearman's correlation coefficient, (ρ, also signified by r_s) measures the strength and direction of association between two ranked variables

3.1.7 Spearman's Rank Correlation procedure

- Assign ranks for sleep and separately;

- when you have two identical values in the data (called a "tie"), you need to take the average of the ranks that they would have otherwise occupied.
- Calculate differences, d_i , between each pair of ranks
- Square differences, d_i^2 between ranks
- Sum squared differences for each variable
- Use shortcut approximation formula

Similar to Pearson correlation, spearman rank-coefficient can be calculated as The formula for when there are no tied ranks is:

$$r_s = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} = \frac{\sum\left(P_i - \bar{P}\right)\left(Q_i - \bar{Q}\right)}{\sqrt{\sum\left(P_i - \bar{P}\right)^2 \sum\left(Q_i - \bar{Q}\right)^2}} \quad r_s = 1 - \frac{6\sum d^2}{n\left(n^2-1\right)}$$

The formula to use when there are tied ranks is:

$$r_s' = \frac{\left[\left(n^3 - n\right)/6\right] - \left(T_x + T_y\right) - \sum d^2}{\sqrt{\left[\left(n^3 - n\right)/6\right] - 2T_x}\sqrt{\left[\left(n^3 - n\right)/6\right] - 2T_y}}$$

T_x (*or* T_y) $= \sum\left(t^3 - t\right)/12$ Where t is the number of $x(or\ y)$ with the same value where

$\mathrm{d_i}$ = difference in paired ranks and n = number of cases.

Example

A project team conducted a survey on sleep and anxiety among sophomores in one high school in Suzhou, Jiangsu Province. In this project, the Pittsburgh Sleep Quality Index Scale (PSQI) and Student Anxiety Scale (SAS) were used to investigate the sleep and anxiety status. Please calculate the Spearman level correlation coefficient rsbetween sleep and anxiety.

	Sleep quality		Anxiety		Rank				
id	score x	rank P_i	score y	rank Q_i	difference $d=P_i-Q_i$	d^2	P_i^2	Q_i^2	$P_i\ Q_i$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	4	6.0	5	7.0	-1.0	1.00	36.00	49.00	42.00
2	8	9.5	8	10.0	-0.5	0.25	90.25	100.00	95.00
3	5	7.0	6	8.0	-1.0	1.00	49.00	64.00	56.00
4	8	9.5	4	5.5	4.	16.00	90.25	30.25	52.25
5	3	4.0	4	5.5	-1.5	2.25	16.00	30.25	22.00
6	2	1.5	3	3.5	-2.0	4.00	2.25	12.25	5.25
7	3	4.0	3	3.5	0.5	0.25	16.00	12.25	14.00
8	7	8.0	7	9.0	-1.0	1.00	64.00	81.00	72.00
9	3	4.0	0	1.0	3.0	9.00	16.00	1.00	4.00
10	2	1.5	1	2.0	-0.5	0.25	2.25	4.00	3.00
Total	45	55.0	41	55.0	-	35.00	382.00	384.00	365.50

Figure 3.7: scores for sleep quality and anxiety among 10 students

$$r_s = \frac{\sum(P_i-\bar{P})(Q_i-\bar{Q})}{\sqrt{\sum(P_i-\bar{P})^2\sum(Q_i-\bar{Q})^2}}$$
$$= \frac{365.5-55\times55/10}{\sqrt{(382-55^2/10)(384-55^2/10)}} = \frac{63}{\sqrt{79.5\times81.5}} = 0.783$$
$$r_s = 1-\frac{6\sum d^2}{n(n^2-1)} = 1-\frac{6\times35.00}{10\times(100-1)} = 0.788$$
$$r_s' = \frac{\left[\left(n^3-n\right)/6\right]-(T_x+T_y)-\sum d^2}{\sqrt{[(n^3-n)/6]-2T_x}\sqrt{[(n^3-n)/6]-2T_y}} = \frac{\left[\left(10^3-10\right)/6\right]-(3+1)-35.00}{\sqrt{[(10^3-10)/6]-2\times3}\sqrt{[(10^3-10)/6]-2\times1}} = 0.783$$

Hypothesis test : Two-Tailed Test $H_0 : \rho = 0 H_a : \rho \neq 0$ t Test for Spearman Rank Correlation
Compute the test statistic

$$t_s = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

which under the null hypothesis of no correlation follows a t distribution with $n-2$ degrees of freedom.

For a two-sided level α test, if $t_s > t_{n-2,1-\alpha/2}$ or $t_s < t_{n-2,\alpha/2} = -t_{n-2,1-\alpha/2}$ then reject H_0 ; otherwise, accept H_0.

The exact p -value is given by
$p = 2 \times$ (area to the left of t_s under a t_{n-2} distribution) if $t_s < 0$
$p = 2 \times$ (area to the right of t_s under a t_{n-2} distribution) if $t_s \geq 0$
This test is valid only if $n \geq 10$.

3.2 Tests For Categorical Data

3.2.1 Chi-square goodness of fit

A Pearson goodness-of-fit test, in general, refers to measuring how well do the observed data correspond to the fitted (assumed) model.
$X_1, X_2, X_3, \ldots X_n$ is a random sample from unknow distribution $F(X), F_0(X)$ is a known distribution. HO : $F(X) = F_0(X); H1 : F(X) \neq F_0(X)$.
Or we can test whether F(X) follow a certain type of distribution(i.e., normal distribution)

$$H0 : F(X) = \mathrm{F}_0(X;\theta);$$
$$H1 : F(X) \neq \mathrm{F}_0(X;\theta)$$

$$\mathrm{H}_0 : \mathrm{p}_1 = \pi_1, \quad \mathrm{p}_2 = \pi_2, \mathrm{p}_3 = \pi_3, \ldots\ldots\ldots \mathrm{p}_\mathrm{k} = \pi_k$$

Pearson Goodness-of-fit Test Statistic

$X^2 = \sum_{j=1}^{k} \frac{(X_j - n\pi_j)}{n\pi_j}$ An easy way to remember it is $X^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)}{E_j}$

Where $O_j = Xj$ is the observed count in cell j , and $Ej = E(Xj) = n\pi j$ is the expected count in cell j under the assumption that null hypothesis is true X_2 measure how closely the model, have an approximate chi-square distribution with k − 1 degrees of freedom when H_0 is true.

$$X^2 \sim \chi^2_{k-1}$$

If the sample proportions p_j are close to the model's π_j , then $O_j \to E_j$, and $X^2 \to 0$.

Proof for df = 1 Suppose a binomial distribution has 2 outcomes $(1, 2)$, of which the probabilities are $\pi_1 and \pi_1$, respectively. $\pi_2 = 1 - \pi_1$ Suppose we draw a sample from this binomial distribution, the sample size is n , the number of outcome 1 and 2 is vl and v2, respectively. According to Pearson Goodness-of-fit Test Statistic

$$X^2 = \frac{(v_1 - n\pi_1)^2}{n\pi_1} + \frac{(v_2 - n\pi_2)^2}{n\pi_2} = \frac{(v_1 - n\pi_1)^2}{n\pi_1} + \frac{[(n - \pi_1) - (n - n\pi_1)]^2}{n(1 - \pi_1)}$$

$$= \frac{(v_1 - n\pi_1)^2}{n\pi_1(1 - \pi_1)} = \left[\frac{v_1 - n\pi_1}{\sqrt{n\pi_1(1 - \pi_1)}}\right]^2$$

De Moivre-Laplace

If $n \to +\infty$, it tends to follow $N(0, 1)$

Extension, H_0 : pi $= \pi i(\theta), i = 1, , k, where \theta$ is unknown parameter Pearson Goodness-of-fit Test Statistic

Use Maximum Likelihood Estimation to estimate θ

Pearson-Fisher theorem

$$X^2 = \sum_{i=1}^{k} \frac{(X_i - n\hat{\pi}_j)}{n\hat{\pi}_i}, \text{ and } X^2 \sim \chi^2_{k-r-1} if n \to \infty$$

where k = the number of groups and r= the number of parameters estimated from the data to compute the expected frequencies

Example

We already known that the height (in cm) of 12-year-old boy in Suzhou follows N(139.48,7.32). Suppose we randomly selected 120 boys aged 12 from Suzhou Industrial Park, and created a frequency table, as follows

We would like to test whether these measurements came from $N(139.48, 7.32)$

128.10	144.40	150.30	146.20	140.60	139.70	134.10	124.30	147.90	143.00
142.70	126.00	125.60	127.70	154.40	142.70	141.20	133.40	131.00	125.40
146.30	146.80	142.70	137.60	136.90	122.70	131.80	147.70	135.80	134.80
139.00	132.30	134.70	138.40	136.60	136.20	141.60	141.00	138.40	145.10
139.90	140.60	140.20	131.00	150.40	142.70	144.30	136.40	134.50	132.30
148.10	139.60	138.90	136.10	135.90	140.30	137.30	134.60	145.20	128.20
140.20	136.60	139.50	135.70	139.80	129.10	141.40	139.70	136.20	138.40
132.90	142.90	144.70	138.80	138.30	135.30	140.60	142.20	152.10	142.40
136.20	135.00	154.30	147.90	141.30	143.80	138.10	139.70	127.40	146.00
141.20	146.40	139.40	140.80	127.70	150.70	157.30	148.50	147.50	138.90
126.00	150.00	143.70	156.90	133.10	142.80	136.80	133.10	144.50	142.40
143.10	130.30	139.10	141.40	152.70	135.90	138.10	142.70	155.80	123.10

