Biostatistics

Jinhong Du,15338039

July 7, 2018

Content

Chapter 1 Overview

1 Hypothesis testing

1.1 Definition

The **null hypothesis**, denoted by H_0, is the hypothesis that is to be tested. The **alternative hypothesis**, denoted by H_1 is the hypothesis that in some sense contradicts the null hypothesis.

In a **one − sample problem**, hypotheses are specified about a single distribution; in a **two − sample problem**, two different distributions are compared.

In a hypothesis test, four possible outcomes can occur:

		Truth	
		H_0	H_1
Decision	Accept H_0	Correctly accept null True Positive Probability: specificity	Type II error False Negative Probability: β
	Reject H_0	Type I error False Positive Probability: α	Correctly reject null True Negative Probability: $Power$ =sensitivity

Table 1: Four Possible Outcomes

The probability of a **type I error** is usually denoted by α and is commonly referred to as the significance level of a test.

The probability of a **type II error** is usually denoted by β.

The **power** of a test is defined as

$$\begin{aligned} Power &= 1-\beta \\ &= 1-\text{probability of a type II error} \\ &= \mathbb{P}(\text{rejecting } H_0 | H_1 \text{true}) \end{aligned}$$

The **acceptance region** is the range of values of the test statistic for which H_0 is accepted.

The **rejection region** is the range of values of the test statistic for which H_0 is rejected.

1.2 Critical-value Method

The critical-value method of hypothesis testing is a general approach in which we compute a **test statistic** and determine the outcome of a test by comparing the test statistic with a **critical value** determined by the type I error.

1.3 p-value Method

The p-value for any hypothesis test is the α level at which we would be indifferent between accepting or rejecting H_0 given the sample data at hand. That is, the p-value is the α level at which the given value of the test statistic (such as t) is on the borderline between the acceptance and rejection regions.

The p-value can also be thought of as the probability of obtaining a test statistic as extreme as or more extreme than the actual test statistic obtained, given that the null hypothesis is true.

1.4 Sensitivity And Specificity

The **sensitivity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is present given that the person has a disease.

The **specificity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is not present given that the person does not have a disease.

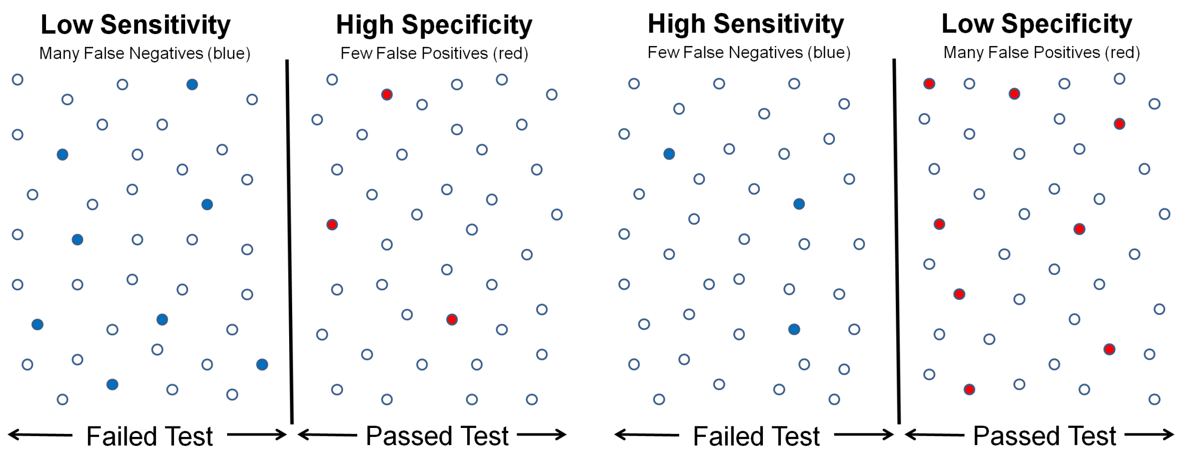

Figure 1: Sensitivity And Specificity

A **false negative** is defined as a negative test result when the disease or condition being tested for is actually present. A **false positive** is defined as a positive test result when the disease or condition being tested for is not actually present.

Given α, we have a table like

$$\begin{aligned} sensitivity &= \mathbb{P}\{\text{rejected}|H_1\} \\ &= \frac{d}{b+d} \\ specificity &= \mathbb{P}\{\text{not rejected}|H_0\} \\ &= \frac{a}{a+c} \end{aligned}$$

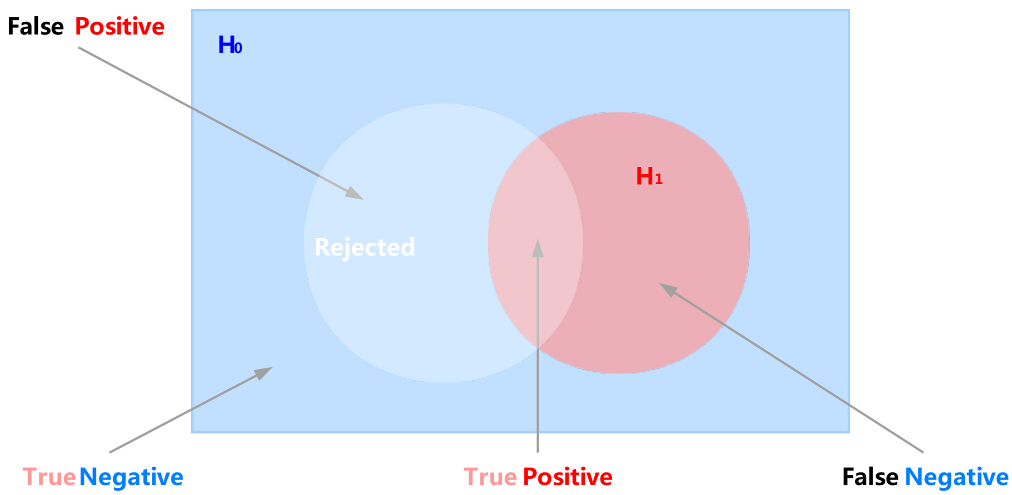

Figure 2: Relationship Between Hypothesis Test And FN,FP,TN,TP

		Truth	
		H_0	H_1
Decision	Accept H_0	a	b
	Reject H_0	c	d

Table 2: Confusion Matrix

1.5 ROC And AUC

A **receiver operating characteristic** (**ROC**) **curve** is a plot of the **sensitivity** versus (**1** − **specificity**) of a screening test, where the different points on the curve correspond to different cutoff points used to designate test-positive.

We start from a very high threshold that any individual in the dataset is tested as negative.

		Truth	
		Yes	No
Test	Yes	0	0
	No	m	n

Then we lower the threshold so that one person tests positive (suppose the one is truth-positive)

		Truth	
		Yes	No
Test	Yes	1	0
	No	$m-1$	n

Then next person tests negative (suppose the one is truth-negative)

		Truth	
		Yes	No
Test	Yes	1	1
	No	$m-1$	$n-1$

Repreat the above procedure, we can get several tables. Plot these sensitivities annd 1−specificities so that we can get the ROC.

Aera under curve (**AUC**), the area under the ROC curve is a reasonable summary of the overall diagnostic accuracy of the test. AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. $AUC = 1.0$ indicates that all true-positives have higher values than any true-negatives.

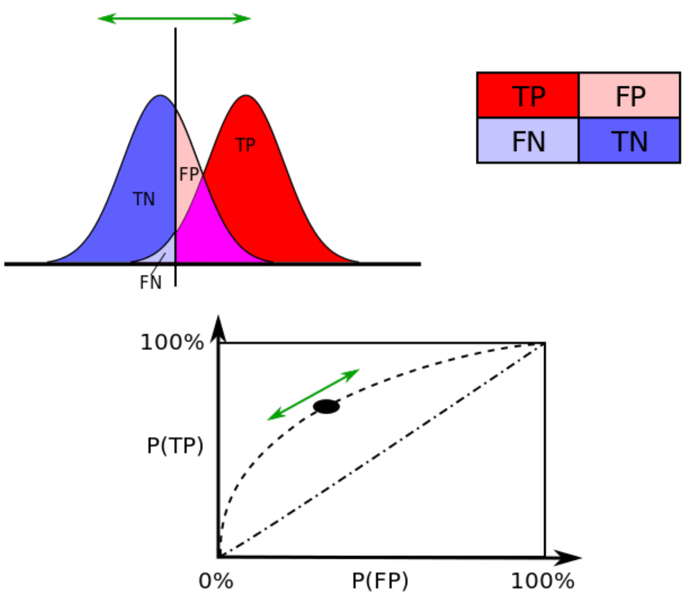

Figure 3: ROC Curves

1.6 Sample Size

Factors Affecting the Sample Size

1. The sample size increases as σ increases.
2. The sample size increases as the significance level is made smaller (α decreases).
3. The sample size increases as the required power increases ($1-\beta$ increases).
4. The sample size decreases as the absolute value of the distance between the null and alternative means ($|\mu_0 - \mu_1|$) increases.

