MATH1312: Lecture Note on Probability Theory and Mathematical Statistics

Shixiao W. Jiang

Institute of Mathematical Sciences, ShanghaiTech University, Shanghai 201210, China

jiangshx@shanghaitech.edu.cn

2024 年 11 月 18 日

Contents

Chapter 1

Confidence Interval and Hypothesis Test

1.1 Introduction to Statistical Inference and Learning

A typical statistical inference question is:
Given a sample $X_1, \dots, X_n \sim F$, how do we infer F? In some cases, we may want to infer only some feature of F such as its mean.

1.1.1 Parametric and Nonparametric models

A **statistical model** $\mathfrak{F}$ is a set of distributions (or densities or regression functions).

A **parametric model** is a set $\mathfrak{F}$ that can be parameterized by a **finite number of parameters**. For example, if we assume that the data come from a Normal distribution, then the model is

$$\mathfrak{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, -\infty < \mu < +\infty, 0 < \sigma < \infty \right\}. \tag{1.1}$$

This is a two-parameter model. We have written the density as $f(x; \mu, \sigma)$ to show that x is a value of the random variable whereas μ and σ are parameters. In general, a parametric model takes the form

$$\mathfrak{F} = \{f(x; \theta) : \theta \in \Theta\}$$

where θ is an unknown parameter (or vector of parameters) that can take values in the **parameter space** Θ. If θ is a vector but we are only interested in one component of θ, we call the remaining parameters **nuisance parameters**.

Example 1.1.1 *(Two-dimensional Parametric Estimation) Suppose that $X_1, \dots, X_n \sim F$ and we assume that the pdf $f \in \mathfrak{F}$ where $\mathfrak{F}$ is given in (1.1). The goal is to estimate the two parameters, μ and σ, from the data. If we are only interested in estimating μ, then μ is the parameter of interest and σ is a nuisance parameter.*

A **nonparametric model** is a set $\mathfrak{F}$ that cannot be parameterized by a finite number of parameters or parameterized by a large amount number of parameters such as Neural Networks. For example, $\mathfrak{F}_{\mathrm{ALL}}$ = {all CDF′s} is nonparametric.

Example 1.1.2 *(Nonparametric density Estimation) Let $X_1, \ldots, X_n$ be independent observations from a CDF F and let $f = F'$ be the pdf. Suppose we want to estimate the pdf f. It is not possible to estimate f assuming only that $F \in \mathfrak{F}_{\text{ALL}}$. We need to assume some smoothness on f. For example, we might assume that $f \in \mathfrak{F}_{\text{DENS}} \cap \mathfrak{F}_{\text{SOB}}$ where $\mathfrak{F}_{\text{DENS}}$ is the set of all probability density functions and*

$$\mathfrak{F}_{\text{SOB}} = \left\{ f : \int \left(f''(x)\right)^2 dx < \infty \right\},$$

is the Sobolev space which is set of functions that are not "too wiggly".

1.1.2 Regression

Example 1.1.3 *(Regression, prediction, and classification). Suppose we observe pairs of data $(X_1, Y_1), \ldots,$ (X_n, Y_n). Perhaps X_i is the blood pressure of subject i and Y_i is how long they live. X is called a predictor or regressor or* ***feature*** *or* ***independent variable****. Y is called the outcome or the* ***response variable*** *or the* ***dependent variable****. We call $r(x) = E(Y|X = x)$ the regression function. If we assume that $r \in \mathfrak{F}$ where $\mathfrak{F}$ is finite dimensional — the set of straight lines for example — then we have a parametric regression model. If we assume that $r \in \mathfrak{F}$ where $\mathfrak{F}$ is not finite dimensional such as Neural Networks then we have a nonparametric regression model. The goal of predicting Y for a new patient based on their X value is called* ***prediction****. If Y is discrete (for example, live or die) then prediction is instead called* ***classification****. If our goal is to estimate the function r, then we call this* ***regression*** *or curve estimation. Regression models are sometimes written as*

$$Y = r(X) + \epsilon,$$

where $E\epsilon = 0$.

1.1.3 Frequentists and Bayesians

Frequentists and Bayesians. There are many approaches to statistical inference. The two dominant approaches are called frequentist inference and Bayesian inference.

Some Notation. If $\mathfrak{F} = f(x; \theta) : \theta \in \Theta$ is a parametric model, we write $P(X \in A) = \int_A f(x; \theta)dx$ and $E(r(X)) = \int r(x)f(x; \theta)dx$. The probability or expectation is with respect to $f(x; \theta)$; it does not mean we are averaging over θ. Similarly, we write Var for the variance.

1.2 Fundamental Concepts in Inference and Point Estimation

1.2.1 Introduction

Many inferential problems can be identified as being one of three types: **point estimation, confidence sets, or hypothesis testing**. The basic idea is to **use point estimation or confidence sets when we know nothing about the parameters** at the beginning. We can **apply hypothesis tests when we know something about the parameters but we have some doubts, suspicions, or requirements for the parameters**. In other words, we can understand hypothesis tests as a disproof approach in the sense of probability view. Parametric hypothesis tests are designed for unknown parameters in parametric models. Nonparametric hypothesis tests are designed for such as distributions and independence in nonparametric models. Here, we give a brief introduction to the ideas.

Table 1.1: Comparison between frequentists and Bayesians.

	frequentists	Bayesians
Hypothesis testing	Set null and alternative hypotheses and use statistical tests to assess evidence against the null.	Consider **prior** beliefs when forming hypotheses.
Probability interpretation	Frame probability in terms of objective, long-term frequencies.	Interpret probabilities subjectively and update them as new data is collected.
Sampling	Emphasize random sampling and often require fixed sample sizes.	Can adapt well to varying sample sizes since Bayesians update their beliefs as more (observed) data comes in.
Assumption	Parameters that you estimate are fixed and are a single point while samples are random variables	There is a probability distribution around both the parameters and the samples.
The regime for application	Law of large number using a large amount of data.	Probability is degree of belief. Applicable when one has limited data, priors, and computing power.