Height	observed value
122-	5
126-	8
130-	10
134-	22
138-	33
142-	20
146-	11
150-	6
154-	5
total	**120**

Figure 3.8: Students's table

height (1)	Observed frequencies (2)	$\Phi\left(\frac{l_i-\mu}{\sigma}\right)$ (3)	$\Phi\left(\frac{u_i-\mu}{\sigma}\right)$ (4)	P_i (5)=(4) - (3)	Expected frequencies (6)	$\frac{(O_i-E_i)^2}{E_i}$ (7)
122.0~	5	0.00832	0.03240	0.02408	2.8900	1.54053
126.0~	8	0.03240	0.09704	0.06463	7.7557	0.00769
130.0~	10	0.09704	0.22642	0.12939	15.5263	1.96698
134.0~	22	0.22642	0.41967	0.19325	23.1898	0.06104
138.0~	33	0.41967	0.63503	0.21536	25.8433	1.98188
142.0~	20	0.63503	0.81411	0.17908	21.4898	0.10328
146.0~	11	0.81411	0.92522	0.11111	13.3331	0.40827
150.0~	6	0.92522	0.97665	0.05143	6.1717	0.00477
154.0~	5	0.97665	0.99441	0.01776	2.1309	3.86289
Total	120			—	—	9.93733

Figure 3.9: Statisic procedure

This assumption can be tested by first computing what the expected frequencies would be in each group if the data did come from an underlying normal distribution and by then comparing these expected frequencies with the corresponding observed frequencies

3.2.2 Two sample test for binomal proportions

Normal-Theory Method (samples large enough, $[n\pi(1-\pi) > 5]$)
Under $\mathrm{H}_0 : \pi_1 = \pi_2 = \pi$

$$p_1 \sim N(\pi, \pi(1-\pi)/\mathrm{n}_1)$$
$$\mathrm{p}_2 \sim N(\pi, \pi(1-\pi)/\mathrm{n}_2)$$

samples are independent. The best estimator for π is based on a weighted average of the sample proportions $\mathrm{p}_1 and \mathrm{p}_2$. then under H_0 ,

$$\mathrm{z} = \frac{p_1 - p_2}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1) \quad \hat{\pi} = \frac{n_1 p1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

$$\mathrm{Z} = \frac{|p_1 - p_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{\pi}(1-\hat{\pi}))\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where } \hat{\pi} = \frac{X_1 + X_2}{n_1 + n_2}$$

For a two-sided level α test, The approximate p -value for this test is given by if $Z > Z_{1-\alpha/2}$ then reject H_0 ;

$$p = 2[1 - \Phi(z)]$$

if $Z \leq Z_{1-\alpha/2}$ then accept H_0 .

Example

Normal-Theory Method (samples large enough, [$n\pi(1-\pi) > 5$])

Drug group	effective yes	effective no	total	Rate of effectiveness(%)	
A	41 X1	4	45 (n1)	91.1	B(n1,π1)
B	24 X2	11	35 (n2)	68.8	B(n2,π2)
Total	65	15	80	81.3	

Figure 3.10: Statisic procedure

One doctor wants to study whether or not there is a difference for the treatment effectiveness of chronic pharyngitis between drug A group and drug B group. He randomly divided 80 homogenous patients suffering from chronic pharyngitis into drug A and B groups, and then observed the treatment effectiveness. The data is shown in Figure 16.
Solution 1 for example 2

$$\hat{\pi} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{65}{80} = 81.3\%, \hat{\pi} = 1 - \hat{\pi} = 18.7\%$$

$$n_1\hat{\pi}(1-\hat{\pi}) = 45 \times 0.813 \times 0.187 = 6.84 > 5$$
$$n_2\hat{\pi}(1-\hat{\pi}) = 35 \times 0.813 \times 0.187 = 5.32 > 5$$

$$\text{z} = \frac{|p1 - p2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{2}{n_2}\right)}}$$

$$= \frac{|0.911 - 0.688| - \left(\frac{1}{2\times 45} + \frac{1}{2\times 35}\right)}{\sqrt{0.813 \times 0.187 \times \left(\frac{1}{35} + \frac{1}{45}\right)}} = 25.47994$$

Contingency-Table Method

Drug group	effective		total	Rate of effectiveness(%)
	yes	no		
A	41	4	45 (row margins)	91.1
B	24	11	35 (row margins)	68.8
Total	65 (column margins)	15 (column margins)	80 (grand total)	81.3

Figure 3.11: Data of effect treated by drug A and drug B

Step1 : **Hypothesis**

- π_1 = The probability or population proportion of effectiveness treated by drug A
- π_2 = the probability or population proportion of effectiveness treated by drug B

$H_0 : \pi_1 = \pi_2$, the difference between p_1 and p_2 is caused by sampling error
$H_1 : \pi_1 \neq \pi_2$, the difference between p_1 and p_2 is caused by sampling error+ treatment effect

Drug group	effective		total	Rate(%)
	yes	no		
A	41(36.6)	4(8.4)	45	91.1
B	24(28.4)	11(6.6)	35	68.8
Total	65	15	80	81.3

➢**Assume H_0:π_1= π_2 =π is true, the best estimate of common rate π is p_c, which is 81.3%.**

Figure 3.12: Expected table for the data in example 1

$$E_{11} = n_1 p_c = 45 \times 81.3\% = 45 \times (65/80) = 36.6$$
$$E_{12} = n_1 (1 - p_c) = 45 \times (100\% - 81.3\%) = 45 \times (15/80) = 8.4$$
$$E_{21} = n_2 p_c = 35 \times 81.3\% = 35 \times (65/80) = 28.4$$
$$E_{22} = n_2 (1 - p_c) = 35 \times (100\% - 81.3\%) = 35 \times (15/80) = 6.6$$

$$E_{rc} = \frac{n_r n_c}{N}$$
n_r correpsonding row margin
n_c correpsonding column margin

How to compare the difference between actual number (or observed number) and theoretical number(or expected number)?

$$\sum \frac{(O - E)^2}{E} \sim \chi^2_{\alpha/2,v}$$

$$df = \mathrm{v} = (\mathrm{r} - 1)(\mathrm{c} - 1), \text{or } \mathrm{df} = \mathrm{k} - 1 - \mathrm{r}, \text{where } \mathrm{k} = \mathrm{rc}$$

- If this sum is small, H_0 is not rejected, because small value of this sum means good agreement between the actual cells and the theoretical cells.
- If this sum is large, H_0 is rejected.

Step 2 Calculate the χ^2 value

$$\chi^2 = \sum \frac{(A-T)^2}{T} = \frac{(41-36.6)^2}{36.6} + \frac{(4-8.4)^2}{8.4} + \frac{(24-28.4)^2}{28.4} + \frac{(11-6.6)^2}{6.6} = 6.565$$
$$v = (r - \mathbf{1})(c - \mathbf{1}) = (\mathbf{2} - \mathbf{1})(\mathbf{2} - \mathbf{1}) = \mathbf{1}$$

Estimate the P value, and draw a conclusion
Step 3

$$\chi^2_{0.05,1} = 3.84, 6.565 > 3.84, P < 0.05$$

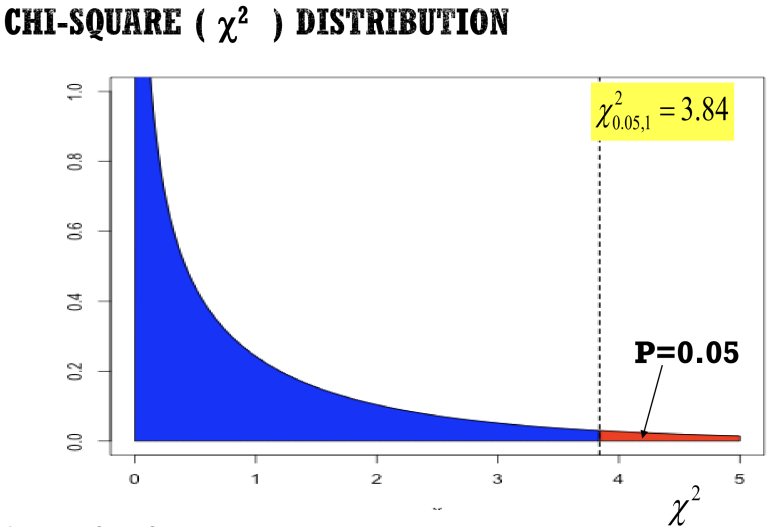


Figure 3.13: χ^2 distribution

co Because P¡0.05 , according to the significance level $\alpha = 0.05$, we have evidence to refuse the H_0 , and accept the H_1. *co*
We can draw a conclusion that the population rate of treatment effectiveness for drug A does not equal to the population rate for drug B.

Short computational form for chi-square test for 2×2 contingency table

group	outcome		total
	+	-	
A	*a*	*b*	*a*+*b*
B	*c*	*d*	*c*+*d*
total	*a*+*c*	*b*+*d*	*n*=*a*+*b*+*c*+*d*

Figure 3.14: General contingency table

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

Drug Group	effective		total
	yes	no	
A	41	4	45
B	24	11	35
Total	65	15	80

Figure 3.15: We use the short computational form to calculate chi-square in example

$$\chi^2 = \frac{(41 \times 11 - 4 \times 24)^2 \times 80}{45 \times 35 \times 65 \times 15} = 6.565$$

Yates-corrected Chi-square test for a 2×2 contingency table

If any of four cells is : $1 \leq T < 5$ and $n \geq 40$, we should compute the **Yates-corrected chi-square**.

$$^2 = \sum \frac{(|O-E|-0.5)^2}{E}$$
$$\chi^2 = \frac{\left(|ad-bc|-\frac{n}{2}\right)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

Example

Cause of death	Type of diet		Total
	High salt	Low salt	
Non-CVD	2 (2.92)	23 (22.08)	25
CVD	5 (4.08)	30 (30.92)	35
Total	7	53	60

Figure 3.16: data concerning the possible association between cause of death and high salt intake

Suppose we want to investigate the relationship between high salt intake and death from cardiovascular disease (CVD). A retrospective study is done among men ages 50-54 in a specific county who died over a 1-month period. The investigators try to include approximately an equal number of men who died from CVD (the cases) and men who died from other causes (the controls). Of 35 people who died from CVD, 5 were on a high-salt diet, whereas of 25 people who died from other causes, 2 were on such a diet

$$T_{11} = 257/60 = 2.92,\ T_{12} = 357/60 = 4.08$$

We find T_{11} and T_{12} are between 1 and 5, so, the Yates-corrected Chi-square test should be used to compare difference between two groups.