1.7 The Relationship Between Hypothesis Testing and Confidence Intervals

Suppose we are testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. H_0 is rejected with a two-sided level α test if and only if the two-sided $100\% \times (1-\alpha)$ CI for μ does not contain μ_0. H_0 is accepted with a two-sided level α test if and only if the two-sided $100\% \times (1-\alpha)$ CI for μ does contain μ_0.

Chapter 2 Parameteric Methods

2 One-Sample Test for the Mean of a Normal Distribution

2.1 Z test (with σ known)

2.1.1 Hypothesis Test

Suppose that $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where σ^2 is known.
The Z test for

$$H_0 : \mu = \mu_0$$

versus

$$\begin{aligned} H_1 &: \mu = \mu_1 < \mu_0 \\ H_2 &: \mu = \mu_1 \neq \mu_0 \\ H_3 &: \mu = \mu_1 > \mu_0 \end{aligned}$$

The test statistic is given by

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \overset{H_0}{\sim} N(0,1)$$

Reject the null hypothesis when

$$\begin{aligned} H_1 :& Z < z_{\alpha} \\ H_2 :& |Z| > z_{1-\alpha/2} \\ H_3 :& Z > z_{1-\alpha} \end{aligned}$$

The p-value is given by

$$\begin{aligned} H_1 :& \mathbb{P}\{Z > z\} = \Phi(Z) \\ H_2 :& 2\mathbb{P}\{|Z| < z\} = 2[1 - \Phi(|Z|)] \\ H_3 :& \mathbb{P}\{Z < z\} = 1 - \Phi(Z) \end{aligned}$$

2.1.2 Power

The *Power* is given by

$$\begin{aligned} H_1 : \ 1 - \beta &= \mathbb{P}\left\{ \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha} \middle| \mu = \mu_1 \right\} \\ &= \mathbb{P}\left\{ \frac{\overline{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha} - \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}} \middle| \mu = \mu_1 \right\} \end{aligned}$$

$$= \Phi\left(z_\alpha - \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= \Phi\left(z_\alpha + \frac{|\mu_1 - \mu_0|}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$H_2: \ 1 - \beta = \mathbb{P}\left\{\left|\frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right| > z_{1-\frac{\alpha}{2}} \middle| \mu = \mu_1\right\}$$

$$= 1 - \mathbb{P}\left\{-z_{1-\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}} < \frac{\overline{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}} \middle| \mu = \mu_1\right\}$$

$$= 1 - \Phi\left(z_{1-\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) + \Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= \Phi\left(-z_{1-\frac{\alpha}{2}} + \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) + \Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= \Phi\left(-z_{1-\frac{\alpha}{2}} + \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) + \Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$\approx \Phi\left(-z_{1-\frac{\alpha}{2}} + \frac{|\mu_1 - \mu_0|}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$H_3: \ 1 - \beta = \mathbb{P}\left\{\frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha} \middle| \mu = \mu_1\right\}$$

$$= \mathbb{P}\left\{\frac{\overline{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}} \middle| \mu = \mu_1\right\}$$

$$= 1 - \Phi\left(z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= \Phi\left(z_\alpha + \frac{|\mu_1 - \mu_0|}{\frac{\sigma}{\sqrt{n}}}\right)$$

2.1.3 Sample Size Based on Power

The sample size estimation is given by

$$H_1: \ n = \frac{\sigma^2 (z_{1-\beta} + z_{1-\alpha})^2}{(\mu_0 - \mu_1)^2}$$
$$H_2: \ n = \frac{\sigma^2 (z_{1-\beta} + z_{1-\frac{\alpha}{2}})^2}{(\mu_0 - \mu_1)^2}$$
$$H_3: \ n = \frac{\sigma^2 (z_{1-\beta} + z_{1-\alpha})^2}{(\mu_0 - \mu_1)^2}$$

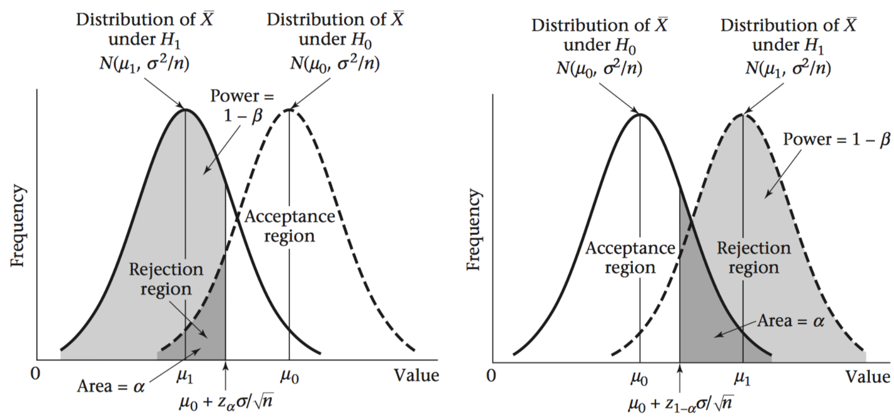

Figure 4: p-value for z test H_1 and H_3

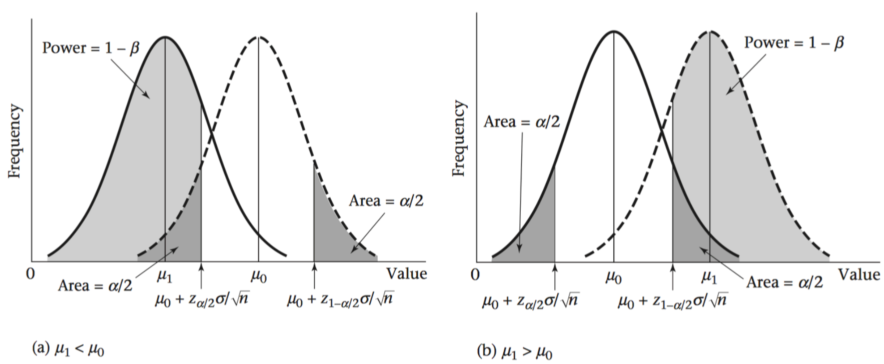

Figure 5: p-value for z test H_2

2.2 t test (with σ unknown)

2.2.1 Hypothesis Test

Suppose that $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where σ^2 is unknown.

The t test for

$$H_0: \mu = \mu_0$$

versus

$$H_1 : \mu = \mu_1 < \mu_0$$
$$H_2 : \mu = \mu_1 \neq \mu_0$$
$$H_3 : \mu = \mu_1 > \mu_0$$

The test statistic is given by

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \overset{H_0}{\sim} t_{n-1}$$

where

$$s = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N-1}}$$

Reject the null hypothesis when

$$H_1 : T < t_{n-1,\alpha}$$
$$H_2 : |T| > t_{n-1,1-\alpha/2}$$
$$H_3 : T > t_{n-1,1-\alpha}$$

The p-value is given by

$$H_1 : \mathbb{P}\{T > t_{n-1}\}$$
$$H_2 : 2\mathbb{P}\{|T| < t_{n-1}\}$$
$$H_3 : \mathbb{P}\{T < t_{n-1}\}$$

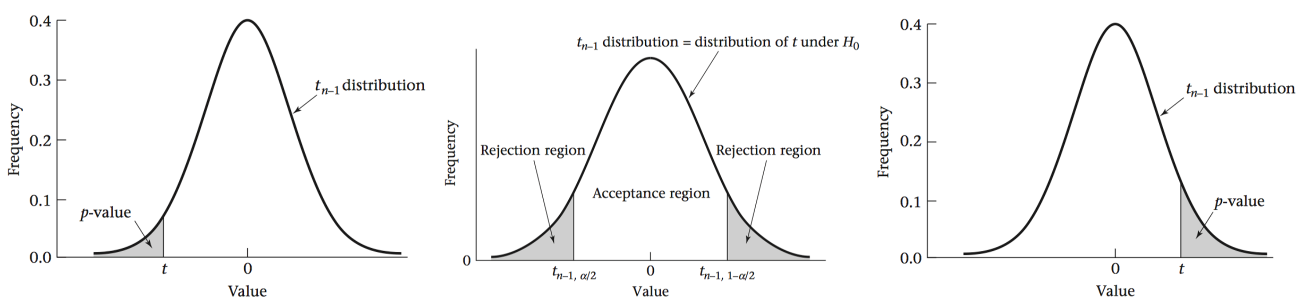

Figure 6: p-value for t test

2.2.2 Sample Size Based on CI Width

Suppose we wish to estimate the mean of a normal distribution with sample variance s^2 and require that the **two − sided** $100\% \times (1-\alpha)$ CI for μ be no wider than L. The number of subjects needed is approximately

$$n = \frac{4 z_{1-\frac{\alpha}{2}}^2 s^2}{L^2}$$

3 One-Sample χ^2 Test for the Variance of a Normal Distribution

3.1 Hypothesis Test

Suppose that $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where σ^2 is unknown.
The χ^2 test for

$$H_0 : \sigma = \sigma_0 \qquad\qquad H_1 : \sigma \neq \sigma_0$$

The test statistic is given by

$$\begin{aligned} X^2 &= \frac{(n-1)s^2}{\sigma^2} \\ &\overset{H_0}{=\!=} \frac{(n-1)s^2}{\sigma_0^2} \\ &\overset{H_0}{\sim} \chi^2_{n-1} \end{aligned}$$

where

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Reject the null hypothesis when

$$X^2 < \chi^2_{n-1,\frac{\alpha}{2}} \qquad \text{or} \qquad X^2 > \chi^2_{n-1,1-\frac{\alpha}{2}}$$