1.2.2 Point Estimation

Point estimation refers to providing a single "best guess" of some quantity of interest. The quantity of interest could be a parameter in a parametric model, a CDF F, a probability density function f, a regression function r, or a prediction for a future value Y of some random variable. Standard estimators include **moments estimator, maximum likelihood estimator (MLE), maximum a posteriori (MAP)**, etc.

By convention, we denote a point estimate of θ *by* $\widehat{\theta}$ *or* $\widehat{\theta}_n$*. Remember that* θ *is a fixed, unknown quantity. The estimate* $\widehat{\theta}$ *depends on the data so* $\widehat{\theta}$ *is a random variable.*

More formally, let $X_1, ..., X_n$ be n i.i.d. data points from some distribution F. A point estimator $\widehat{\theta}_n$ of a parameter θ is some function of $X_1, ..., X_n$:

$$\widehat{\theta}_n = g(X_1, ..., X_n).$$

The bias of an estimator is defined by

$$\text{bias}(\widehat{\theta}_n) = E(\widehat{\theta}_n) - \theta.$$

We say that $\widehat{\theta}_n$ is unbiased if $E(\widehat{\theta}_n) = \theta$. Unbiasedness used to receive much attention but these days is considered less important; many of the estimators we will use are biased.

A reasonable requirement for an estimator is that it should converge to the true parameter value as we collect more and more data. This requirement is quantified by the following definition:

Definition 1.2.1 *A point estimator* $\widehat{\theta}_n$ *of a parameter* θ *is (weakly)* ***consistent*** *if* $\widehat{\theta}_n \xrightarrow{P} \theta$*.*

The distribution of $\widehat{\theta}_n$ is called the **sampling distribution**. The standard deviation of $\widehat{\theta}_n$ is called the **standard error**, denoted by σ:

$$\sigma = \sigma(\widehat{\theta}_n) = \sqrt{Var(\widehat{\theta}_n)}.$$

Often, the standard error depends on the unknown F. In those cases, σ is an unknown quantity but we usually can estimate it. The estimated standard error is denoted by $\widehat{\sigma}$.

Example 1.2.2 *Let $X_1, ..., X_n \sim$ Bernoulli(p) with unknown p and let $\widehat{p}_n = n^{-1}\sum_i X_i := \overline{X}_n$. Then $E(\widehat{p}_n) = n^{-1}\sum_i E(X_i) = p$ so $\widehat{p}_n$ is unbiased. The standard error is $\sigma = \sqrt{V(\widehat{p}_n)} = \sqrt{p(1-p)/n}$. The estimated standard error is $\widehat{\sigma} = \sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}$.*

The quality of a point estimate is sometimes assessed by the **mean squared error, or MSE** defined by

$$\text{MSE} = E(\widehat{\theta}_n - \theta)^2.$$

Keep in mind that $E(\cdot)$ refers to expectation with respect to the distribution

$$f(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta),$$

that generated the data.

Theorem 1.2.3 *The MSE can be written as*

$$\text{MSE} = \text{bias}^2(\widehat{\theta}_n) + Var(\widehat{\theta}_n).$$

Proof. Let $\overline{\theta}_n = E(\widehat{\theta}_n)$. Then

$$\begin{aligned} E(\widehat{\theta}_n - \theta)^2 &= E(\widehat{\theta}_n - \overline{\theta}_n + \overline{\theta}_n - \theta)^2 \\ &= E(\widehat{\theta}_n - \overline{\theta}_n)^2 + 2(\overline{\theta}_n - \theta)E(\widehat{\theta}_n - \overline{\theta}_n) + E(\overline{\theta}_n - \theta)^2 \\ &= (\overline{\theta}_n - \theta)^2 + E(\widehat{\theta}_n - \overline{\theta}_n)^2 \\ &= \text{bias}^2(\widehat{\theta}_n) + Var(\widehat{\theta}_n). \end{aligned}$$

There is a bias-variance tradeoff. ■

Theorem 1.2.4 *If* bias $\to 0$ *and* $\sigma \to 0$ *as* $n \to \infty$ *then* $\widehat{\theta}_n$ *is consistent, that is,* $\widehat{\theta}_n \xrightarrow{P} \theta$.

Proof. If bias $\to 0$ and $\sigma \to 0$ then, by above Theorem, MSE $\to 0$. It follows that $\widehat{\theta}_n \xrightarrow{qm} \theta$. Thus the result follows. ■

Example 1.2.5 *Returning to the coin flipping example, we have that $E(\widehat{p}_n) = n^{-1}\sum_i E(X_i) = p$ so $\widehat{p}_n$ is unbiased. The standard error is $\sigma = \sqrt{V(\widehat{p}_n)} = \sqrt{p(1-p)/n} \to 0$. Hence, $\widehat{p}_n$ is a consistent estimator.*

Many of the estimators we will encounter will turn out to have, approximately, a normal distribution.

Definition 1.2.6 *An estimator is* ***asymptotically normal*** *if*

$$\frac{\widehat{\theta}_n - \theta}{\sigma} \xrightarrow{d} N(0,1). \tag{1.2}$$

1.3 Confidence Sets

1.3.1 Definition

（置信度为$1-\alpha$的置信区间）

Definition 1.3.1 *A $1-\alpha$ confidence interval for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, ..., X_n)$ and $b = b(X_1, ..., X_n)$ are two functions or two statistics of the random samples $X_1, \ldots, X_n$ from a distribution such that*

$$P(\theta \in C_n) \geq 1-\alpha, \quad \textit{for all } \theta \in \Theta.$$

In words, (a, b) traps θ with probability $1-\alpha$. We call $1-\alpha$ the coverage of the confidence interval.