- Hypothesis
  - H_0: $\pi_1 = \pi_2$, the population proportion of high-salt intake among people who died from CVD equals to proportion high-salt intake among people who died from other diseases
  - $H_1 : \pi_1 \neq \pi_2$, two population proportions don't equal to each other $\alpha = 0.05$
- Calculate the corrected chi-square

$$\chi^2 = \frac{(|2 \times 30 - 23 \times 5| - 60/2)^2 \times 60}{25 \times 35 \times 7 \times 53} = 0.116 \quad v = (2-1)(2-1) = 1$$

- Conclusion

$$\chi^2 = 0.116 < 3.84, P > 0.05,\ \text{we can't reject } H_0.$$

Summary
We Have Several Requirements For The Fourfold Table

- If $T \geq 5$ and $n \geq 40$

$$\chi^2 = \sum \frac{(A-T)^2}{T} \quad \text{or} \quad \chi^2 = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

- If any $1 \leq T < 5$ and $n \geq 40$

$$\chi^2 = \sum \frac{(|A-T| - 0.5)^2}{T} \text{ or } \quad \chi^2 = \frac{\left(|ad-bc| - \frac{n}{2}\right)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

- If any $T < 1$ or $n < 40$, we should use Fisher's exact test to compute the P value

3.2.3 Fisher's exact test

When any of theoretical value is less than 1(T¡1) or n¡40, we need to use fisher's exact method to estimate the P value and make the statistical inference.

We can use R, SPSS or SAS to get the P value

Poison group	Death outcome		total	Rate of death(%)
	Yes	No		
A	1	9	10	10.0
B	5	5	10	50.0
Total	6	14	20	30.0

Figure 3.17: Comparison of mouse death rate for poison A and B

3.2.4 Two Sample Test ForBinomial Proportions ForMatched Data(McNemar's Test)

Matched pair design patterns :

- Two subjects in each pair receive different treatments (different subjects match), it needs to match the pair for the subjects by some conditions such as same age, sex, weight or others before randomizing
- One subject receives two different treatments (within-subjects design) Measuring the same objects at two different time points Different parts of the same object

Assumptions

- You must have one nominal variable with two categories (i.e. dichotomous variables) and one independent variable with two connected groups.
- The two groups in your the dependent variable must be mutually exclusive. In other words, participants cannot appear in more than one group.
- Your sample must be a random sample.

Matched pair design

Example The standard screening test for Down's syndrome is based on a combination of maternal age and the levels of serum AFP and Free $\beta - hCG$.

A new test (Non-invasive Prenatal Testing) is proposed that may be better or worse than the standard test. To assess their relative efficacy, both tests are used on the same subjects and compared with the true diagnosis. Let + equals to correct assessment, - equals to incorrect assessment. The results are given in Figure 24.

McNemar chi-square Test

- **Hypothesis** : Let the capital letter " B " denote population parameter of b , and C denote population parameter of c .

Standard test	New test	N
+	+	82
+	-	5
-	+	10
-	-	3

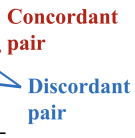


Figure 3.18: Comparison of two screening tests for Down's syndrome,There're four different combinations 85(82+3) pairs have the same outcomes, and 15(5+10) pairs have different outcomes.

Standard test	**New test**		**total**
	+	**-**	
+	**82(a)**	**5(b)**	**87**
-	**10(c)**	**3(d)**	**13**
total	**92**	**8**	**100**

Figure 3.19: Comparison of two screening tests for Down's syndrome, The same outcomes don't contribute to comparing the difference between two methods. If the new test and standard test are equally effective, an equal number of new test and standard test discordant pairs would be expected.

- **Compute test statistic** If $\mathbf{H}_0$ is true, the theoretical value =(b+c) / 2

  (1) When $b + c \geq 40$

$$\chi^2 = \frac{\left[b - \frac{b+c}{2}\right]^2}{b+c} + \frac{\left[c - \frac{b+c}{2}\right]^2}{b+c} = \frac{(b-c)^2}{b+c}$$

  (2) when b+c <40

$$\chi^2 = \frac{\left[\left|b - \frac{b+c}{2}\right| - \frac{1}{2}\right]^2}{\frac{b+c}{2}} + \frac{\left[\left|c - \frac{b+c}{2}\right| - \frac{1}{2}\right]^2}{\frac{b+c}{2}} = \frac{(|b-c|-1)^2}{b+c}$$

- For this example, $b + c$<40

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} = \frac{(|5-10|-1)^2}{5+10} = 1.07$$
$$\mathbf{v} = (\mathbf{2-1})(\mathbf{2-1}) = \mathbf{1}$$

- **Conclusion** $\chi^2 = 1.07 < 3.84, \mathrm{P} > 0.05, \mathrm{H}_0$ can not be rejected.

3.2.5 R×C contingency table

An R×C contingency table is a table with R rows and C columns. It displays the relationship between two variables, where the variable in the rows has R categories and the variable in the columns has C categories. The number of R or C is more than 2.

Example1

Obesity is an important risk factor for many diseases. However, in studying the effects of obesity, it is important to be aware of other risk factors that may be potentially related to obesity. One commonly used measure of obesity is body mass index (BMI) (kg/m^2),which is often categorized as follows

normal = BMI <25, overweight = BMI 25.0-29.9and obesity = $BMI \geq 30.0$. The data in table 8 were got from a study relating education to BMI categories. What test can be used to compare the percentages of individuals with at least a high-school education in the three BMI groups?
Generalizing our experience from the 2×2 situation, the expected number for every cell can be formed in the same way as 2×2 table

- **Hypothesis**
  - $H_0 : \pi_1 = \pi_2 = \pi_3$, Population percentages of people with at least high school education are same among three BMI groups
  - H_1 : At least two groups' π are different $\alpha = 0.05$.

BMI category	≥High-school education	<High-school education	Total	%≥High-school education
Normal	70	7	77	90.9
Overweight	105	15	120	87.5
Obese	53	11	64	77.1
total	228	33	261	87.36

Figure 3.20: Relationship between BMI category and education level (n=261)

- **Compute test statistic**

$$E_{11} = (77 \times 228)/261 = 67.26$$
$$E_{12} = (77 \times 33)/261 = 9.74$$
$$E_{21} = (120 \times 228)/261 = 104.83$$
$$E_{22} = (120 \times 33)/261 = 15.17$$
$$E_{31} = (64 \times 228)/261 = 55.91$$
$$E_{32} = (64 \times 33)/261 = 8.09$$

Basic formula

$$\chi^2 = \sum \frac{(O-E)^2}{O} = \frac{(70-67.26)^2}{67.26} + \frac{(7-9.74)^2}{9.74} + \frac{(105-104.83)^2}{104.83} + \frac{(15-15.17)^2}{15.17} + \frac{(53-55.91)^2}{55.91} + \frac{(11-8.09)^2}{8.09} = 2.08$$

The special formula

$$\chi^2 = n\left[\sum \frac{O^2}{n_r n_c} - 1\right] = 261\left[\frac{70^2}{77\times228} + \frac{7^2}{77\times33} + \frac{105^2}{120\times228} = \frac{15^2}{120\times33} + \frac{53^2}{64\times228} + \frac{11^2}{64\times33} - 1\right] = 2.08$$
$$v = (3-1)(2-1) = 2$$

Conclusion

$$\chi^2 = 2.08 < \chi^2_{0.05,2} = 5.99, \mathrm{P} > 0.05, \mathrm{H}_0 \text{ can not be rejected.}$$

Example2

One doctor wants to know whether there is difference in the proportions distribution of blood type of acute lymphatic leukemia (ALL) patients between children and adults. He collected some data in Figure 27.

group	A(%)	B(%)	O(%)	AB(%)	Total(%)
children	30(26.8)	38(33.9)	32(28.6)	12(10.7)	112(100.0)
adults	19(24.7)	30(39.0)	19(24.7)	9(11.7)	77(100.0)
total	49	68	51	21	189

Figure 3.21: Blood type distribution between children and adults

- $\Rightarrow$ H_0 : the population proportions of the blood type are equal for children and adults ALL patients.
- $\Rightarrow$ H_1 : the population proportions of the blood type are not equal for children and adults ALL patients

$$\alpha = \mathbf{0.05}$$

$$\chi 2 = n\left(\sum \frac{O^2}{n_r \cdot n_c} - 1\right) = 189 \times \left(\frac{30^2}{112\times 49} + \frac{38^2}{112\times 68} + \frac{32^2}{112\times 51} + \frac{12^2}{112\times 21} + \frac{19^2}{77\times 49} + \frac{30^2}{77\times 68} + \frac{19^2}{77\times 51} + \frac{9^2}{77\times 71} - 1\right) = 0.695$$

$$v = (r-1)(c-1) = (4-1)(2-1) = 3$$

$$\text{Conclusion } \chi^2_{0.05,3} = 7.81, 0.695 < 7.81, P > 0.05$$

We can't reject $\mathbf{H}_0$, there is no significant difference in proportions of blood between children and adults ALL patients, which means the blood type does not relate to age of ALL patients.
$\star$ Cautions in Chi-square test for R $\times$ C table Use this test only if both of the following two conditions are satisfied:

(1) No more than 1 / 5 of the cells have expected values < 5 .