The p-value is given by

$$\begin{cases} 2\mathbb{P}\{X^2 > \chi^2_{n-1}\} & \text{, if } s^2 \leq \sigma_0^2 \\ 2\mathbb{P}\{X^2 < \chi^2_{n-1}\} & \text{, if } s^2 > \sigma_0^2 \end{cases}$$

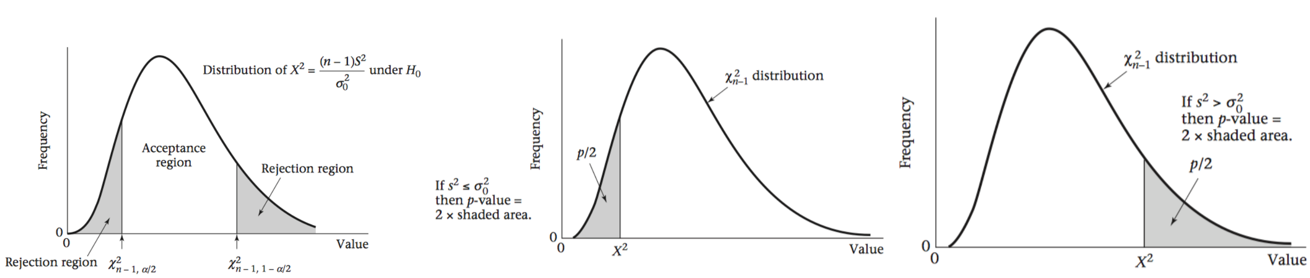

Figure 7: p-value for χ^2 test

4 One-Sample Test under Binominal Distribution

4.1 Normal-Theory Method

Suppose that $X_1, \cdots, X_n \overset{iid}{\sim} Binomial(1, p)$ where p is the prevalence rate of a disease.

Then the Binomial test is to test

$$H_0 : p = p_0 \qquad\qquad H_1 : p \neq p_0$$

When $np_0q_0 \geq 5$, normal approximation to the binomial distribution is valid.

The test statistic is given by

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0q_0}{n}}} \overset{H_0}{\sim} N(0,1)$$

where $\hat{p}$ is the sample proportion of cases.

Then the z test is appied.

4.2 Exact Method

Suppose that $X_1, \cdots, X_n \overset{iid}{\sim} Binomial(1, p)$ where p is the prevalence rate of a disease.

Then the Binomial test is to test

$$H_0 : p = p_0 \qquad\qquad H_1 : p \neq p_0$$

The test statistic is given by

$$n\hat{p} = n\overline{X} \overset{H_0}{\sim} Binomial(n, p_0)$$

The p-value is given by

$$\begin{cases} 2\mathbb{P}\{X \leq n\overline{X} | H_0\} & , \hat{p} \leq p_0 \\ 2\mathbb{P}\{X \geq n\overline{X} | H_0\} & , \hat{p} > p_0 \end{cases}$$

where $X \sim Binomial(n, p_0)$, i.e.,

$$\begin{cases} \min\left\{2\sum_{k=0}^{n\overline{X}} \binom{n}{k} p_0^k (1-p_0)^{n-k}, 1\right\} & , \hat{p} \leq p_0 \\ \min\left\{2\sum_{k=n\overline{X}}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k}, 1\right\} & , \hat{p} > p_0 \end{cases}$$

5 Two-Sample Paired t Test

Two samples are said to be **paired** when each data point in the first sample is matched and is related to a unique data point in the second sample.

Two samples are said to be **independent** when the data points in one sample are unrelated to the data points in the second sample.

Suppose that $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $Y_1, \cdots, Y_n \overset{iid}{\sim} N(\mu + \Delta, \sigma^2)$ are paired with σ^2 unkonwn. Then their differences are given by $d_i = Y_i - X_i$ (if X_i is the baseline).

The paired t test for

$$H_0 : \Delta = 0 \qquad\qquad H_1 : \Delta \neq 0$$

The test statistic is given by

$$T = \frac{\overline{d}}{\frac{s_d}{\sqrt{n}}} \overset{H_0}{\sim} t_{n-1}$$

where

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}}$$

Reject the null hypothesis when

$$|T| > t_{n-1,1-\frac{\alpha}{2}}$$

The p-value is given by

$$2\mathbb{P}\{|T| < t_{n-1}\}$$

6 Two-Sample t-test for independent Samples with Equal Variances

6.1 Hypothesis Test

Suppose that $X_1, \cdots, X_{n_1} \overset{iid}{\sim} N(\mu_1, \sigma^2)$, $Y_1, \cdots, Y_{n_2} \overset{iid}{\sim} N(\mu_2, \sigma^2)$ with σ^2 unkonwn.
The paired t test for

$$H_0 : \mu_1 = \mu_2 \qquad\qquad H_1 : \mu_1 \neq \mu_2$$

The test statistic is given by

$$Z = \frac{\overline{X} - \overline{Y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{H_0}{\sim} t_{n_1+n_2-2}$$

where

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

$$s_1 = \sqrt{\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2}$$

$$s_2 = \sqrt{\frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2}$$

Reject the null hypothesis when

$$|T| > t_{n_1+n_2-2,1-\frac{\alpha}{2}}$$

The p-value is given by

$$2\mathbb{P}\{|T| < t_{n_1+n_2-2}\}$$

6.2 Power

To test the hypothesis $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$ for the specific alternative $|\mu_1 - \mu_2| = \Delta$, with significance level α,

$$Power = \Phi\left(-z_{1-\frac{\alpha}{2}} + \frac{\Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)$$

6.3 Sample Size Based on Power

$$n_1 = \frac{(\sigma_1^2 + \frac{1}{k}\sigma_2^2)(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{\Delta^2}$$
$$n_2 = \frac{(k\sigma_1^2 + \sigma_2^2)(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{\Delta^2}$$

where $\Delta = |\mu_2 - \mu_1|$; (μ_1, σ_1^2), (μ_2, σ_2^2), are the means and variances of the two respective groups and $k = \frac{n_2}{n_1} =$ the projected ratio of the two sample sizes.

Chapter 3 Nonparametric Methods

7 One-sample (Matched) Sign Test

7.1 Normal-Theory Method

Suppose that $d_i = X_i - Y_i$, Δ is the population median of the d_i or the 50th percentile of the underlying distribution of the d_i. The actual d_i cannot be observed; we can only observe whether $d_i > 0$, $d_i < 0$, or $d_i = 0$. The people for whom $d_i = 0$ will be excluded.

To test the hypothesis

$$H_0 : \Delta = 0 \qquad\qquad H_1 : \Delta \neq 0$$

where the number of nonzero d_i's $= n \geq 20$.

The test statistic is given by

$$C = |\{d_i | d_i > 0, i = 1, \cdots, n\}| \overset{H_0}{\sim} Binomial\left(n, \frac{1}{2}\right)$$

Approximately,

$$\frac{C - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \overset{H_0}{\sim} N(0,1)$$

Reject the null hypothesis when

$$C > \frac{n}{2} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{n}{4}} \qquad \text{or} \qquad C < \frac{n}{2} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{n}{4}}$$

The p-value is given by

$$\begin{cases} 2\left[1 - \Phi\left(\frac{C - \frac{n}{2}}{\sqrt{\frac{n}{4}}}\right)\right] & \text{, if } C > \frac{n}{2} \\ 2\Phi\left(\frac{C - \frac{n}{2}}{\sqrt{\frac{n}{4}}}\right) & \text{, if } C < \frac{n}{2} \\ 1 & \text{, if } C = \frac{n}{2} \end{cases}$$

alternatively and equivalently,

$$\begin{cases} 2\left[1 - \Phi\left(\frac{C - D}{\sqrt{n}}\right)\right] & \text{, if } C \neq D \\ 1 & \text{, if } C = D \end{cases}$$

where $D = |\{d_i | d_i < 0, i = 1, 2, \cdots, n\}|$.