Remark 1.3.2 *Warning! C_n is random and θ is fixed. There is much confusion about how to interpret a confidence interval. A confidence interval is not a probability statement about θ since θ is a fixed quantity, not a random variable.*

Remark 1.3.3 *Commonly, people use 95 percent confidence intervals, which corresponds to choosing $\alpha = 0.05$. If θ is a vector then we use a confidence set (such as a sphere or an ellipse) instead of an interval.*

（枢轴量）shu zhou

Definition 1.3.4 *A function $T = T(X_1, X_2, \ldots, X_n, \theta)$ is called a pivotal quantity if it is bijective in θ and has a completely known distribution.*

Usually, a pivotal quantity is not a statistic. However, besides the parameter θ that we want to estimate, a pivotal quantity T is not allowed to contain other unknown parameters.

The procedure for find the confidence interval is as follows:

(1) Find a suitable pivotal quantity $T(X_1, X_2, \ldots, X_n, \theta)$.

(2) Given the coefficient α, find the corresponding quantile from the distribution of T such that the probability between two quantiles is $1-\alpha$.

(3) Transform the inequality and calculate the confidence interval $[a, b]$ for θ.

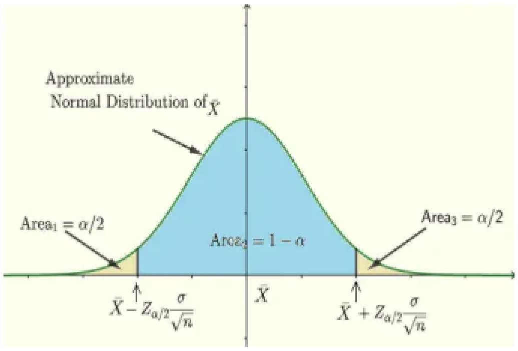


Figure 1.1: The confidence interval for the normal distribution.

Example 1.3.5 *(Z-distribution). Suppose that $X \sim N(\mu, \sigma^2)$, where σ^2 is a known constant but μ is an unknown parameter. Let $X_1, \ldots, X_n$ be a random sample from X. Can we find a coefficient $1-\alpha$ confidence interval for μ?*

Solution. Note that $\overline{X}$ is an unbiased estimator of μ, and $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. (Here σ is a known constant, and Z does not contain unknown parameters other than μ.)

Note that the pdf of a standard normal distribution is an even function. We have

$$P\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| \geq z_{\alpha/2}\right) = 2P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \geq z_{\alpha/2}\right) = 2 \cdot \frac{\alpha}{2} = \alpha.$$

See Fig. 1.1. Thus a $1-\alpha$ confidence interval for μ is

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

Remark 1.3.6 *A coefficient $1-\alpha$ confidence interval for θ may be not unique. In above example, the numbers $z_{1-\alpha/2}$ and $z_{\alpha/2}$ can be replaced by any numbers such that $0 < \alpha_1, \alpha_2 < 1$ and $\alpha_1 - \alpha_2 = 1-\alpha$. Indeed,*

$$P(z_{\alpha_1} < Z < z_{\alpha_2}) = \alpha_1 - \alpha_2 = 1-\alpha.$$

However, in this example, the choice of $\alpha_1 = 1-\alpha/2$ and $\alpha_2 = \alpha/2$ gives the shortest coefficient $1-\alpha$ confidence interval (which can be deduced from the symmetry of the pdf or seen from the graph).

Proof. Solution. Notice that

$$P(z_\beta \leq Z \leq z_{1-\alpha+\beta}) = 1-\alpha+\beta-\beta = 1-\alpha,$$

where β is not unique to be $\alpha/2$.

We want to find the shortest interval,

$$\min_\beta \left(z_{1-\alpha+\beta} - z_\beta\right).$$

We have that

$$\Phi(z_\beta) = \beta, \quad \Phi(z_{1-\alpha+\beta}) = 1-\alpha+\beta.$$

where Φ is the distribution function. We take the derivative w.r.t. β:

$$\begin{aligned} \varphi(z_\beta)\frac{dz_\beta}{d\beta} &= 1, \\ z'_\beta &= \frac{1}{\varphi(z_\beta)}. \end{aligned}$$

Similarly,

$$z'_{1-\alpha+\beta} = \frac{1}{\varphi(z_{1-\alpha+\beta})}.$$

The minimization problem is equivalent to

$$z'_{1-\alpha+\beta} - z'_\beta = \frac{1}{\varphi(z_{1-\alpha+\beta})} - \frac{1}{\varphi(z_\beta)} = 0.$$

Hence

$$\varphi(z_{1-\alpha+\beta}) = \varphi(z_\beta).$$

Since pdf φ is an even function, this gives $z_{1-\alpha+\beta} = -z_\beta$. Thus, $\beta = \frac{\alpha}{2}$ gives the shortest interval. ■

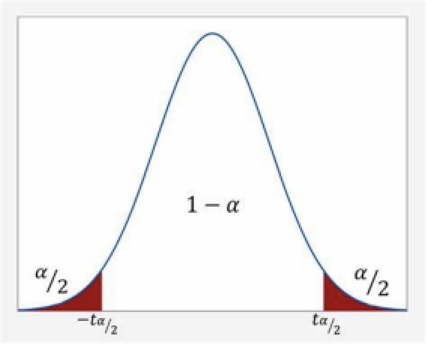


Figure 1.2: The confidence interval for the t-distribution.