(2) No cell has a theoretical or an expected value < 1 .

If these two conditions are not satisfied, there is no continuity correction, we can use Fisher's exact test to compare differences among groups.

3.2.6 Correlation analysis of categorical variable data

Example

There are standard screening test and new test for Down's syndrome.
Phi coefficient

$$r_n = \frac{(ad-bc)}{\sqrt{(a+b)(a+c)(c+d)(b+d)}} = \frac{82\times 3 - 5\times 10}{\sqrt{87\times 13\times 92\times 8}} = 0.214826$$

Standard test	New test		total
	+	-	
+	**82(a)**	**5(b)**	**87**
-	**10(c)**	**3(d)**	**13**
total	**92**	**8**	**100**

Figure 3.22: Comparison of two screening tests for Down's syndrome

Hypothesis test, H_0: $r_n = 0$

$$\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} = \frac{(82\times 3 - 5\times 10)^2 \times 100}{87\times 13\times 92\times 8} = 4.615$$

$\chi^2 = 4.615 > 3.84$, therefore the results of two screening tests is connected

KAPPA STAITSIC

The Kappa Statistic

(1) If a categorical variable is reported at two surveys by each of n subjects, then the Kappa statistic (κ) is used to measure reproducibility between surveys, where

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

p_o = observed probability of concordance between the two surveys
p_e = expected probability of concordance between the two surveys
$= \sum a_t b_t$
where $a_{i'} b_i$ are the marginal probabilities for the i th category in the $c \times c$ contingency table relating response at the two surveys.

(2)Furthermore,

$$se(\kappa) = \sqrt{\frac{1}{n\left(1-p_e\right)^2} \times \left\{ p_e + p_e^2 - \sum_{i=1}^{c} \left[a_l b_i \left(a_l + b_l\right)\right] \right\}}$$

To test the one-sided hypothesis $H_0 : \kappa = 0$ vs. $H_1 : \kappa > 0$, use the test statistic

$$Z = \frac{\kappa}{se(\kappa)}$$

which follows an $N(0,1)$ distribution under H_0 . (3) Reject H_0 at level *alpha* if $z > z_{1-\alpha}$ and accept H_0 otherwise.

(4) The exact p -value is given by $p = 1 - \Phi(z)$.

Guidelines for Evaluating Kappa
$\kappa > 75$ denotes excellent reproducibility.
$.4 \leq \kappa \leq .75$ denotes good reproducibility.
$0 \leq \kappa < .4$ denotes marginal reproducibility.

$$p_o = \sum_{i=1}^{k} O_{ii}/N = \frac{82+3}{100} = 0.85$$
$$p_e = \sum_{i=1}^{k} n_{i+}n_{+i}/N^2$$
$$= \frac{92\times 87+8\times 13}{100^2} = 0.8108$$
$$K = \frac{p_o - p_e}{1-p_e} = 0.207188$$

summary for the section

- This chapter discussed the most widely used techniques for analyzing qualitative or categorical data
- For large sample, we can use chi-square test to compare binomial proportions or rates from two independent samples. For the small-sample case, Fisher's exact test is used to compare binomial proportions in two independent samples.
- To compare binomial proportions in the paired samples, McNemar's test for proportions should be used
- For R×C contingency table, a chi-square test is a direct generalization of 2 ×2 contingency table test.

Chapter 4

Regression analysis

4.1 Simple linear regression

4.1.1 model introduction

Assume that we only have information on observation and we observe n pairs (y_i, x_i).
Specifying the model: given $(x_1, \ldots, x_n)$
we assume that

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\varepsilon \sim \left(0, \sigma^2 \cdot I_{n\times n}\right)$$

Fitting the model: Estimate (β_0, β_1)
Least squares

$$\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = (\beta_0, \beta_1) argmin \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

Computation: Find $\left(\widehat{\beta}_0, \widehat{\beta}_1\right)$
Modelling the y (y- dependent variable) on x (x - independent variable) alone we only had **one covariate**: hence it is a "simple" model. In the model

$$E(y \mid x) = \beta_0 + \beta_1 x$$

i.e. the conditional expectation of y given x is linear in x . Hence it is a linear regression model.
In general, a linear regression model for an outcome y and covariates $x_1, \ldots, x_p$ states that

$$E\left(y \mid x_1, \ldots, x_p\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

Could also be a linear combination of known functions of x_j - maybe polynomials, etc.

4.1.2 Least Square : Comuputation

In regression model, least squares regression chooses the line that minimizes

$$SSE\left(\beta_0,\beta_1\right)=\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2.$$

Normal equations:

$$\begin{array}{c}\frac{\partial SSE}{\partial\beta_0}=-2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)=0\\ \frac{\partial SSE}{\partial\beta_1}=-2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)\cdot x_i=0\end{array}$$

$$\left\{\begin{array}{l}\overline{y}=\widehat{\beta}_0+\widehat{\beta}_1\overline{x},\\ \overline{xy}=\widehat{\beta}_0\overline{x}+\widehat{\beta}_1\overline{x^2}\end{array}\right.\Longrightarrow\left\{\begin{array}{l}\widehat{\beta}_1=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\overline{x^2}-\bar{x}^2},\\ \widehat{\beta}_0=\bar{y}-\widehat{\beta}_1\bar{x}.\end{array}\right.$$

Define

$$\begin{array}{c}\ell_{xx}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2,\\ \ell_{yy}=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2,\\ \ell_{xy}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right).\end{array}$$

We thus have

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(x_i-\bar{x}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(x_i-\bar{x}\right)}=\frac{\ell_{xy}}{\ell_{xx}}=\frac{1}{\ell_{xx}}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)y_i$$

4.1.3 Expectation and Variance

Assumption A1: $E\left[\epsilon_i\right]=0, i=1,\ldots,n$. Under Assumption A1, $\hat{\beta}_0,\hat{\beta}_1$ are **unbiased estimators** for β_0,β_1 , respectively.
Proof:

$$\begin{aligned}E\left[\hat{\beta}_1\right]&=\tfrac{1}{\ell_{xx}}\textstyle\sum_{i=1}^{n}\left(x_i-\bar{x}\right)E\left[y_i\right]\\ &=\tfrac{1}{\ell_{xx}}\textstyle\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(\beta_0+\beta_1x_i\right)\\ &=\tfrac{\beta_0}{\ell_{xx}}\textstyle\sum_{i=1}^{n}\left(x_i-\bar{x}\right)+\tfrac{\beta_1}{\ell_{xx}}\textstyle\sum_{i=1}^{n}\left(x_i-\bar{x}\right)x_i\\ &=\tfrac{\beta_1}{\ell_{xx}}\textstyle\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(x_i-\bar{x}\right)\\ &=\beta_1\end{aligned}$$

$$E\left[\hat{\beta}_0\right]=E[\bar{y}-\hat{\beta}_1\bar{x}\Big]=E[\bar{y}]-\beta_1\bar{x}=\beta_0+\beta_1\bar{x}-\beta_1\bar{x}=\beta_0$$

Assumption A2: $Cov\left(\epsilon_i,\epsilon_j\right)=\sigma^2 1\{i=j\}$

Under Assumption A2, we have

$$Var\left[\hat{\beta}_0\right] = \left(\frac{1}{n} + \frac{\bar{x}^2}{\ell_{xx}}\right)\sigma^2$$
$$Var\left[\hat{\beta}_1\right] = \frac{\sigma^2}{\ell_{xx}}$$
$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{-\bar{x}}{\ell_{xx}}\sigma^2$$

Since $Cov(\epsilon_i, \epsilon_j) = 0$ for any $i \neq j, Cov(y_i, y_j) = 0$. We thus have

$$Var\left[\hat{\beta}_1\right]$$
$$= \frac{1}{\ell_{xx}^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 Var[y_i]$$
$$= \frac{\sigma^2}{\ell_{xx}^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sigma^2}{\ell_{xx}}$$

We next show that $Cov\left(\bar{y}, \hat{\beta}_1\right) = 0$.

$$Cov\left(\bar{y}, \hat{\beta}_1\right)$$
$$= Cov\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \frac{1}{\ell_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})\, y_i\right)$$
$$= \frac{1}{n\ell_{xx}} Cov\left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n}(x_i - \bar{x})\, y_i\right)$$
$$= \frac{1}{n\ell_{xx}}\sum_{i=1}^{n} Cov\left(y_i, (x_i - \bar{x})\, y_i\right)$$
$$= \frac{1}{n\ell_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})\,\sigma^2 = 0$$

$$Var\left[\hat{\beta}_0\right] = Var[\bar{y} - \hat{\beta}_1\bar{x}\Big] = Var[\bar{y}] + Var\left[\hat{\beta}_1\bar{x}\right] = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\ell_{xx}}$$

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = Cov\left(\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1\right) = -\bar{x}Var\left[\hat{\beta}_1\right] = \frac{-\bar{x}}{\ell_{xx}}\sigma^2.$$

4.1.4 Estimate of the variance of the error term

For Assumption $A2$, it is common that the variance σ^2 is unknown. The next theorem gives an unbiased estimate of σ^2 .