7.2 Exact Method

When $n < 20$, exact binomial probabilities rather than the normal approximation must be used to compute the p-value.

$$\begin{cases} 2\sum\limits_{k=C}^{n}\binom{n}{k}\left(\frac{1}{2}\right)^n & , \text{ if } C > \frac{n}{2} \\ 2\sum\limits_{k=0}^{C}\binom{n}{k}\left(\frac{1}{2}\right)^n & , \text{ if } C < \frac{n}{2} \\ 1 & , \text{ if } C = \frac{n}{2} \end{cases}$$

8 One-sample (Matched) Wilcoxon Signed-Rank Test

8.1 Ranking Procedure

1. Arrange the differences d_i in order of absolute value.
2. Count the number of differences with the same absolute value.
3. Ignore the observations where $d_i = 0$, and rank the remaining observations from 1 for the observation with the lowest absolute value, up to n for the observation with the highest absolute value.
4. If there is a group of several observations with the same absolute value, then find the lowest rank in the range $= 1+R$ and the highest rank in the range $= G+R$, where $R =$ the highest rank used prior to considering this group and $G =$ the number of differences in the range of ranks for the group. Assign the average rank = (lowest rank in the range + highest rank in the range)/2 as the rank for each difference in the group.

8.2 Hypothesis Test

Suppose that $d_1, \cdots, d_n$ have an underlying continuous symmetric distribution and R_1 is the rank sum of the positive differences.

To test the hypothesis

$$H_0 : \Delta = 0 \qquad\qquad H_1 : \Delta \neq 0$$

where the number of nonzero d_i's $= n \geq 16$.

Under H_0 and if there are no ties,

$$\mathbb{E}R_1 = \frac{1}{2} \cdot \frac{n(n+1)}{2}$$
$$VarR_1 = \frac{1}{2^2} \cdot \frac{n(n+1)(2n+1)}{6}$$

For large sample size n, R_1 follows a normal distribution.

The test statistic is given by

$$Z = \frac{R_1 - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}} \overset{H_0}{\sim} N(0,1)$$

Reject the null hypothesis when

$$|Z| > z_{1-\frac{\alpha}{2}}$$

The p-value is given by

$$2[1-\Phi(|Z|)]$$

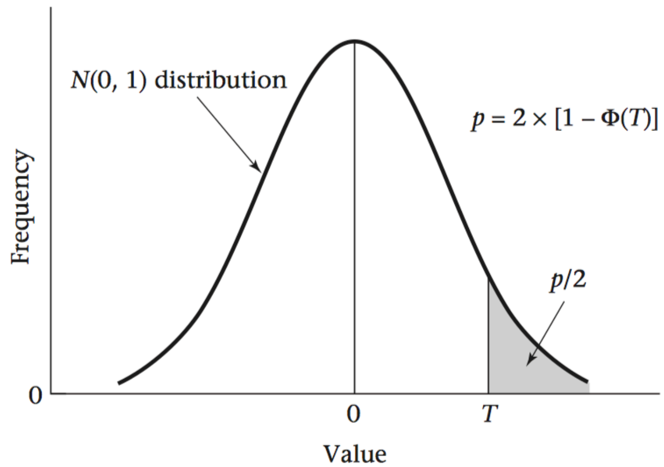

Figure 8: p-value for Wilcoxon Signed-Rank Test And Rank-Sum Test

9 Two-Sample Wilcoxon Rank-Sum Test

9.1 Ranking Procedure

1. Combine the data from the two groups, and order the values from lowest to highest.
2. Assign ranks to the individual values.
3. If a group of observations has the same value, then compute the range of ranks for the group, as was done for the signed-rank test and assign the average rank for each observation in the group.

9.2 Hypothesis Test

Suppose that $X_1, \cdots, X_{n_1} \sim F_X$ and $Y_1, \cdots, Y_{n_2} \sim F_Y$ where both n_1 and n_2 are at least 10.
To test the hypothesis

$$H_0 : F_X = F_Y \qquad H_1 : F_X(x) = F_Y(x-\Delta)$$

where $\Delta \neq 0$.

Alternatively, to test the hypothesis

$$H_0 : \Delta = 0 \qquad H_1 : \Delta \neq 0$$

The test statistic for this test is the sum of the ranks in the first sample

$$R_1 = \sum_{i=1}^{n_1} rank(X_i)$$

Under H_0 and if there are no ties,

$$\mathbb{E}R_1 = \frac{n_1(n_1+n_2+1)}{2}$$
$$VarR_1 = \frac{n_1 n_2(n_1+n_2+1)}{12}$$

For large sample size n, R_1 follows a normal distribution.

The test statistic is given by

$$Z = \frac{R_1 - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2}{12}(n_1+n_2+1)}} \overset{H_0}{\sim} N(0,1)$$

Reject the null hypothesis when

$$|Z| > z_{1-\frac{\alpha}{2}}$$

The p-value is given by

$$2[1-\Phi(|Z|)]$$

Chapter 4 Categorical Data

10 Two-Sample Test for Binomial Proportions

10.1 Normal-Theory Method

10.1.1 Hypothesis Test

Suppose that $X_1, \cdots, X_{n_1} \overset{iid}{\sim} Binomial(1, p_1)$ and $Y_1, \cdots, Y_{n_2} \overset{iid}{\sim} Binomial(1, p_2)$.
To test the hypothesis

$$H_0 : p_1 = p_2 \qquad\qquad H_1 : p_1 \neq p_2$$

Since $n_1\hat{p}_1 \sim Binomial(n_1, p_1)$, $n_2\hat{p}_2 \sim Binomial(n_2, p_2)$,

$$\begin{aligned} n_1\hat{p}_1 &\dot{\sim} N(n_1p_1, n_1p_1q_1) \\ n_2\hat{p}_2 &\dot{\sim} N(n_2p_2, n_2p_2q_2) \\ \hat{p}_1 - \hat{p}_2 &\dot{\sim} N\left(p_1 - p_2, \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}\right) \end{aligned}$$

The test statistic is given by

$$Z_{Wald} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{H_0}{\sim} N(0,1)$$

Since p is unkonwn, we estimate p by

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Therefore when $n_1\hat{p}\hat{q} \geq 5$ and $n_2\hat{p}\hat{q} \geq 5$, approximately

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{H_0}{\sim} N(0,1)$$

Reject the null hypothesis when

$$|Z| > z_{1-\frac{\alpha}{2}}$$

The p-value is given by

$$2[1 - \Phi(|Z|)]$$

10.1.2 Power

$$\begin{aligned} 1 - \beta &= \mathbb{P}\left\{ |Z| > z_{1-\frac{\alpha}{2}} \middle| H_1 \right\} \\ &= 1 - \mathbb{P}\left\{ -z_{1-\frac{\alpha}{2}} < \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{pq}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < z_{1-\frac{\alpha}{2}} \middle| H_1 \right\} \end{aligned}$$

$$
\begin{aligned}
&= 1 - \mathbb{P}\left\{ -\frac{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} z_{1-\frac{\alpha}{2}} - \frac{p_1-p_2}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} < \frac{\hat{p}_1-\hat{p}_2-(p_1-p_2)}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} \right. \\
&\qquad\qquad\qquad \left. < \frac{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} z_{1-\frac{\alpha}{2}} - \frac{p_1-p_2}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} \middle| H_1 \right\} \\
&= 1 - \Phi\left(-\frac{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} z_{1-\frac{\alpha}{2}} - \frac{p_1-p_2}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} \right) + \Phi\left(\frac{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} z_{1-\frac{\alpha}{2}} - \frac{p_1-p_2}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} \right) \\
&= \Phi\left(\frac{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1}-\frac{1}{n_2}\right)}}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} z_{1-\frac{\alpha}{2}} - \frac{p_1-p_2}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} \right) + \Phi\left(\frac{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} z_{1-\frac{\alpha}{2}} - \frac{p_1-p_2}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} \right) \\
&\approx \Phi\left(-\frac{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} z_{1-\frac{\alpha}{2}} + \frac{|p_1-p_2|}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} \right)
\end{aligned}
$$

where $\overline{p} = \dfrac{n_1p_1+n_2p_2}{n_1+n_2}$ and $\overline{q} = 1-\overline{p}$.