Example 1.3.7 *(t-distribution). Suppose that $X \sim N(\mu, \sigma^2)$, where μ and σ^2 are both unknown parameters. Let $X_1, \ldots, X_n$ be a random sample from X. Can we find a coefficient $1-\alpha$ confidence interval for μ?*

Solution. See Fig. 1.2. Note that $\overline{X}$ is an unbiased estimator of μ, and $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. We replace σ in Z by its unbiased estimator S_n^*. Note that

$$\frac{\overline{X}-\mu}{S_n^*/\sqrt{n}} = \frac{\overline{X}-\mu}{S_n/\sqrt{n-1}} \sim t(n-1).$$

Note that the pdf of t distribution is an even function. We have

$$\begin{aligned} P\left(\left|\frac{\overline{X}-\mu}{S_n^*/\sqrt{n}}\right| \geq t_{\alpha/2}(n-1)\right) &= \alpha, \\ P\left(\left|\frac{\overline{X}-\mu}{S_n^*/\sqrt{n}}\right| < t_{\alpha/2}(n-1)\right) &= 1-\alpha. \end{aligned}$$

Calculation reveals that a $1-\alpha$ confidence interval is

$$\left(\overline{X} - t_{\alpha/2}(n-1)\frac{S_n^*}{\sqrt{n}}, \overline{X} + t_{\alpha/2}(n-1)\frac{S_n^*}{\sqrt{n}}\right).$$

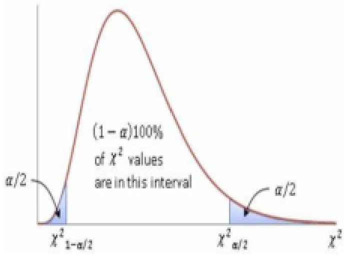


Figure 1.3: The confidence interval for the χ^2 distribution.

Example 1.3.8 *Suppose that $X \sim N(\mu, \sigma^2)$, where μ and σ^2 are both unknown parameters. Let $X_1, \ldots, X_n$ be a random sample from X. Can we find a coefficient $1-\alpha$ confidence interval for σ^2?*

Solution. See Fig. 1.3. We consider the following:

$$\frac{(n-1)S_n^{*2}}{\sigma^2} = \frac{nS_n^2}{\sigma^2} \sim \chi^2(n-1).$$

We would like

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

So a confidence interval for σ^2 is

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

Remark 1.3.9 *In above example, if μ is a known constant, we can use $\chi^2 = \frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2}$, which has distribution $\chi^2(n)$.*

Example 1.3.10 *We now estimate the $1-\alpha$ joint confidence sets of μ and σ^2 for the normal distribution. Since $\overline{X}$ and S_n^{*2} (or S_n^2) are independent, we construct two independent pivotal quantities containing only two unknowns μ and σ^2,*

$$U = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \text{ and } \chi^2 = \frac{(n-1)S_n^{*2}}{\sigma^2},$$

which have distributions of $N(0,1)$ and $\chi^2(n-1)$, respectively. We woule like to find a, c_1, c_2 in the following,

$$P\left(-a < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < a, c_1 < \frac{(n-1)S_n^{*2}}{\sigma^2} < c_2\right) = 0.95.$$

Due to the independence, we only need to solve

$$P\left(-a < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < a\right) P\left(c_1 < \frac{(n-1)S_n^{*2}}{\sigma^2} < c_2\right) = 0.95.$$

This is equivalent to

$$P\left(-a < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < a\right) = \beta_1, \quad P\left(c_1 < \frac{(n-1)S_n^{*2}}{\sigma^2} < c_2\right) = \beta_2,$$

where $\beta_1\beta_2 = 0.95$. There are infinitely many β_1 and β_2 satisfying the relation. For convenience, ignoring the best approximation, we just take $\beta_1 = \beta_2 = 0.975$. Hence, with $\alpha = 0.025$,

$$P\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 0.975,$$
$$P\left(\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S_n^{*2}}{\sigma^2} < \chi^2_{\alpha/2}(n-1)\right) = 0.975,$$

*Notice that the confidence set is **not a rectangle**, which is shown by the shadow region in Fig. 1.4. We can eventually obtain the result,*

$$P\left(\left(\overline{X}-\mu\right)^2 < \frac{\sigma^2 z^2_{\alpha/2}}{n}, \frac{(n-1)S_n^{*2}}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S_n^{*2}}{\chi^2_{1-\alpha/2}(n-1)}\right) = 0.95.$$

*Note that the confidence set is **not unique determined** in this example based on the choice of β_1 and β_2. For symmetric pdf, one can find the shortest confidence interval centered at the mean value. However, for nonsymmetric pdf, it is not easy to find the shortest confidence interval.*

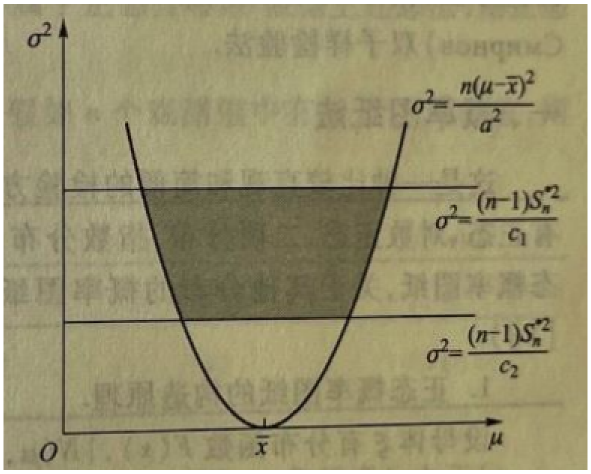

σ^2
$\sigma^2=\frac{n(\mu-\bar{x})^2}{a^2}$
$\sigma^2=\frac{(n-1)S_n^{*2}}{c_1}$
$\sigma^2=\frac{(n-1)S_n^{*2}}{c_2}$
O
$\bar{x}$
μ

Figure 1.4: Shown by the shadow is the joint confidence set for the parameter (μ, σ^2).