The sum of squared errors (SSE) is defined by

$$S_e^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Let

$$\hat{\sigma}^2 := \frac{Q\left(\hat{\beta}_0, \hat{\beta}_1\right)}{n-2} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2} = \frac{S_e^2}{n-2}.$$

Under Assumptions $A1$ and $A2$, we have $E\left[\hat{\sigma}^2\right] = \sigma^2$.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1$$

$$x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) .$$

$$\begin{aligned} E\left[Q\left(\hat{\beta}_0, \hat{\beta}_1\right)\right] &= \sum_{i=1}^n E\left[(y_i - \hat{y}_i)^2\right] = \sum_{i=1}^n Var\left[y_i - \hat{y}_i\right] + \left(E\left[y_i\right] - E\left[\hat{y}_i\right]\right)^2 \\ &= \sum_{i=1}^n \left[Var\left[y_i\right] + Var\left[\hat{y}_i\right] - 2Cov\left(y_i, \hat{y}_i\right)\right] \end{aligned}$$

$$\begin{aligned} Var\left[\hat{y}_i\right] = Var\left[\hat{\beta}_0 + \hat{\beta}_1 x_i\right] &= Var\left[\bar{y} + \hat{\beta}_1 (x_i - \bar{x})\right] \\ &= Var[\bar{y}] + (x_i - \bar{x})^2 Var\left[\hat{\beta}_1\right] \\ &= \frac{\sigma^2}{n} + \frac{(x_i - \bar{x})^2 \sigma^2}{\ell_{xx}} \end{aligned}$$

$$\begin{aligned} Cov\left(y_i, \hat{y}_i\right) &= Cov\left(\beta_0 + \beta_1 x_i + \epsilon_i, \bar{y} + \hat{\beta}_1 (x_i - \bar{x})\right) \\ &= Cov\left(\epsilon_i, \bar{y}\right) + (x_i - \bar{x}) Cov\left(\epsilon_i, \hat{\beta}_1\right) \\ &= \frac{\sigma^2}{n} + \frac{(x_i - \bar{x})^2 \sigma^2}{\ell_{xx}}. \end{aligned}$$

$$E\left[Q\left(\hat{\beta}_0, \hat{\beta}_1\right)\right] = \sum_{i=1}^n \left[\sigma^2 - \frac{\sigma^2}{n} - \frac{(x_i - \bar{x})^2 \sigma^2}{\ell_{xx}}\right] = (n-2)\sigma^2$$

4.1.5 Sampling distribution theorem

Assumption B: $\epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right), i = 1, \ldots, n$
Assumption B leads to Assumptions $A1$ and $A2$. Under Assumption B , we have

(1) $\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{\ell_{xx}}\right)\sigma^2\right)$.

(2) $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\ell_{xx}}\right)$.

(3) $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{S_\epsilon^2}{\sigma^2} \sim \chi^2(n-2)$.

(4) $\hat{\sigma}^2$ is independent of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$.

Under Assumption B, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$ independently. Both $\hat{\beta}_0, \hat{\beta}_1$ are linear combinations of y_i s. Consequently, they are normally distributed. We have known their expected values and variances from Theorems 4.1 and 4.2. The claims (1) and (2) are thus verified. The proofs of claims (3) and (4) are deferred to the general case.
It is n-2 degrees of freedom because we have fit two parameters to the n data points.

4.1.6 Hypothesis Testing

For known σ we can make tests and confidence intervals using

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{\ell_{xx}}} \sim N(0,1)$$

The $100(1-\alpha)\%$ confidence interval for β_1 is given by $\hat{\beta}_1 \pm u_{1-\alpha/2}\sigma/\sqrt{\ell_{xx}}$. For testing
H_0 :

$$\beta_1 = \beta_1^* \text{ vs. } H_1 : \beta_1 \neq \beta_1^*$$

we reject $H_0 if \left|\hat{\beta}_1 - \beta_1^*\right| > u_{1-\alpha/2}\sigma/\sqrt{\ell_{xx}}$ with the most popular hypothesized value being $\beta_1^* = 0$ (i.e., the regession function is significant or not at significance level α .) In the more realistic setting of unknown σ , so long as $n \geq 3$, using claims (2-4) gives

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{\ell_{xx}}} \sim t(n-2).$$

The $100(1-\alpha)\%$ confidence interval for β_1 is $\hat{\beta}_1 \pm t_{1-\alpha/2}(n-2)\hat{\sigma}/\sqrt{\ell_{xx}}$. For testing
H_0 :

$$\beta_1 = \beta_1^* \text{ vs. } H_1 : \beta_1 \neq \beta_1^*$$

we reject H_0 if $\left|\hat{\beta}_1 - \beta_1^*\right| > t_{1-\alpha/2}(n-2)\hat{\sigma}/\sqrt{\ell_{xx}}$. For drawing inferences about β_0 , we can use

$$\frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{1/n + \bar{x}^2/\ell_{xx}}} \sim N(0,1),$$
$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{1/n + \bar{x}^2/\ell_{xx}}} \sim t(n-2).$$

The $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left[\frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}(n-2)}, \frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha/2}(n-2)}\right] = \left[\frac{S_e^2}{\chi^2_{1-\alpha/2}(n-2)}, \frac{S_e^2}{\chi^2_{\alpha/2}(n-2)}\right]$$

4.2 Multiple linear regression

4.2.1 Model introduction

$$Y = g(x_1, x_2, \ldots, x_p) + \epsilon$$

where the deterministic function $g(x_1, x_2, \ldots, x_p)$ indicates the relationship between Y and $x_1, x_2, \ldots, x_p$ and the error term ϵ comes from the variability. We have discussed the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where $\epsilon \sim N\left(0, \sigma^2\right)$. We now extend this basic model to include multiple independent variables $x_1, x_2, \ldots, x_p$ with$p \geq 2$.
The extended model including multiple independent variables $x_1, x_2, \ldots, x_p$ is called multiple linear regression model. It has the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

- where $\epsilon \sim N\left(0, \sigma^2\right)$
- There are now p+1 unknown (but fixed) regression parameters $\beta_0, \beta_1, \beta_2, \ldots \beta_p$
- Y is still called dependent variable (random), and $x_1, x_2, \ldots, x_p$ are called independent variables (fixed).
- Error term ϵ is random (normal) and unknown, as well.

4.2.2 Least Square : Computation

Review

In simple linear regression, we use Method of Least Squares (LS) to fit the regression line. LS estimates the value of $\beta_0 and \beta_1$ by minimizing the sum of squared distance between each observed $\mathrm{Y}_i and its population value \beta_0 + \beta_1 x_i for each x_i$.

$$Q\left(\beta_0, \beta_1\right) = \sum_{i=1}^{n} \left[Y_i - \left(\beta_0 + \beta_1 x_i\right)\right]^2$$

In multiple linear regression, we plan to use the same method to estimate regression parameters $\beta_0, \beta_1, \beta_2, \ldots \beta_p$. It is easier to derive the estimating formula of the regression parameters by the **form of matrix**. So, before uncover the formula, let's take a look of the matrix representation of the multiple linear regression function.

Matrix form

Define

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{n \times 1} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n \times (p+1)}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1} \qquad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n\times 1}$$

where x_{ij} is the measurement on the jth independent variable for the ith individual, for $i = 1, 2, \cdots, nandj = 1, 2, \cdots, p$.
With these definitions, the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

for $i = 1, 2, \cdots, n$, can be expressed equivalently as

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

- $\mathbf{Y}$ is an $n \times 1$ (random) vector of responses.
- $\mathbf{X}$ is an $n \times (p+1)$ (fixed) matrix of independent variable measurements.
- β is a $p \times 1$ (fixed) vector of unknown population regression parameters
- ϵ is an $n \times 1$ (random) vector of unobserved errors.

The notion of least squares is the same in multiple linear regression as it was in simple linear regression. Specifically, we want to find the values of $\beta_0, \beta_1, \beta_2, \ldots \beta_p$ that minimize

$$Q\left(\beta_0, \beta_1, \beta_2, \ldots \beta_p\right) = \sum_{i=1}^{n} \left[Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})\right]^2$$

Recognize that

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

is the inner (dot) product of the ith row of $\mathbf{X}$ and β , e.g. $\mathbf{X}\beta$. Therefore,

$$Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$$

is the ith entry in the difference vector $\mathbf{Y} - \mathbf{X}\beta$.
The objective function Q can be expressed by

$$Q(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta)$$

the inner (dot) product of $\mathbf{Y} - \mathbf{X}\beta$ with itself.

- Remark: For any vector A , the notation A^T represents the transpose of A . It makes the columns of the new matrix A^T the rows of the original A .