10.1.3 Sample Size Based on Power

$$
\begin{aligned}
n_1 &= \frac{\left(\sqrt{\frac{p_1q_1}{1}+\frac{p_2q_2}{k}}\, z_{1-\beta} + \sqrt{\bar{p}\bar{q}\left(\frac{1}{1}+\frac{1}{k}\right)}\, z_{1-\frac{\alpha}{2}}\right)^2}{(p_1-p_2)^2} \\
n_2 &= kz_1
\end{aligned}
$$

10.2 Contingency-Table Method

	Exposed	**Unexposed**	Total
Case	O_{11}	O_{12}	n_1
Control	O_{21}	O_{22}	n_2
Total	m_1	m_2	n

Table 6: A Contingency Table

The expected number of units in the (i, j) cell is given by

$$E_{ij} = \frac{n_i m_j}{n}$$

. The test is used only when $E_{ij} > 5 \; \forall \; i, j$.

It can be used as

1. A test for homogeneity of binomial proportions. In this situation, one set of margins is fixed (e.g., the rows) and the number of successes in each row is a random variable.

2. A test of independence or a test of association between two characteristics. In this setting, both sets of margins are assumed to be fixed. The number of units in one particular cell of the table [e.g., the (1, 1) cell] is a random variable, and all other cells can be determined from the fixed margins and the (1, 1) cell. Two binary random variables $X = \mathbb{1}_{\{Exposed\}}$ and $Y = \mathbb{1}_{\{Case\}}$.

To test the hypothesis

$$H_0 : p_1 = p_2 \qquad\qquad H_1 : p_1 \neq p_2$$

The test statistic is given by

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{H_0}{\sim} \chi_1^2$$

Reject the null hypothesis when

$$X^2 > \chi_{1,1-\alpha}^2$$

The p-value is given by

$$\mathbb{P}\{X^2 < \chi_1^2\}$$

The χ^2 statistic can be used

1. the rows are fixed (binomial)

2. the columns are fixed (binomial)

3. the total sample size is fixed (multinomial)

4. none are fixed (Poisson)

10.3 Fisher's Exact Test

Suppose that X and Y are the number of exposed in the case and control group respectively.

To test the hypothesis

$$H_0 : p_1 = p_2 = p$$

versus

$$H_1 : p_1 < p_2$$

$$H_2: p_1 \neq p_2$$
$$H_3: p_1 > p_2$$

Since

$$X \overset{H_0}{\sim} Binomial(n_1, p)$$
$$Y \overset{H_0}{\sim} Binomial(n_2, p)$$

we have

$$X+Y \overset{H_0}{\sim} Binomial(n_1+n_2, p)$$

The test statistic is given by

$$X|(X+Y=z) \overset{H_0}{\sim} Hypergeometric(n_1, n_2, z)$$

$$\begin{aligned}
\mathbb{P}\{X=x|X+Y=m_1\} &= \frac{\mathbb{P}\{X=x, Y=m_1-x\}}{\mathbb{P}\{X+Y=m_1\}} \\
&= \frac{\mathbb{P}\{X=x\}\mathbb{P}\{Y=m_1-x\}}{\mathbb{P}\{X+Y=m_1\}} \\
&= \frac{\binom{n_1}{x} p^x (1-p)^{n_1-x} \binom{n_2}{m_1-x} p^{m_1-x} (1-p)^{n_2-m_1+x}}{\binom{n_1+n_2}{m_1} p^{m_1} (1-p)^{n_1+n_2-m_1}} \\
&= \frac{\binom{n_1}{x}\binom{n_2}{m_1-x}}{\binom{n_1+n_2}{m_1}}
\end{aligned}$$

$$\mathbb{E}(X|X+Y=m_1) = \frac{m_1 n_1}{n}$$
$$Var(X|X+Y=m_1) = \frac{m_1 m_2 n_1 n_2}{n^2(n-1)}$$

where $m_1 + m_2 = n_1 + n_2 = n$.

Rearrange the rows and columns of the observed table so the smaller row total is in the first row and the smaller column total is in the first column, i.e. $m_1 < m_2$ and $n_1 < n_2$. Then the table is given by

a	b	n_1
c	d	n_2
m_1	m_2	n

Table 7: The Rearranged Observed Contingency Table

Enumerate $k = \min\{m_1, n_1\}$ all possible tables with the same row and column margins as the observed table. Suppose that we get k tables and each table looks like

a_i	b_i	n_1
c_i	d_i	n_2
m_1	m_2	n

Table 8: The ith Contingency Table

Then compute the exact probability of each table enumerated:

$$P_i = \mathbb{P}\{X = a_i | X + Y = m_1\} = \frac{\binom{n_1}{a_i}\binom{n_2}{m_1 - a_i}}{\binom{n_1+n_2}{m_1}}$$

Suppose the observed table is the kth table.

The p-value is given by

$$H_1: \sum_{i=0}^{a} P_i$$

$$H_2: \min\{2\sum_{i=0}^{a} P_i, 2\sum_{i=a}^{k} P_i, 1\}$$

$$H_3: \sum_{i=a}^{k} P_i$$

11 Two-Sample Test for Binomial Proportions for Matched-Pair Data (McNemar’s Test)

11.1 Matched Pair

A **concordant pair** is a matched pair in which the outcome is the same for each member of the pair.

A **discordant pair** is a matched pair in which the outcomes differ for the members of the pair.

A **type A discordant pair** is a discordant pair in which the treatment A member of the pair has the event and the treatment B member does not. Similarly, a **type B discordant pair** is a discordant pair in which the treatment B member of the pair has the event and the treatment A member does not.

11.2 Normal-Theory Test

11.2.1 Hypothesis Test

Type A	Type B		Total
	Case	**Control**	
Case	n_{11}	n_{12}	n_1
Control	n_{21}	n_{22}	n_2
Total	m_1	m_2	n

Table 9: A Contingency Table With The Matched Pair

Suppose that of n_D discordant pairs, n_A are type A. Let $p =$ the probability that a discordant pair is of type A.

$$n_D = n_{12} + n_{21}$$
$$n_A = n_{12}$$

To test the hypothesis

$$H_0 : p = \frac{1}{2} \qquad H_1 : p \neq \frac{1}{2}$$

Given that the observed number of discordant pairs $= n_D$, under H_0

$$\mathbb{E}n_A = \frac{n_D}{2}$$
$$Varn_A = \frac{n_D}{4}$$

When $n_D \geq 20$, the normal approximation gives the test statistic

$$X^2 = \frac{\left(n_A - \frac{n_D}{2}\right)^2}{\frac{n_D}{4}} \overset{H_0}{\sim} \chi_1^2$$

Reject the null hypothesis when

$$X^2 > \chi^2_{1,1-\alpha}$$

The p-value is given by

$$\mathbb{P}\{X^2 \leq \chi_1^2\}$$

11.2.2 Power

Since approximately

$$\hat{p}_A \sim N\left(p_A, \frac{p_A q_A}{n_D}\right)$$

,

$$Power = 1 - \beta$$
$$= \mathbb{P}\left\{\left|\frac{n_A - \frac{n_D}{2}}{\sqrt{\frac{n_D}{2}}}\right| > z_{1-\frac{\alpha}{2}} \middle| H_1\right\}$$

$$
\begin{aligned}
&= \mathbb{P}\left\{ \left| \frac{\hat{p}_A - \frac{1}{2}}{\sqrt{\frac{1}{4np_D}}} \right| > z_{1-\frac{\alpha}{2}} \middle| H_1 \right\} \\
&= 1 - \mathbb{P}\left\{ \left| \frac{\hat{p}_A - \frac{1}{2}}{\sqrt{\frac{1}{4np_D}}} \right| < z_{1-\frac{\alpha}{2}} \middle| H_1 \right\} \\
&= 1 - \mathbb{P}\left\{ -\frac{1}{2\sqrt{p_A q_A}} z_{1-\frac{\alpha}{2}} - \frac{p_A - \frac{1}{2}}{2\sqrt{\frac{p_A q_A}{np_D}}} < \frac{\hat{p}_A - p_A}{\sqrt{\frac{p_A q_A}{np_D}}} < \frac{1}{2\sqrt{p_A q_A}} z_{1-\frac{\alpha}{2}} - \frac{p_A - \frac{1}{2}}{2\sqrt{\frac{p_A q_A}{np_D}}} \middle| H_1 \right\} \\
&= \Phi\left(-\frac{1}{2\sqrt{p_A q_A}} z_{1-\frac{\alpha}{2}} + \frac{p_A - \frac{1}{2}}{\sqrt{\frac{p_A q_A}{np_D}}} \right) + \Phi\left(-\frac{1}{2\sqrt{p_A q_A}} z_{1-\frac{\alpha}{2}} - \frac{p_A - \frac{1}{2}}{\sqrt{\frac{p_A q_A}{np_D}}} \right) \\
&\approx \Phi\left(-\frac{1}{2\sqrt{p_A q_A}} z_{1-\frac{\alpha}{2}} + \frac{\left| p_A - \frac{1}{2} \right|}{\sqrt{\frac{p_A q_A}{np_D}}}) \right)
\end{aligned}
$$

where n is the number of pairs.