1.3.2 Different Ways to Construct Confidence Intervals

Let $X_1, ..., X_n \sim \text{Bernoulli}(p)$. This has wide applications in qualification rate of products, passage rate in exams, market satisfaction, etc.

Example 1.3.11 *In the coin flipping setting, let $X_1, ..., X_n \sim \text{Bernoulli}(p)$. Let $C_n = (\widehat{p}_n - \epsilon_n, \widehat{p}_n + \epsilon_n)$. We saw that Chebyshev's inequality yielded*

$$P(|\overline{X}_n - p| > \epsilon_n) \leq \frac{E|\overline{X}_n - p|^2}{\epsilon_n^2} = \frac{pq/n}{\epsilon_n^2} = \alpha.$$

Take $p = 0.5, n = 100$ and $\alpha = 0.05$. Then

$$\epsilon_n = \sqrt{\frac{pq}{n\alpha}} = 0.2236.$$

Example 1.3.12 *In the coin flipping setting, let $C_n = (\widehat{p}_n - \epsilon_n, \widehat{p}_n + \epsilon_n)$. Using Hoeffding's inequality (see details in the following Appendix)*

$$P(|\overline{X}_n - p| > \epsilon_n) \leq 2e^{-2n\epsilon_n^2},$$

it follows that $\epsilon_n^2 = \frac{\log(2/\alpha)}{2n}$ and

$$P(p \in C_n) \geq 1 - \alpha,$$

for every p. Hence, C_n is a $1-\alpha$ confidence interval. Take $\alpha = 0.05$ and $n = 100$, then $\epsilon_n = \sqrt{\frac{\log(2/\alpha)}{2n}} = 0.1358$. (More comments here. Hoeffding's inequality gives us a simple way to create a confidence interval for a binomial parameter p. Fix $\alpha > 0$ and let

$$\epsilon_n = \sqrt{\frac{\log(2/\alpha)}{2n}}.$$

By Hoeffding's inequality,

$$P(|\overline{X}_n - p| > \epsilon_n) \leq 2e^{-2n\epsilon_n^2} = \alpha.$$

Let $C_n = (\widehat{p}_n - \epsilon_n, \widehat{p}_n + \epsilon_n)$ where $\widehat{p}_n = \overline{X}_n$.Then, $P(p \notin C_n) = P(|\overline{X}_n - p| > \epsilon_n) \leq \alpha$. Hence, $P(p \in C_n) \geq 1 - \alpha$, that is, the random interval C_n traps the true parameter value p with probability $1-\alpha$; we call C_n a $1-\alpha$ confidence interval.)

As mentioned earlier, point estimators often have a limiting normal distribution based on Central Limit Theorem, meaning that equation (1.2) holds, that is, $\widehat{\theta}_n \approx N(\theta, \widehat{\sigma}^2)$. In this case we can construct (approximate) confidence intervals as follows.

Theorem 1.3.13 *(**Normal-based Confidence Interval**). Suppose that $\widehat{\theta}_n \approx N(\theta, \widehat{\sigma}^2)$. Let Φ be the CDF of a standard Normal and let $z_{\alpha/2} = \Phi^{-1}(1-(\alpha/2))$, that is, $P(Z > z_{\alpha/2}) = \alpha/2$ and $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$ where $Z \sim N(0,1)$. Let*

$$C_n = (\widehat{\theta}_n - z_{\alpha/2}\widehat{\sigma}, \widehat{\theta}_n + z_{\alpha/2}\widehat{\sigma}).$$

Then

$$P(\theta \in C_n) \to 1-\alpha,$$

converges in distribution.

Proof. Let $Z_n = (\widehat{\theta}_n - \theta)/\widehat{\sigma}$. By assumption $Z_n \xrightarrow{d} Z$ where $Z \sim N(0,1)$. Hence

$$\begin{aligned} P(\theta &\in C_n) = P(\widehat{\theta}_n - z_{\alpha/2}\widehat{\sigma} < \theta < \widehat{\theta}_n + z_{\alpha/2}\widehat{\sigma}) \\ &= P(-z_{\alpha/2} < \frac{\widehat{\theta}_n - \theta}{\widehat{\sigma}} < z_{\alpha/2}) \\ &\to P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha. \end{aligned}$$

■

Remark 1.3.14 *For 95 percent confidence intervals, $\alpha = 0.05$ and $z_{\alpha/2} = 1.96 \approx 2$ leading to the approximate 95 percent confidence interval $\widehat{\theta}_n \pm 2\widehat{\sigma}$.*

Remark 1.3.15 *Our definition of confidence interval requires that $P(\theta \in C_n) \geq 1-\alpha$ for all $\theta \in \Theta$. A **pointwise asymptotic** confidence interval requires that*

$$\lim_{n\to\infty} \inf P(\theta \in C_n) \geq 1-\alpha,$$

for all $\theta \in \Theta$. A uniform asymptotic confidence interval requires that

$$\lim_{n\to\infty} \inf \inf_{\theta\in\Theta} P(\theta \in C_n) \geq 1-\alpha.$$

The approximate Normal-based interval is a pointwise asymptotic confidence interval.