- Fact 1: For any two matrices A and B, $(AB)^T = B^T A^T$. For example, $(\mathbf{X}\beta)^T = \beta^T \mathbf{X}^T$.
- Fact 2: For any two vectors C and D, $C^T D = D^T C$

Derivation ot Least squares Estimator Let's expend the $Q(\beta)$:

$$Q(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta)$$

$$= \left[\mathbf{Y}^T - (\mathbf{X}\beta)^T\right] [\mathbf{Y} - \mathbf{X}\beta]$$

$$= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T(\mathbf{X}\beta) - (\mathbf{X}\beta)^T\mathbf{Y} + (\mathbf{X}\beta)^T(\mathbf{X}\beta)$$

$$= \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T(\mathbf{X}\beta) + (\mathbf{X}\beta)^T(\mathbf{X}\beta) \quad \text{(Fact 2)}$$

$$= \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta \quad \text{(Fact 1)}$$

In order to find the value of β to minimize $Q(\beta)$, we take derivative and set it to zero.

$$\frac{\partial Q(\beta)}{\partial \beta} = \frac{\partial \left(\mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta\right)}{\partial \beta} \equiv 0$$

$$\frac{\partial Q(\beta)}{\partial \beta} = \frac{\partial \left(\mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta\right)}{\partial \beta}$$

$$= -2\left(\mathbf{Y}^T\mathbf{X}\right)^T + 2\mathbf{X}^T\mathbf{X}\beta$$

$$= -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\beta \quad \text{(Fact 1)}$$

$$\equiv 0$$

The last equation gives $\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{Y}$, leading to the LS estimator of β to be Given $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T Y$, the function we use to predict $\mathbf{Y}$ is

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ is called the "hat-matrix" $(n \times n)$, because it helps $\mathbf{Y}$ to wear a hat!
Remark: $\mathbf{H}^T = \mathbf{H}(symmetric)$ and $\mathbf{HH} = \mathbf{H}$ (idempotent). Residuals: $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, where $\mathbf{I}$ is called identity matrix, which looks like

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}_{n \times n}$$

4.2.3 Estimate of the variance of the error term

In multiple linear regression, we have

$$\begin{aligned}
\text{SSE } &= \textstyle\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}}) \\
&= (\mathbf{Y} - \mathbf{X}\hat{\beta})^T (\mathbf{Y} - \mathbf{X}\hat{\beta}) \\
&= (\mathbf{Y} - \mathbf{H}\mathbf{Y})^T (\mathbf{Y} - \mathbf{H}\mathbf{Y}) \\
&= [(\mathbf{I} - \mathbf{H})\mathbf{Y}]^T [(\mathbf{I} - \mathbf{H})\mathbf{Y}] \\
&= \mathbf{Y}^T (\mathbf{I} - \mathbf{H})^T (\mathbf{I} - \mathbf{H})\mathbf{Y} \quad \text{Because } (AB)^T = B^T A^T \\
&= \mathbf{Y}^T (\mathbf{I} - \mathbf{H})\mathbf{Y}
\end{aligned}$$

Remark: I- H is symmetric and idempotent as well.
Fact: $E(SSE) = (n - p - 1)\sigma^2$ In multiple linear regression, define $MSE = \frac{SSE}{n-p-1} MSE$ is an unbiased estimator of σ^2

$$E(MSE) = E\left(\frac{SSE}{n-p-1}\right) = \frac{(n-p-1)\sigma^2}{n-p-1} = \sigma^2$$

Therefore,

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-p-1} == \frac{\mathbf{Y}^\top (\mathbf{I} - \mathbf{H})\mathbf{Y}}{n-p-1}$$

and

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-p-1}} = \sqrt{\frac{\mathbf{Y}^\top (\mathbf{I} - \mathbf{H})\mathbf{Y}}{n-p-1}}$$

4.2.4 Inference for Individual Regression Parameters

In the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

where $\epsilon \sim N\left(0, \sigma^2\right)$, we interested in writing confidence intervals for individual regression parameters β_j , and we also want to test whether $H_0 : \beta_j = 0$, or not.
It can help us access the importance of using the independent variable x_j in a model including the other independent variables. Remark: inference regarding the β_j is always conditional on the other variables being included in the model. Under our linear regression model assumptions, $a100(1-\alpha)\%$ confidence interval for $\beta_j, j = 0, 1, 2, \ldots, p$, is given by

$$\hat{\beta}_j \pm t_{n-p-1,\alpha/2}\sqrt{MSE \times c_{jj}}$$

- $\hat{\beta}_j$ is the least square estimate of β_j (the j^{th} element in $\hat{\beta}$ vector).
- $MSE = \frac{SSE}{n-p-1} = \frac{\mathbf{Y}^\top(\mathbf{I}-\mathbf{H})\mathbf{Y}}{n-p-1}$
- $c_{jj} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{jj}$ is the corresponding j^{th} diagonal element of the $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ matrix.

Interpretation: We are $100(1-\alpha)\%$ confident that the population parameter β_j is in this interval.

4.2.5 Hypothesis test for the β_j

F test for joint hypothesis

We would like to test the overall hypothesis that the predictors when considered together have significant impact on the outcome.

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \text{ versus } H_a : \text{least one } \beta_j \neq 0.$$

This is the overall hypothesis that at least some of the β_j 's are different from zero, but without specifying which one is different. Similar to the F test for SLR
Estimate the regression parameters using the method of least squares, and compute Reg SS and Res SS

$$\begin{gathered}\text{Res SS} = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \\ Reg\text{SS} = \text{ Total SS } - \text{ Res SS} \\ \text{Total SS} = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \\ \hat{y}_i = a + \sum_{j=1}^{k} b_j x_{ij}\end{gathered}$$

$x_{ij} = jth$ independent variable for the ith subject, $j = 1, \ldots, k; i = 1, \ldots, n$

- Fstatistic :

$$F = \frac{(RegSS)/k}{(ReSSS)/(n-k-1)} = \frac{RegMS}{ResMS}$$

- Null distribution:Let $e \sim N\left(0, \sigma^2\right)$. If H_0 is true, then $F \sim F_{k,n-k-1}$, the centralized F distribution with $(k, n-k-1)$ degrees of freedom.
- The exact p -value of the observed F value is given by p= $Pr\left(F_{k,n-k-1} > F\right)$
- $Pr\left(F_{k,n-k-1} > F\right)$ Decision rule:Reject the H_0 at nominal level $\alpha \in (0,1)$ if $F > F_{k,n-k-1,1-\alpha}$ (equivalently, $p < \alpha$), where the critical value $F_{k,n-k-1,1-\alpha}$, is the $1-\alpha$ percentile of $F_{k,n-k-1}$.

Table 2: Analysis of variance (ANOVA)

Source	DF	Sum of Squares	Mean Squares	F
Regression	k	RegSS	$\text{RegMS} = \frac{\text{RegSS}}{k}$	$F = \frac{\text{RegMS}}{\text{ResMS}}$.
Residual	$n-k-1$	ResSS	$\text{ResMS} = \frac{\text{ResSS}}{n-k-1}$	
Total	$n-1$	TotSS	The p value of a realized F is: $p = \Pr(F \sim F_{k,n-k-1} > F)$.	

Figure 4.1: Analysis of variance (ANOVA)

for a partial hypothesis

Often, we want to know whether an individual predictor x_j has a significant effect on outcome y after controlling for the other predictors. The partial hypothesis on β_j is
Partial T test

$$H_{0j} : \beta_j = 0 \text{ versus } H_{aj} : \beta_j \neq 0$$

we assume other β is making a contribution under either hypothesis

- The t - statistic:

$$t\left(x_j \mid \text{ other } x's\right) = \frac{\hat{\beta}_j}{se\left(\hat{\beta}_j\right)}.$$

- Null distribution: Let $e \sim N\left(0, \sigma^2\right)$. If $H_{0j} : \beta_j = 0 istrue, then\, (x_j \mid$ other $x's) \sim t_{n-k-1}$, the centralized t distribution with n-k-1 degrees of freedom.

- The exact two-tailed p value is $p = 2 \times Pr\left(t_{n-k-1} > \mid t\left(x_j \mid \text{ other } x's\right) \mid\right)$.

- Decision rule: Reject H_{0j} at nominal level $\alpha \in (0,1)$ if $\mid t\left(x_j \mid$ other $x's) \mid > t_{n-k-1,1-\alpha/2}$ (equivalently, $p < \alpha$).

Partial F test
Aforesaid partial t test for the effect of one particular predictor adjusts for the contribution of the other predictor. To better understand this point, let us 'develop' partial F test for

$$H_{02}: \beta_2 = 0 \text{ vs. } H_{a2}: \beta_2 \neq 0$$

in full model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

after adjusting for the contribution of x_1 .

- If H_{02} is false, then we have the full model and $RegSS(x_1, x_2)$.
- The extra sum of squares due to x_2 after adjusting for x_1 is

$$RegSS(x_2 \mid x_1) = RegSS(x_1, x_2) - RegSS(x_1).$$

- Hence

$$RegMS(x_2 \mid x_1) = \frac{RegSS(x_2 \mid x_1)}{1} = RegSS(x_2 \mid x_1)$$

- The F test statistic is

$$F(x_2 \mid x_1) = \frac{RegMS(x_2 \mid x_1)}{ResMS(x_1, x_2)}$$

- Null distribution: Let $e \sim N(0, \sigma^2)$. If H_{02} is true, then $F(x_2 \mid x_1) \sim F_{1,n-3}$.
- The exact p value of a realized $F(x_2 \mid x_1)$ is given by $p = Pr(F_{1,n-3} > F(x_2 \mid x_1))$
- Decision rule: Reject H_{02} when the realized $F(x_2 \mid x_1) > F_{1,n-3,1-\alpha}$(equivalently, $p < \alpha$)

4.2.6 Multiple and partial correlations

For a given data set $\{(y_i, x_i = (x_{i1}, \ldots, x_{ik})) : i = 1, \ldots, n\}$, we fit the genuine data generating mechanism by aforesaid multiple linear model. Thus far, we have learnt to:

1 estimate the unknown parameters by the method of Least Squares;

2 use $R^2 = \frac{\text{RegSS}}{\text{Total SS}}$ to measure the strength of linear relationship;

3 use $R^2 = \frac{\text{RegSS}}{\text{Total SS}}$ to measure the strength of linear relationship;

4 perform partial t test or partial F test for partial hypothesis $H_{0j}: \beta_j = 0 versus H_{aj}: \beta_j \neq 0$.