11.2.3 Sample Size

The number of pair is

$$
n = \frac{(z_{1-\frac{\alpha}{2}} + 2z_{1-\beta}\sqrt{p_A q_A})^2}{4\left(p_A - \frac{1}{2}\right)^2 p_D}
$$

or equivalently the number of individuals should be $2n$.

11.3 Exact Test

To test the hypothesis

$$
H_0 : p = \frac{1}{2} \qquad\qquad H_1 : p \neq \frac{1}{2}
$$

The test statistic is given by

$$
n_A \overset{H_0}{\sim} Binomial\left(n_D, \frac{1}{2}\right)
$$

The p-value is given by

$$
2\min\left\{ \sum_{i=0}^{n_A} \binom{n_D}{i} \left(\frac{1}{2}\right)^{n_D}, \sum_{i=n_A}^{n_D} \binom{n_D}{i} \left(\frac{1}{2}\right)^{n_D} \right\}
$$

Chapter 5 Design and Analysis Techniques for Epidemiologic Studies

12 Study Design

	Disease	**No Disease**	Total
Exposed	a	b	n_1
Not Exposed	c	d	n_2
Total	m_1	m_2	n

Table 10: Hypothetical Exposure–Disease Relationship

1. Prospective study, or cohort study
2. Retrospective study, or case–control study
3. Cross-sectional study, or prevalence study

13 Measures of Effect for Categorical Data

Let $p_1 =$ probability of developing disease for exposed individuals and $p_2 =$ probability of developing disease for unexposed individuals.

13.1 Risk Difference

The **risk difference** (**RD**) is defined as $p_1 - p_2$.

When $n_1\hat{p}_1\hat{q}_1 \geq 5,\ n_2\hat{p}_2\hat{q}_2 \geq 5$, approximately,

$$\hat{p}_1 \sim N\left(p_1, \frac{p_1q_1}{n_1}\right)$$
$$\hat{p}_2 \sim N\left(p_2, \frac{p_2q_2}{n_2}\right)$$

we have

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}\right)$$

13.2 Relative Risk

The **risk ratio** (**RR**) or **relative risk** is defined as $\dfrac{p_1}{p_2}$.

By delta method,

$$Var[f(X)] \approx [f'(X)]^2 Var(X)$$

we have

$$\begin{aligned}
Var(\ln \hat{p}_1) &\approx \frac{1}{\hat{p}_1^2} Var(\hat{p}_1) \\
&= \frac{\hat{q}_1}{\hat{p}_1 n_1} \\
&= \frac{b}{a n_1} \\
Var(\ln \hat{p}_2) &\approx \frac{1}{\hat{p}_2^2} Var(\hat{p}_2) \\
&= \frac{\hat{q}_2}{\hat{p}_2 n_2} \\
&= \frac{d}{c n_2}
\end{aligned}$$

It follows that

$$Var(\ln \hat{RR}) = \frac{b}{a n_1} + \frac{d}{c n_2}$$

13.3 Odds Ratio

13.3.1 Point and Interval Estimation

If the probability of a success $= p$, then **the odds in favor of success** $= \dfrac{p}{1-p}$.

The **Odds Ratio** is defined as

$$OR = \frac{\frac{p_1}{q_1}}{\frac{p_2}{q_2}} = \frac{p_1 q_2}{p_2 q_1}$$

and it is estimated by

$$\hat{OR} = \frac{\hat{p}_1 \hat{q}_2}{\hat{p}_2 \hat{q}_1} = \frac{ad}{bc}$$

By Woolf method, approximately,

$$Var(\ln \hat{OR}) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

$$\begin{aligned}
\ln \hat{OR} &= \ln \frac{\hat{p}_1}{\hat{q}_1} + \ln \frac{\hat{p}_2}{\hat{q}_2} \\
&= \ln \frac{\hat{p}_1}{1-\hat{p}_1} + \ln \frac{\hat{p}_2}{1-\hat{p}_2} \\
\frac{\mathrm{d}}{\mathrm{d}x}\left(\ln \frac{x}{1-x}\right) &= \frac{1-x}{x} \cdot \frac{1}{(1-x)^2} \\
&= \frac{1}{x(1-x)} \\
Var\left(\ln \frac{\hat{p}_1}{1-\hat{p}_1}\right) &= \frac{1}{\hat{p}_1^2 \hat{q}_1^2} Var(\hat{p}_1)
\end{aligned}$$

$$
\begin{aligned}
&\approx \frac{1}{\hat{p}_1^2\hat{q}_1^2}\frac{\hat{p}_1\hat{q}_1}{a+b} \\
&= \frac{1}{(a+b)\hat{p}_1\hat{q}_1} \\
&= \frac{1}{a}+\frac{1}{b} \\
Var\left(\ln\frac{\hat{p}_2}{1-\hat{p}_2}\right) &= \frac{1}{\hat{p}_2^2\hat{q}_2^2}Var(\hat{p}_2) \\
&\approx \frac{1}{\hat{p}_2^2\hat{q}_2^2}\frac{\hat{p}_2\hat{q}_2}{c+d} \\
&= \frac{1}{(c+d)\hat{p}_2\hat{q}_2} \\
&= \frac{1}{c}+\frac{1}{d} \\
Var(\ln\hat{OR}) &\approx \frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}
\end{aligned}
$$

In a prospective or a cross-sectional study, if $n_1\hat{p}_1\hat{q}_1 \geq 5$ and $n_2\hat{p}_2\hat{q}_2 \geq 5$; In a case-control study, if $m_1\hat{p}_1^*\hat{q}_1^* \geq 5$ and $m_2\hat{p}_2^*\hat{q}_2^* \geq 5$, then the approximate $100\%(1-\alpha)$ CI for OR is given by

$$(e^{c_1}, e^{c_2})$$

where

$$
\begin{aligned}
c_1 &= \ln\hat{OR} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \\
c_2 &= \ln\hat{OR} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}
\end{aligned}
$$

and $\hat{p}_1^* = \frac{a}{m_1}$, $\hat{p}_2^* = \frac{b}{m_2}$.

If the disease under study is rare, then OR and its associated $100\% \times (1-\alpha)$ CI can be interpreted as approximate point and interval estimates of the RR. This is particularly important in case–control studies in which no direct estimate of the RR is available.

14 Inference for Stratified Categorical Data

14.1 The Mantel-Haenszel Test

	Exposure		
	Yes	No	Total

		Exposure		
		Yes	No	Total
Disease	Yes	a_i	b_i	a_i+b_i
	No	c_i	d_i	c_i+d_i
		a_i+c_i	b_i+d_i	n_i

Table 11: The ith Contingency Table

Based on our work on Fisher's exact test, the distribution of a_i follows a hypergeometric distribution. Since

$$\begin{aligned}\mathbb{E}O_i &= E_i \\ &= \frac{(a_i+b_i)(a_i+c_i)}{n_i} \\ Var(O_i) &= \frac{(a_i+b_i)(a_i+c_i)(b_i+d_i)(c_i+d_i)}{n_i^2(n_i-1)}\end{aligned}$$

then the expectation and the variance of $O = \sum_{i=1}^{k} O_i$ are given by $E = \sum_{i=1}^{k} E_i$ and $V = \sum_{i=1}^{k} V_i$ respectively.

When $V > 5$, Mantel-Haenszel Test is to test

$$H_0 : OR = 1 \qquad H_1 : OR \neq 1$$

after controlling for one or more confounding variables.