Example 1.3.16 *Let $X_1, ..., X_n \sim \text{Bernoulli}(p)$ with unknown p and let $\widehat{p}_n = n^{-1}\sum_i X_i$. Then $E(\widehat{p}_n) = n^{-1}\sum_i E(X_i) = p$ so $\widehat{p}_n$ is unbiased. The standard error is $\sigma = \sqrt{V(\widehat{p}_n)} = \sqrt{p(1-p)/n}$. The estimated standard error is $\widehat{\sigma} = \sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}$. By the Central Limit Theorem, $\widehat{p}_n \approx N(p, \widehat{\sigma}^2)$. Therefore, an approximate $1-\alpha$ confidence interval is*

$$\widehat{p}_n \pm z_{\alpha/2}\widehat{\sigma} = \widehat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}.$$

*Take $\alpha = 0.05$, $\widehat{p}_n \approx 0.5$, and $n = 100$, then $z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}} \approx 0.1$. Compare this with the confidence interval in Example 1.3.11 and Example 1.3.12. **The Normal-based interval is shorter but it only has approximately (large sample) correct coverage.** When $n \geq 30$, the application of normal-based Confidence Interval is pretty good.*

Example 1.3.17 *Here is another way to compute based on the central limit theorem. We have*

$$\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1).$$

This reveals that

$$P\left(z_{\alpha/2} \leq \sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \leq z_{1-\alpha/2}\right) \approx 1-\alpha,$$

from which we can solve the approximate $1-\alpha$ *confidence interval,*

$$\left[\frac{2n\overline{X}_n + z^2 - \sqrt{z^4 + 4n\overline{X}_n z^2 - 4n\overline{X}_n^2 z^2}}{2(n+z^2)}, \frac{2n\overline{X}_n + z^2 + \sqrt{z^4 + 4n\overline{X}_n z^2 - 4n\overline{X}_n^2 z^2}}{2(n+z^2)}\right], \tag{1.3}$$

where $z = z_{1-\alpha/2}$. *Take* $\alpha = 0.05$, $n = 100$, $\overline{X}_n = \widehat{p}_n \approx 0.5$, *we obtain the confidence interval*

$$[0.4038, 0.5962].$$

The interval (1.3) is called the Wilson interval approximation, which can be applied to the cases that the number of samples n *is greater than* 30.

Example 1.3.18 *Agresti and Coull (1998) proposed another interval approximation for* p,

$$\left[\widetilde{p} - z_{\alpha/2}\sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{\widetilde{n}}}, \widetilde{p} + z_{\alpha/2}\sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{\widetilde{n}}}\right], \tag{1.4}$$

where $\widetilde{n} = n + z_{\alpha/2}^2$ *and* $\widetilde{p} = \frac{1}{\widetilde{n}}\left(\sum_{i=1}^n X_i + \frac{1}{2}z_{\alpha/2}^2\right)$. *Take* $\alpha = 0.05$, $n = 100$, $\overline{X}_n = \widehat{p}_n \approx 0.5$, *we obtain*

$$[0.4038, 0.5962].$$

This interval is called Agresti and Coull interval approximation. Numerical results show that the coverage of (1.4) is a little bit larger than that of (1.3). When $n \geq 30$, *both (1.3) and (1.4) are suggested. When* $n < 30$, *both can be used for interval approximations as a reference.*

1.3.3 Appendix Probability Inequalities

Hoeffding's inequality is similar in spirit to Markov's inequality and Chebyshev's inequality but it is a **sharper** inequality. We present the result here in two parts.

Theorem 1.3.19 *(Hoeffding's Inequality). Let* $Y_1, ..., Y_n$ *be independent observations such that* $E(Y_i) = 0$ *and* $a_i \leq Y_i \leq b_i$ *(The boundedness is important). Let* $\epsilon > 0$. *Then, for any* $t > 0$,

$$P\left(\sum_{i=1}^n Y_i \geq \epsilon\right) \leq e^{-t\epsilon}\prod_{i=1}^n e^{t^2(b_i-a_i)^2/8}. \tag{1.5}$$

Devroye et al. (1996) is a good reference on probability inequalities and their use in statistics and pattern recognition. The following proof of Hoeffding's inequality is from that text.

Proof. Proof of Hoeffding's Inequality. We will make use of the exact form of Taylor's theorem: if g is a smooth function, then there is a number $\xi \in (0, u)$ such that $g(u) = g(0) + ug'(0) + \frac{u^2}{2}g''(\xi)$.

Proof of Theorem 1.3.19. For any $t > 0$, we have, from Markov's inequality, that

$$\begin{aligned} P\left(\sum_{i=1}^n Y_i \geq \epsilon\right) &= P\left(t\sum_{i=1}^n Y_i \geq t\epsilon\right) = P\left(e^{t\sum_{i=1}^n Y_i} \geq e^{t\epsilon}\right) \\ &\leq e^{-t\epsilon} E\left(e^{t\sum_{i=1}^n Y_i}\right) = e^{-t\epsilon}\prod_{i=1}^n E(e^{tY_i}). \end{aligned} \tag{1.6}$$

Since $a_i \leq Y_i \leq b_i$, we can write Y_i as a convex combination of a_i and b_i, namely, $Y_i = \alpha b_i + (1-\alpha)a_i$ where $\alpha = (Y_i - a_i)/(b_i - a_i)$. So, by the convexity of e^{ty} we have

$$e^{tY_i} \leq \frac{Y_i - a_i}{b_i - a_i} e^{tb_i} + \frac{b_i - Y_i}{b_i - a_i} e^{ta_i}.$$

Take expectations of both sides and use the fact that $E(Y_i) = 0$ to get

$$Ee^{tY_i} \leq -\frac{a_i}{b_i - a_i} e^{tb_i} + \frac{b_i}{b_i - a_i} e^{ta_i} = e^{g(u)},$$

where $u = t(b_i - a_i)$, $g(u) = -\gamma u + log(1 - \gamma + \gamma e^u)$ and $\gamma = -a_i/(b_i - a_i)$.

Note that $g(0) = g'(0) = 0$. Also, $g''(u) \leq 1/4$ for all $u > 0$. By Taylor's theorem, there is a $\xi \in (0, u)$ such that

$$\begin{aligned} g(u) &= g(0) + ug'(0) + \frac{u^2}{2} g''(\xi) \\ &= \frac{u^2}{2} g''(\xi) \leq \frac{u^2}{8} = \frac{t^2(b_i - a_i)^2}{8}. \end{aligned}$$

Hence,

$$Ee^{tY_i} \leq e^{g(u)} \leq e^{t^2(b_i - a_i)^2/8}.$$

The result follows from (1.6). ■

Theorem 1.3.20 *Let* $X_1, ..., X_n \sim$ Bernoulli(p). *Then, for any* $\epsilon > 0$,

$$P(|\overline{X}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2},$$

where $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$.