- Note 1: In multiple linear regression, we often assume normality $\left(e \sim N\left(0, \sigma^2\right)\right)$, linearity $(\mathrm{E}\left(y \mid x_1, \ldots, x_k\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$, homoscedasticity (σ^2 does not depends on x values) and independence (i.i.d. sample). In other words, real model is assumed to be identical to postulated model, up to unknown parameters.
- Note 2: For given data points, we can compute the estimates even if all the assumptions are not true. In general, postulated statistical models are not identical to the data generating mechanism. The discrepancy between the real model and the linear model determines the performance of the estimators.
- Note 3: As illustrated, $t^2\left(x_j \mid .\right.$ other $\left.x's\right) = F\left(x_j \mid .\right.$ other $\left.x's\right)$ for arbitrary specific data points (n¿k+1) , no matter all the assumptions are true.
- Note 4: If all the assumptions are true, then under the null, $t^2\left(x_j \mid other\ x's\right) = F\left(x_j \mid .\right.$ other $\left.x's\right) \sim F_{1,n-k-1}$. Even without normality, the statistics may approximately follows $F_{1,n-k-1}$ for large sample size.

Multiple Correlation

Sometimes, we are interested in existence and strength of the association (linear relationship) between one variable y and a set of variables $(x_1, \ldots, x_k)$ when considered as a group () other than the effects themselves. The multiple correlation is a useful measure for such association.

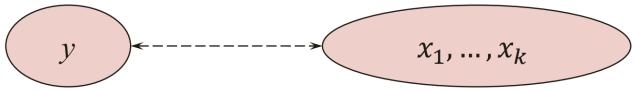


Figure 4.2: Existence and strength of the linear relationship between variable y and a set of variables $(x_1, \ldots, x_k)$.

Definition Definition 1.1: The coefficient of the multiple correlation (ϱ) between variable y and a set of variables $\{x_1, \ldots, x_k\}$ is defined as the maximum possible correlation coefficient between y and a linear combination of $\{x_1, \ldots, x_k\}$. Mathematically, we have

$$\varrho = \sqrt{c' \Omega^{-1} c}$$

where $c = \left(\rho_{y,x_1}, \ldots, \rho_{y,x_k}\right)'$ is the $k \times 1$ vector of cross-correlationsand $\Omega = \left(\rho_{x_i x_j}\right)$ is the correlation matrix of inter-correlations between variable $x_1, \ldots, x_k$.

Note 1 : This population level definition allows for arbitrary variable y and an arbitrary set of variables $\{x_1, \ldots, x_k\}$. It does not require normality of any of the variables, linearity relationship or even the dependence assumption.

A large ϱ value indicates a strong association (linear relationship) between y and the set of predictors. If $\varrho = 0$, then there is no linear relationship between y and $\{x_1, \ldots, x_k\}$.

Note 2 : For given data points $\{(y_i, x_i = (x_{i1}, \ldots, x_{ik})) : i = 1, \ldots, n\}$, the multiple correlation coefficient is estimated by the Pearson correlation between $(y_1, \ldots, y_n)\, and\, (\hat{y}_1, \ldots, \hat{y}_n)$, where $\hat{y}_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, and \beta_1, \ldots, \beta_k$ are the least-squares estimates of $\beta_1, \ldots, \beta_k$. The sample multiple correlation coefficientequals to

$$\varrho = \sqrt{R^2} = \sqrt{(RegSS)/(\text{ TotSS })}.$$

F test for multiple correlation

- The hypothesis is $H_0 : \varrho = 0 versus H_a : \varrho \neq 0$
- F statistic: $F = \frac{R^2}{1-R^2} \frac{n-k-1}{k}$.
- Null distribution:
  - (i) If (y_i, x_i) 's are i.i.d. and (y_i, x_i) follows a joint normal distribution, then under the null, $F \sim F_{k,n-k-1}$, the centralized F distribution with $(k, n-k-1)$ degrees of freedom.
  - (ii) If (y_i, x_j) 's are i.i.d. and the sample size is large, then under the null F approximately $F_{k,n-k-1}$.

Cont'd

- The exact p-value of the observed F value is given by $p = Pr\,(F_{k,n-k-1} > F)$.
- Decision rule: Reject the H_0 at nominal level $\alpha \in (0,1) if F > F_{k,n-k-1,1-\alpha}$ (equivalently, $p < \alpha$), where the critical value $F_{k,n-k-1,1-\alpha}$, is the $1-\alpha$ percentile of $F_{k,n-k-1}$

Note 3: The above F statistic is mathematically the F statistic for the joint hypothesis $H_0 : \beta_1 = \cdots = \beta_k = 0 \text{versus} H_a : \text{least one} \beta_j \neq 0$.

Prove it by the definitions of RegSS, TotSS, R^2 and F statistics.

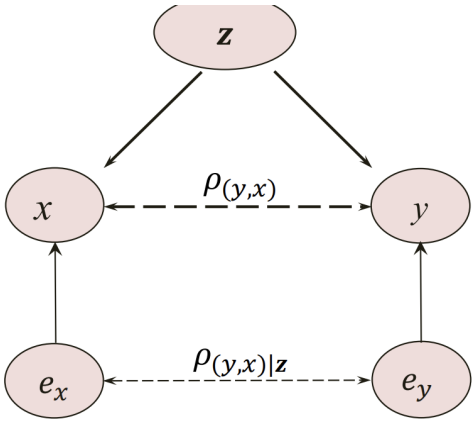


Figure 4.3: Correlation and partial correlation

Partial Correlation

It is important to assess the degree of association between two variables (y, x_j) after controlling for other variables $\{x_1, \ldots x_{j-1}, x_{j+1}, x_k\}$

In context, I rewrite x_j as x , and rewrite variables $\{x_1, \ldots x_{j-1}, x_{j+1}, x_k\}$ as partial $z = (z_1, \ldots, z_{k-1})$ correlation (Figure 4.3) accomplish this goal.
2.1. Definition

Definition 2 2: The partial correlation coefficient two variables (y, x) after controlling for z is defined as $\rho_{(y,x)|}$ | **z** the Pearson correlation between two residual variables $\mathrm{e}_x and e_y in the multiple linear arregressions \mathrm{x}=\gamma_0+\gamma_1 z_1+\cdots+\gamma_k z_{k-1}+e_x$ and $y = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_k z_{k-1} + e_y$.

Note 4: Partial Correlation Coefficient is a measure of the strength of the association between two variables (y, x) after controlling for the effects of other variables z .

Note 6: In particular, for three random variables (y, x_1, x_2) , the Partial Correlation Coefficient between y and x_1 is define as

$$\rho_{(y,x_1)|x_2} = \frac{\rho_{y,x_1} - \rho_{y,x_2}\rho_{x_1,x_2}}{\sqrt{\left(1-\rho^2_{y,x_2}\right)\left(1-\rho^2_{x_1,x_2}\right)}}$$

where the notation ρ .. denotes the Pearson Moment Product Correlation Coefficient between two random variables.
The Sample Partial Correlation Coefficient between y and x_1 is define as

$$r_{(y,x_1)|x_2} = \frac{r_{y,x_1} - r_{y,x_2}r_{x_1,x_2}}{\sqrt{\left(1-r^2_{y,x_2}\right)\left(1-r^2_{x_1,x_2}\right)}}$$

where the notation r_i denotes the Sample Pearson Moment Product Correlation Coefficient between two random variables.
Definition 3 : The partal coefficlent of determination of x after controlling for z is defined as the proportion that the variability in y not explained by z is now explained by x :

$$R^2_{(y,x)|\mathbf{z}} = \frac{RegSS(x \mid \mathbf{z})}{ResSS(\mathbf{z})}$$

Note 5: For a given data set, $r^2_{(y,x)|\mathbf{z}^2} = R^2_{(y,x)|\mathbf{z}}$

Hypothesis testing for partial correlation

We want to know whether an individual variable x is significantly associated with y after controlling for the other variables z .

The hypothesis is $H_0 : \rho_{(y,x)|z} = 0 versus H_a : \rho_{(y,x)|z} \neq 0$.

4.3 Logistic regression

4.3.1 Motivations

In health sciences, researchers often need to investigate the association between categorical (dichotomous) dependent variables with categorical and/or quantitative independent variables.
Examples: Typical dichotomous dependent variables include disease categories (affected, unaffected), death status (dead, alive), and remission status (in remission, not in remission).

Simple and multiple linear regressions do NOT apply.

For such scenarios, traditional linear regression techniques are not appropriate. Some fundamental assumptions are severely violated. In particular, the normality assumption (on residual and thus dependent variable) does not hold any more.

Contingency analysis is insufficient

Analysis for $R * C$ contingency could deal with some simple situations, where both dependent and independent variables are categorical variables.

Dichotomization of quantitative independent variables may result in loss of information and thus reduce statistical power.

Another important issue is whether a disease–exposure relationship is influenced by confounders (covariates). Contingency table analysis would be able to do nothing about adjustment of confounders

$$\theta^T \mathbf{X} = \sum_{i=1}^{n} \theta_i X_i = \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_n X_n$$
$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Logistic Regression is a classification algorithm that works by trying to learn a function that approximates $P(Y \mid X)$. It makes the central assumption that $P(Y \mid X)$ can be approximated as a **sigmoid function** applied to a linear combination of input features. Mathematically, for a single training datapoint $(\mathbf{x}, y)$ Logistic Regression assumes:

$$\begin{matrix} P(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \sigma(z) \\ z = \theta_0 + \sum_{i=1}^{m} \theta_i x_i \end{matrix} \quad c$$

This assumption is often written in the equivalent forms:

$$P(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \sigma\left(\theta^T \mathbf{x}\right)$$
$$P(Y = 0 \mid \mathbf{X} = \mathbf{x}) = 1 - \sigma\left(\theta^T \mathbf{x}\right)$$

Using these equations for probability of $Y \mid X$, we can create an algorithm that select values of θ that maximize that probability for all data. I am first going to state the log probability function and partial derivatives with respect to θ .