The test statistic is given by

$$X^2 = \frac{(O-E)^2}{V} \overset{H_0}{\sim} \chi_1^2$$

Reject the null hypothesis when

$$X^2 > \chi_{1,1-\alpha}^2$$

The p-value is given by

$$\mathbb{P}\{X^2 < \chi_1^2\}$$

14.2 Estimation of the Odds Ratio for Stratified Data

The Mantel-Haenszel estimator of the common OR is given by

$$\hat{OR}_{MH} = \frac{\sum_{i=1}^{k} \frac{a_i d_i}{n_i}}{\sum_{i=1}^{k} \frac{b_i c_i}{n_i}}$$

14.3 Effect Modification

Chi-Square Test for Homogeneity of ORs over Different Strata (Woolf Method) is to test

$$H_0 : OR_1 = \cdots = OR_k \qquad H_1 : \exists\ i \neq j,\ OR_i \neq OR_j$$

The test statistic is given by

$$
\begin{aligned}
X_{HOM}^2 &= \sum_{i=1}^{k} w_i(\ln \hat{OR}_i - \overline{\ln OR})^2 \\
&= \sum_{i=1}^{k} w_i(\ln \hat{OR}_i) - \frac{\left(\sum_{i=1}^{k} w_i \ln \hat{OR}_i\right)^2}{\sum_{i=1}^{k} w_i} \\
&\overset{H_0}{\sim} \chi_{k-1}^2
\end{aligned}
$$

where

$$
\begin{aligned}
\overline{\ln OR} &= \frac{\sum_{i=1}^{k} w_i \ln \hat{OR}_i}{\sum_{i=1}^{k} w_i} \\
w_i &= \frac{1}{\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}}
\end{aligned}
$$

Reject the null hypothesis when

$$X^2 > \chi_{k-1,1-\alpha}^2$$

The p-value is given by

$$\mathbb{P}\{X^2 < \chi_{k-1}^2\}$$

15 Multiple Logistic Regression

15.1 General Model

If $x_1, \cdots, x_k$ are a collection of independent variables and y is a binomial-outcome variable with probability of success $= p$, then the multiple logistic-regression model is given by

$$logit(p) = \ln \frac{p}{1-p} = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

or, equivalently, if we solve for p, then the model can be expressed in the form

$$p = \frac{e^{\alpha + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\alpha + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

Suppose we consider two individuals with different values of the independent variables, where only the jth independent variable is different.

$$
\begin{aligned}
logit(p_A) &= \alpha + \beta_1 x_1 + \cdots + \beta_j(x_j + \Delta) + \cdots + \beta_k x_k \\
logit(p_B) &= \alpha + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_k x_k
\end{aligned}
$$

then

$$logit(p_A) - logit(p_B) = \beta_j \Delta$$

i.e.,

$$\frac{\frac{p_A}{1-p_A}}{\frac{p_B}{1-p_B}} = e^{\beta_j \Delta}$$

i.e.,

$$OR = \frac{Odds_A}{Odds_B} = e^{\beta_j \Delta}$$

and

$$\hat{O}R = e^{\hat{\beta}_j \Delta}$$

the estimated OR in favor of success for the A versus B.

When x_j is a dichotomous variable, then $\Delta = 1$.

15.2 Relationship Between Logistic-Regression Analysis and Contingency-Table Analysis

Suppose we have a dichotomous disease variable D and a single dichotomous exposure variable E, derived from either a prospective, cross-sectional, or case−control study design, and that the 2×2 table relating disease to exposure is given by

		E		
		Yes	No	Total
D	Yes	a	b	$a+b$
	No	c	d	$c+d$
		$a+c$	$b+d$	n

Table 12: The Contingency Table

We can estimate the OR relating D to E in either of two equivalent ways:

1. We can compute the OR directly from the 2×2 table $= \dfrac{ad}{bc}$
2. We can set up a logistic-regression model of the form

$$\ln \frac{p}{1-p} = \alpha + \beta E$$

where $p =$ probability of disease D given exposure status E and where we estimate the OR by $e^{\hat{\beta}}$.

For prospective or cross-sectional studies, we can estimate the probability of disease among exposed (p_E) subjects and unexposed ($p_{\overline{E}}$) subjects in either E of two equivalent ways:

1. From the 2×2 table, we have

$$p_E = \frac{a}{a+c}$$
$$p_{\overline{E}} = \frac{b}{b+d}$$

2. From the logistic-regression model,

$$p_E = \frac{e^{\hat{\alpha}+\hat{\beta}}}{1+e^{\hat{\alpha}+\hat{\beta}}}$$
$$p_{\overline{E}} = \frac{e^{\hat{\alpha}}}{1+e^{\hat{\alpha}}}$$

For case−control studies, it is impossible to estimate absolute probabilities of disease unless the sampling fraction of cases and controls from the reference population is known, which is almost always not the case.

16 Equivalence Studies

16.1 Definition

Equivalence study is to show approximate equivalence of two experimental treatments.

Suppose p_1 is the survival rate for the standard treatment and p_2 is the survival rate for the experimental treatment.

When $\delta > 0$, the equivalence study can be descrbed as

1. **Superiority test**

   Higher means better,

   $$H_0 : p_2 \leq p_1 + \delta \qquad H_1 : p_2 > p_1 + \delta$$

   Higher means worse,

   $$H_0 : p_2 \geq p_1 - \delta \qquad H_1 : p_2 < p_1 - \delta$$

2. **Non − inferiority test**

   Higher means better,

   $$H_0 : p_2 \leq p_1 - \delta \qquad H_1 : p_2 > p_1 - \delta$$

   Higher means worse,

   $$H_0 : p_2 \geq p_1 + \delta \qquad H_1 : p_2 < p_1 + \delta$$

16.2 Superiority Test

Here we only consider the case that higher means better. For the other case, just switch p_1 and p_2, $\hat{p}_1$ and $\hat{p}_2$ respectively.

16.2.1 Inference Based on Confidence-Interval Estimation

Take one case as an example, to test

$$H_0 : p_2 \le p_1 + \delta \qquad\qquad H_1 : p_2 > p_1 + \delta$$

The test statistic is gven by

$$Z = \frac{(\hat{p}_2 - \hat{p}_1) - \delta}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \overset{H_0}{\sim} N(0,1)$$

approximately.

Since

$$\begin{aligned}
\frac{(\hat{p}_2 - \hat{p}_1) - \delta}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} &= \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1) + (p_2 - p_1) - \delta}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \\
&\overset{H_1}{=\!=} N(0,1) + \frac{(p_2 - p_1) - \delta}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}
\end{aligned}$$

under H_1, $p_1 - p_2 - \delta > 0$, we reject H_0 when Z is great enough. Therefore the rejection area is

$$Z > z_{1-\alpha}$$

A lower one-sided $100\% \times (1-\alpha)$ CI for $p_1 - p_2$ is given by

$$p_1 - p_2 > \hat{p}_1 - \hat{p}_2 - z_{1-\alpha}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Reject H_0 if the lower bound of this one-sided CI exceeds δ, i.e. we will consider the treatments as equivalent.

16.2.2 Sample-Size Estimation for Equivalence Studies

Since

$$\begin{aligned}
1 - \beta &= \mathbb{P}\{\hat{p}_1 - \hat{p}_2 - z_{1-\alpha}\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} \ge \delta\} \\
&= \mathbb{P}\left\{\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} \ge \frac{\delta - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} + z_{1-\alpha}\right\}
\end{aligned}$$

we have

$$z_\beta = \frac{\delta - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} + z_{1-\alpha}$$

Therefore,

$$\frac{1}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} = \frac{z_{1-\alpha} + z_{1-\beta}}{(p_1 - p_2) - \delta}$$

If we assume the experimental treatment sample size (n_2) is k times as large as the standard treatment sample size (n_1), we obtain

$$\frac{1}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{kn_1}}} = \frac{z_{1-\alpha} + z_{1-\beta}}{(p_1 - p_2) - \delta}$$

Solve for n_1 yields

$$n_1 = \frac{\left(p_1q_1 + \frac{p_2q_2}{k}\right)(z_{1-\alpha} + z_{1-\beta})^2}{[\delta - (p_1 - p_2)]^2}$$
$$n_2 = kn_1$$

16.3 Non-inferiority test

Here we only consider the case that higher means better. For the other case, just switch p_1 and p_2, $\hat{p}_1$ and $\hat{p}_2$ respectively.

16.3.1 Inference Based on Confidence-Interval Estimation

Take one case as an example, to test

$$H_0 : p_2 \leq p_1 - \delta \qquad H_1 : p_2 > p_1 - \delta$$

The test statistic is gven by

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \delta}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \overset{H_0}{\sim} N(0,1)$$

approximately.

Since

$$\begin{aligned}\frac{(\hat{p}_1 - \hat{p}_2) - \delta}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} &= \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2) + (p_1 - p_2) - \delta}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \\ &\overset{H_1}{=\!=} N(0,1) + \frac{(p_1 - p_2) - \delta}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}\end{aligned}$$

under H_1, $p_1 - p_2 - \delta < 0$, we reject H_0 when Z is small enough. Therefore the rejection area is

$$Z < -z_{1-\alpha}$$

A lower one-sided $100\% \times (1-\alpha)$ CI for $p_1 - p_2$ is given by

$$p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + z_{1-\alpha}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Reject H_0 if the upper bound of this one-sided CI does not exceed δ, i.e. we will consider the treatments as equivalent.