Proof. Proof of Theorem 1.3.20. Let $Y_i = (1/n)(X_i - p)$. Then $E(Y_i) = 0$ and $a \leq Y_i \leq b$ where $a = -p/n$ and $b = (1-p)/n$. Also, $(b-a)^2 = 1/n^2$. Applying Theorem 1.3.19 we get

$$P(\overline{X}_n - p > \epsilon) = P\left(\sum_{i=1}^n Y_i \geq \epsilon\right) \leq e^{-t\epsilon} e^{t^2/(8n)}.$$

The above holds for any $t > 0$. In particular, take $t = 4n\epsilon$ and we get $P(\overline{X}_n - p > \epsilon) \leq e^{-2n\epsilon^2}$. By a similar argument we can show that $P(\overline{X}_n - p < -\epsilon) \leq e^{-2n\epsilon^2}$. Putting these together we get

$$P(|\overline{X}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

■

The following inequality is useful for bounding probability statements about Normal random variables.

Theorem 1.3.21 *(Mill's Inequality). Let $Z \sim N(0,1)$. Then,*

$$P(|Z| > t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}.$$

Proof. Intuitively, this kind of tail bound is useful because we can get exponentially-fast decay without calculating the distribution function directly. The broad strokes of the proof follow Aliyah Ahmed's response to a post on StackExchange. We begin by observing that density of Z is symmetric about the origin, therefore:

$$P\{|Z| > t\} = 2P\{Z > t\}$$

We then observe that by playing with distribution functions and expectations, we get the following upper bound:

$$\begin{aligned} t \cdot P\{Z &> t\} = t \int_t^\infty dF(x) \\ &\leq \int_t^\infty xdF(x) \\ &= \int_t^\infty x \cdot \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\} \\ &= \frac{1}{\sqrt{2\pi}} \exp\{-\frac{t^2}{2}\}. \end{aligned}$$

In the process using sneaky way to introduce a quantity that has a nice, clean closed-form integral. Closer examination shows that this is in fact a tighter version of Markov's Inequality; rather than taking EX, we take $E[X \cdot \mathbf{1}\{X > t\}]$. This implies that:

$$\begin{aligned} P\{Z &> t\} \leq \frac{1}{\sqrt{2\pi}} \frac{1}{t} \exp\{-\frac{t^2}{2}\}, \\ P\{|Z| &> t\} \leq \sqrt{\frac{2}{\pi}} \frac{1}{t} \exp\{-\frac{t^2}{2}\}. \end{aligned}$$

■

Remark 1.3.22 *Let $Z \sim N(0, \sigma^2)$. Then,*

$$P(|Z| > t) \leq \sqrt{\frac{2}{\pi}} \frac{\sigma}{t} e^{-t^2/(2\sigma^2)}.$$

Remark 1.3.23 *This result can be extended to the maximum of m Gaussian random variables by way of the union bound. Suppose $\{Z_i\}_{i=1}^m \sim N(0, \sigma^2)$. Then the union bound implies:*

$$P(\max_{1 \leq i \leq m} |Z_i| > t) \leq m \cdot \sqrt{\frac{2}{\pi}} \frac{\sigma}{t} e^{-t^2/(2\sigma^2)}.$$

Example 1.3.24 *Let $X_1, ..., X_n \sim \text{Bernoulli}(p)$. Let $p = 0.5, n = 100$ and $\epsilon = 0.2$. We saw that Chebyshev's inequality yielded*

$$P(|\overline{X}_n - p| > \epsilon) \leq \frac{E|\overline{X}_n - p|^2}{\epsilon^2} = \frac{pq/n}{\epsilon^2} = 0.0625.$$

According to Hoeffding's inequality,

$$P(|\overline{X}_n - p| > \epsilon) \leq 2e^{-2(100)(0.2)^2} = 0.00067.$$

By the Central Limit Theorem,

$$P(|\overline{X}_n - p| > \epsilon) \le P(\frac{|\overline{X}_n - p|}{\sqrt{pq/n}} > \frac{\epsilon}{\sqrt{pq/n}}) \approx P\left(|Z| > \frac{0.2}{\frac{0.5}{10}} = 4\right) < 6 \times 10^{-5},$$

where $Z \sim N(0,1)$. By the table for Φ, we know that the probability is smaller at least than 0.0004 since 4 *is too large for a standard normal. By Mill's inequality, we can see that the probability is not greater than* 6×10^{-5}*. One can see CLT is most tight bound, however, CLT only provides an estimation of the probability. This CLT bound is valid only when n is really large. In contrast, the other two bounds are reliable for any number of samples n. Moreover, Hoeffding's bound is tighter than Chebyshev's. As I know, when n is large one can use Hoeffding's bound whereas when n is small one can use Chebyshev's bound.*

1.3.4 More Topics about Confidence Interval

unknown variances for both subsamples

If both σ_1^2 and σ_2^2 are unknown and also we do not know whether they are equal, then the problem to estimate the confidence interval of the difference mean $\mu_1 - \mu_2$ is the famous Behren-Fisher problem in statistics. If both σ_1^2 and σ_2^2 were known, we could use the confidence interval,

$$\left[\overline{X} - \overline{Y} - z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \overline{X} - \overline{Y} + z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right].$$