4.3.2 Likelihood function

$$L(\theta) = \prod_{i=1}^{n} P\left(Y = y^{(i)} \mid X = \mathbf{x}^{(i)}\right)$$
$$= \prod_{i=1}^{n} \sigma\left(\theta^T \mathbf{x}^{(i)}\right)^{y^{(i)}} \cdot \left[1 - \sigma\left(\theta^T \mathbf{x}^{(i)}\right)\right]^{\left(1-y^{(i)}\right)}$$

4.3.3 MLE method

$$LL(\theta) = \sum_{i=1}^{n} \log\left(f\left(\mathbf{x}^{(i)}, y^{(i)} \mid \theta\right)\right) = \sum_{i=1}^{n} \log\left(f\left(\mathbf{x}^{(i)} \mid \theta\right) P\left(y^{(i)} \mid \mathbf{x}^{(i)}, \theta\right)\right)$$
$$= \sum_{i=1}^{n} \log\left(f\left(\mathbf{x}^{(i)}\right) f\left(y^{(i)} \mid \mathbf{x}^{(i)}, \theta\right)\right)$$
$$\theta_{MLE} = argmax LL(\theta) = argmax\left(\sum_{i=1}^{n} \log f\left(\mathbf{x}^{(i)}\right) + \log f\left(y^{(i)} \mid \mathbf{x}^{(i)}, \theta\right)\right)$$
$$= argmax\left(\sum_{i=1}^{n} \log f\left(y^{(i)} \mid \mathbf{x}^{(i)}, \theta\right)\right)$$

In general, we can write it in this form

$$Y_i \sim Binom\left(p_i\right)$$
$$logit(p) = \log\left(\frac{p}{1-p}\right), 0 \leq p \leq 1$$
$$\eta_i = \log\left(\frac{p_i}{1-p_i}\right) = \theta_0 + \theta_1 X_{i1} + \ldots + \theta_k X_{ik}$$
$$OR = e^{\theta}$$

4.3.4 Odd - Ratios

To simplify the analysis with our table. we only talk about the simple logisitc regression's odd - ratio.

Note that $0 \leq p(x) \leq 1$ for any specific value x of X . The independent variable X can be either quantitative or categorical.

In the SLRM, θ_1 represents the average change in the log odds for every one-unit change in x .

Inverting with an exponential function, we see the odds in favor of success represented as a function of x :

$$ODDS(x) = \frac{p(x)}{1-p(x)} = e^{\theta_0+\theta_1 x}$$

Often, OR=we compare the odds in favor of success (Y= 1) at two distinct values of an independent variable X .
Definition. (OR-Odds Ratio): Odds ratio (OR) between the odds at two fixed values x_i, x_j of X is defined as

$$\text{OR} = \frac{ODDS\,(x_i)}{ODDS\,(x_j)}$$

- $OR = 1$: the probability of success is the same for individuals at $X = x_i$ and $X = x_j$
- $OR > 1$: the probability of success is greater for those with X=x$_i$*thanthosewith*X=x$_j$.
- $OR < 1$: the probability of success is greater for those with X $= x_j$ than those with $X = x_i$.

Under the SLRM, we have

$$OR = \frac{ODDS\,(x_i)}{ODDS\,(x_j)} = \frac{e^{\beta_0+\beta_1 x_i}}{e^{\beta_0+\beta_1 x_j}} = e^{\beta_1(x_i-x_j)}$$

The Maximum Likelihood Estimation Method is needed to fit the logistic regression model, i.e., compute the maximum likelihood estimates (MLEs) of the regression coefficients.
In general, this MLE method requires an intensive iterative process for optimization. We will rely on statistical packages for this purpose. Using the MLEs, we can estimate odds and odds ratio accordingly. 2. Connection with Contingency-Iable Inalysis In particular, if X is a categorical variable at two categories with dummy variable coding: X = 1 for category 1 ;

	case	control
Exposure (x = 1)	a	b
Non-exposure (x = 0)	c	d

X = 0 for the other category.

$$ODDS(x=1) = \frac{p(x=1)}{1-p(x=1)} = \frac{a/(a+b)}{1-a/(a+b)} = a/b$$
$$ODS(x=0) = \frac{p(x=0)}{1-p(x=0)} = \frac{c/(c+d)}{1-c/(c+d)} = c/d$$
$$OR = \frac{\mathbf{ODDS(X=1)}}{\mathbf{ODDS(X=0)}} = \frac{e^{\theta_0+\theta_1\times 1}}{e^{\theta_0+\theta_1\times 0}} = e^{\theta_1}.$$
$$OR = \frac{odds(x=1)}{odds(x=0)} = \frac{ad}{bc}$$

We can estimate the O R relating Y to X in either of two equivalent ways:

a . We can estimate the OR directly from the 2×2 table: (ad)/(bc).

b . We can set up a logistic-regression model of the form

$$\log[p/(1-p)] = \theta_0 + \theta_1 X$$

where p= probability of Y=1 given X=1 and where we estimate the OR by $\exp\left(\widehat{\theta}_1\right)$. 2. For simple random samples (prospective or cross-sectional studies), we can estimate the $Pr(\mathrm{Y}=1 \mid \mathrm{X}=1)$ and $Pr(\mathrm{Y}=1 \mid \mathrm{X}=0)$ in either of two equivalent ways: a. From the 2×2 table, we have

$$Pr(Y=1 \mid X=1) = \frac{a}{a+c},$$
$$Pr(Y=1 \mid X=0) = \frac{b}{b+d}.$$

From the logistic-regression model,

$$Pr(Y=1 \mid X=1) = \frac{e^{\widehat{\theta}_0+\widehat{\theta}_1}}{1+e^{\widehat{\theta}_0+\widehat{\theta}_1}}, Pr(Y=1 \mid X=0) = \frac{e^{\widehat{\theta}_0}}{1+e^{\widehat{\theta}_0}}.$$

4.4 Generalized Linear Models*

All of the regression models we have considered (including multiple linear, logistic, and Poisson) actually belong to a family of models called generalized linear models. (In fact, a more "generalized" framework for regression models is called general regression models, which includes any parametric regression model.) Generalized linear models provides a generalization of ordinary least squares regression that relates the random term (the response Y) to the systematic term (the linear predictor $\mathbf{X}\beta$) via a link function (denoted by $g(\cdot)$). Specifically, we have the relation

$$\mathrm{E}(Y) = \mu = g^{-1}(\mathbf{X}\beta)$$

so $g(\mu) = \mathbf{X}\beta$. Some common link functions are: The identity link:

$$g(\mu) = \mu = \mathbf{X}\beta$$

which is used in traditional linear regression. The logit link:

$$\begin{gathered} g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \mathbf{X}\beta \\ \Rightarrow \mu = \frac{e^{\mathbf{X}\beta}}{1+e^{\mathbf{X}\beta}} \end{gathered}$$

which is used in logistic regression. The log link:

$$\begin{gathered} g(\mu) = \log(\mu) = \mathbf{X}\beta \\ \Rightarrow \mu = e^{\mathbf{X}\beta} \end{gathered}$$

which is used in Poisson regression. The probit link:

$$\begin{gathered} g(\mu) = \Phi^{-1}(\mu) = \mathbf{X}\beta \\ \Rightarrow \mu = \Phi(\mathbf{X}\beta), \end{gathered}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. This link function is also sometimes called the normit link. This can also be used in logistic regression.
The complementary log-log link:

$$\begin{gathered} g(\mu) = \log(-\log(1-\mu)) = \mathbf{X}\beta \\ \Rightarrow \mu = 1 - \exp\left\{-e^{\mathbf{X}}\beta\right\} \end{gathered}$$

which can also be used in logistic regression. This link function is also sometimes called the gompit link. The power link:

$$\begin{gathered} g(\mu) = \mu^{\lambda} = \mathbf{X}\beta \\ \Rightarrow \mu = (\mathbf{X}\beta)^{1/\lambda} \end{gathered}$$

where $\lambda \neq 0$. This is used in other regressions which we do not explore (such as gamma regression and inverse Gaussian regression). Also, the variance is typically a function of the mean and is often written as

$$Var(Y) = V(\mu) = V\left(g^{-1}(\mathbf{X}\beta)\right).$$

The random variable Y is assumed to belong to an exponential family distribution where the density can be expressed in the form

$$q(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\},$$

where $a(\cdot), b(\cdot)$, and $c(\cdot)$ are specified functions, θ is a parameter related to the mean of the distribution, and ϕ is called the dispersion parameter. Many probability distributions belong to the exponential family. For example, the normal

distribution is used for traditional linear regression, the binomial distribution is used for logistic regression, and the Poisson distribution is used for Poisson regression. Other exponential family distributions lead to gamma regression, inverse Gaussian (normal) regression, and negative binomial regression, just to name a few.
The unknown parameters, β , are typically estimated with maximum likelihood techniques (in particular, using iteratively reweighted least squares), Bayesian methods, or quasi-likelihood methods. The quasi-likelihood is a function which possesses similar properties to the log-likelihood function and is most often used with count or binary data. There are also tests using likelihood ratio statistics for model development to determine if any predictors may be dropped from the model.