16.3.2 Sample-Size Estimation for Equivalence Studies

Since

$$1 - \beta = \mathbb{P}\{\hat{p}_1 - \hat{p}_2 + z_{1-\alpha}\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} \leq \delta\}$$

$$= \mathbb{P}\left\{ \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} \leq \frac{\delta - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} - z_{1-\alpha} \right\}$$

we have

$$z_{1-\beta} = \frac{\delta - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} - z_{1-\alpha}$$

Therefore,

$$\frac{1}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} = \frac{z_{1-\alpha} + z_{1-\beta}}{\delta - (p_1 - p_2)}$$

If we assume the experimental treatment sample size (n_2) is k times as large as the standard treatment sample size (n_1), we obtain

$$\frac{1}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{kn_1}}} = \frac{z_{1-\alpha} + z_{1-\beta}}{\delta - (p_1 - p_2)}$$

Solve for n_1 yields

$$n_1 = \frac{\left(p_1q_1 + \frac{p_2q_2}{k}\right)(z_{1-\alpha} + z_{1-\beta})^2}{[\delta - (p_1 - p_2)]^2}$$

$$n_2 = kn_1$$

Chapter 6 Multisample Inference

17 One-Way Analysis of Variance

17.1 Fixed-Effects Model

17.1.1 Interpretation

Suppose there are k groups with n_i observations in the ith group. The jth observation in the ith group will be denoted by y_{ij}. Let's assume the following model.

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

where μ is a constant, α_i is a constant specific to the ith group, and $e_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ is an error term.

1. μ represents the underlying mean of all groups taken together.
2. α_i represents the difference between the mean of the ith group and the overall mean.
3. e_{ij} represents random error about the mean $\mu + \alpha_i$ for an individual observation from the ith group.

17.1.2 Hypothesis Test

Overall F Test for One-Way ANOVA is to test the hypothesis

$$H_0 : \alpha_1 = \cdots = \alpha_k = 0 \qquad H_1 : \exists\ k_0,\ s.t.\ \alpha_{k_0} \neq 0$$

$$\begin{aligned}
SST &= SSW + SSB \\
SST &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \\
SSW &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \\
SSB &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 \\
MSW &= \frac{SSW}{n-k} \\
MSB &= \frac{SSB}{k-1} \\
n &= \sum_{i=1}^{k} n_i
\end{aligned}$$

The test statistic is given by

$$F = \frac{MSB}{MSW} \overset{H_0}{\sim} F(k-1, n-k)$$

Reject the null hypothesis when

$$F > F_{k-1,n-k,1-\alpha}$$

The p-value is given by

$$\mathbb{P}\{F < F_{k-1,n-k}\}$$

17.2 Random-Effects Model

17.2.1 Interpretation

Suppose there are k groups with n_i observations in the ith group. The jth observation in the ith group will be denoted by y_{ij}. Let's assume the following model.

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

where μ is a constant, $\alpha_i \sim N(0, \sigma_A^2)$ is specific to the ith group, and $e_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ is an error term.

1. μ represents the underlying mean of all groups taken together.
2. α_i is a random variable representing between-subject variability, which is assumed to follow an $N(0, \sigma_A^2)$ distribution
3. e_{ij} is a random variable representing within-subject variability, which follows an $N(0, \sigma^2)$ distribution and is independent of α_i and any of the other e_{ij}

17.2.2 Hypothesis Test

$$\begin{aligned} \mathbb{E}MSW &= \sigma^2 \\ \mathbb{E}MSB &= \sigma^2 + \frac{1}{k-1}\left(\sum_{i=1}^{k} n_i - \frac{\sum_{i=1}^{k} n_i^2}{\sum_{i=1}^{k} n_i}\right)\sigma_A^2 \end{aligned}$$

When $n_1 = \cdots = n_k$, we have

$$\mathbb{E}MSB = \sigma^2 + \frac{n}{k}\sigma_A^2$$

To test the hypothesis

$$H_0 : \sigma_A^2 = 0 \qquad H_1 : \sigma_A^2 > 0$$

The test statistic is given by

$$F = \frac{MSB}{MSW} \overset{H_0}{\sim} F\left(k-1, n-k\right)$$

Reject the null hypothesis when

$$F > F_{k-1,n-k,1-\alpha}$$

The p-value is given by

$$\mathbb{P}\{F < F_{k-1,n-k}\}$$

Also, the estimators are given by

$$\hat{\sigma}^2 = MSW$$

$$\hat{\sigma}_A^2 = \max\left\{\frac{MSB - MSW}{\dfrac{1}{k-1}\left(\sum_{i=1}^k n_i - \dfrac{\sum_{i=1}^k n_i^2}{\sum_{i=1}^k n_i}\right)}, 0\right\}$$

18 Meta Analysis

Study	True value of targeted parameter (unknwon)	Results	
		Estimation	Variance (known)
1	θ_1	$\hat{\theta}_1$	σ_1
2	θ_2	$\hat{\theta}_2$	σ_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	θ_k	$\hat{\theta}_k$	σ_k

Table 13: Result From Different Studies

$$\hat{\theta}_i = \theta_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma_i^2)$, $i = 1, 2, \cdots, k$ are independent.

18.1 Fixed-Effect Model

The fixed-effects model assumes that $\theta_1 = \cdots = \theta_k = \theta$, so that the model becomes

$$\hat{\theta}_i = \theta + \varepsilon_i \sim N(\theta, \sigma_i^2)$$

To test the hypothesis that

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0$$

The MLE of θ is given by

$$\hat{\theta} = \overline{\theta}_w = \frac{\sum_{i=1}^k w_i \hat{\theta}_i}{\sum_{i=1}^k w_i} \overset{H_0}{\sim} N\left(\theta_0, \frac{1}{\sum_{i=1}^k w_i}\right)$$

where

$$w_i = \frac{1}{\sigma_i^2}$$

The test statistic is given by

$$Z = \frac{\overline{\theta}_w - \theta_0}{\sqrt{\frac{1}{\sum_{i=1}^{k} w_i}}} \overset{H_0}{\sim} N(0,1)$$

Reject the null hypothesis when

$$|Z| > Z_{1-\frac{\alpha}{2}}$$

The p-value is given by

$$2[1 - \Phi(|Z|)]$$

18.2 Random-Effect Model

The random-effects model assumes that $\theta_1, \cdots, \theta_k \overset{iid}{\sim} N(\theta, \sigma_A^2)$, so that the model becomes

$$\hat{\theta}_i = \theta_i + \varepsilon_i \sim N(\theta, \sigma_A^2 + \sigma_i^2)$$

where σ_A^2 is unknown and needs to be estimated.

To test the hypothesis that

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0$$

The MLE of θ is given by

$$\hat{\theta} = \overline{\theta}_{w^*} = \frac{\sum_{i=1}^{k} w_i^* \hat{\theta}_i}{\sum_{i=1}^{k} w_i^*} \overset{H_0}{\sim} N\left(\theta_0, \frac{1}{\sum_{i=1}^{k} w_i^*}\right)$$

where $w_i^* = \dfrac{1}{\hat{\sigma}_A^2 + \sigma_i^2}$ and $\hat{\sigma}_A^2$ is the MLE of σ_A^2.

The test statistic is

$$Z = \frac{\overline{\theta}_{w^*} - \theta_0}{\sqrt{\frac{1}{\sum_{i=1}^{k} w_i^*}}} \overset{H_0}{\sim} N(0,1)$$

It can be shown that the best estimate of σ_A^2 is given by

$$\hat{\sigma}_A^2 = \max\left\{0, \frac{Q_w - (k-1)}{\left(\sum_{i=1}^{k} w_i\right)^2 - \sum_{i=1}^{k} w_i^2} \sum_{i=1}^{k} w_i\right\}$$

where

$$Q_w = \sum_{i=1}^{k} w_i (\theta_i - \overline{\theta}_w)^2$$

Reject the null hypothesis when

$$|Z| > Z_{1-\frac{\alpha}{2}}$$

The p-value is given by

$$2[1-\Phi(|Z|)]$$

18.3 Test of Odds Ratio

In either fixed-effects or random-effects model, we substitute σ_i^2 by

$$s_i^2 = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

the estimated variance of $\log OR$ for the ith study.

18.4 Test of Homogeneity of Odds Ratios

To test the hypothesis

$$H_0 : \theta_1 = \cdots = \theta_k \qquad\qquad H_1 : \exists\ i, \neq j,\ \theta_i \neq \theta_j$$

where $\theta_i =$ estimated $\log OR$ in the ith study.

The test statistic is

$$Q_w = \sum_{i=1}^{k} w_i (y_i - \bar{y}_w)^2 \overset{H_0}{\sim} \chi^2_{k-1}$$

Reject the null hypothesis when

$$Q_w > \chi^2_{k-1,1-\alpha}$$

The p-value is given by

$$\mathbb{P}\{Q_w < \chi^2_{k-1}\}$$