Now since σ_1^2 and σ_2^2 are unknown, we can replace them with unbiased $S_{n,1}^{*2}$ and $S_{n,2}^{*2}$ and also replace $z_{1-\alpha/2}$ with $t_{1-\alpha/2}(l)$ (with l degrees of freedom) to obtain

$$\left[\overline{X} - \overline{Y} - t_{1-\alpha/2}(l)\sqrt{\frac{S_{n,1}^{*2}}{n} + \frac{S_{n,2}^{*2}}{m}}, \overline{X} - \overline{Y} + t_{1-\alpha/2}(l)\sqrt{\frac{S_{n,1}^{*2}}{n} + \frac{S_{n,2}^{*2}}{m}}\right],$$

where l is the closest integer to the following l^*,

$$l^* = \frac{\left(\frac{S_{n,1}^{*2}}{n} + \frac{S_{n,2}^{*2}}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_{n,1}^{*2}}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_{n,2}^{*2}}{m}\right)^2}.$$

Here are two issues to construct the pivotal quantity. First, the sum of χ^2 distributions with different non-integer variances may not give rise to a χ^2 distribution since the exponential indices may be different. Second, it is hard to cancel out both σ_1^2 and σ_2^2 in constructing the pivotal quantity.

单侧置信限

one-sided confidence interval

So far we only discuss the two-sided confidence interval. In practice, we only interested in the unknown parameter not smaller than or not greater than some value. For instance, we hope the lifespan of some products as long as possible, the standard deviation of the size of some products as small as possible, etc. We now need the concept for one-sided confidence interval.

Definition 1.3.25 *Let θ be a parameter and let $a = a(X_1, ..., X_n)$ and $b = b(X_1, ..., X_n)$ be two functions or two statistics of the random samples $X_1, \ldots, X_n$ from a distribution. Then a is said to be a $1-\alpha$ lower one-sided confidence interval for θ if*

$$P(\theta \geq a) \geq 1-\alpha, \quad \textit{for all } \theta \in \Theta.$$

In words, (a, ∞) traps θ with probability $1-\alpha$. On the other hand, b is said to be a $1-\alpha$ upper one-sided confidence interval for θ if

$$P(\theta \leq b) \geq 1-\alpha, \quad \textit{for all } \theta \in \Theta.$$

In words, $(-\infty, b)$ traps θ with probability $1-\alpha$.

置信水平为$1-\alpha$的单侧置信下限（上限）

Example 1.3.26 *Suppose that $X \sim N(\mu, \sigma^2)$, where μ and σ^2 are both unknown parameters. Let $X_1, \ldots, X_n$ be a random sample from X. Can we find a coefficient $1-\alpha$ lower one-sided confidence interval for μ?*
Solution. Note that $\overline{X}$ is an unbiased estimator of μ, and $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. We replace σ in Z by its unbiased estimator S_n^*. Note that

$$\frac{\overline{X}-\mu}{S_n^*/\sqrt{n}} = \frac{\overline{X}-\mu}{S_n/\sqrt{n-1}} \sim t(n-1).$$

Then we have

$$P\left(\frac{\overline{X}-\mu}{S_n^*/\sqrt{n}} < t_\alpha(n-1)\right) = 1-\alpha.$$

Calculation reveals that a $1-\alpha$ lower one-sided confidence interval is

$$[\overline{X} - t_\alpha(n-1)\frac{S_n^*}{\sqrt{n}}, +\infty).$$

relationship between confidence interval and hypothesis test for standard problems

For the problems of estimating means, variances, and differences of means, ratios of variances for normal distributions, the construction of pivotal quantities for confidence intervals is similar to the application of test statistics in hypothesis tests. In the following section, we will discuss more about this topic.

This is not hard to understand since there are close relations between confidence interval approximaitons and hypothesis tests. Let us consider the example for $X \sim N(\mu, \sigma^2)$, where μ is unknown and σ^2 is known. Then the statement that μ_0 belongs to the $1-\alpha$ confidence interval of μ is equivalent to the statement that we cannot reject the null hypothesis at the level α for the hypothesis test problem H_0:$\mu = \mu_0$, H_1:$\mu \neq \mu_0$. The reason is given as follows. The $1-\alpha$ confidence interval is

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

If this set constains μ_0, then we have $\left|\overline{X} - \mu_0\right| \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. This means that $\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right| \leq z_{\alpha/2}$. Thus at the significance level α, the subsampling μ_0 is NOT located in the rejection region of the test. Therefore, for the hypothesis test problem, we cannot reject the null hypothesis at the significance level of α.

1.4 Hypothesis Testing

1.4.1 Definition

Keywords: Parametric Hypothesis Test（参数假设检验）， significance level（显著性水平）， p-value（p值）

In general, let us consider a statistical problem involving a population X whose distribution has a unknown parameter θ. The value of θ is unknown but must lie in a certain parameter space Θ. Suppose now that Θ can be partitioned into two disjoint subsets Θ_0 and Θ_1, i.e.,

$$\Theta = \Theta_0 \cup \Theta_1, \quad \Theta_0 \cap \Theta_1 = \varnothing.$$

And the statistician is interested in whether θ lies in Θ_0 or in Θ_1. Denote by H_i the hypothesis that $\theta \in \Theta_i$ $(i = 0, 1)$. The statistician must decide which of the hypotheses H_0 or H_1 appears to be true. A problem of this type is called a problem of testing hypotheses.

零假设，备择假设

In **hypothesis testing**, we start with some default theory — called a **null hypothesis** H_0 — and we ask if the data provide sufficient evidence to reject the theory. If not we retain the null hypothesis. The term "retaining the null hypothesis" is due to Chris Genovese. Other terminology is "accepting the null" or "failing to reject the null." In other words, for **hypothesis testing**, we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1.$$

Then H_0 is the **null hypothesis** and H_1 is the **alternative hypothesis**. if we decide that θ lies in Θ_1, then we reject the null hypothesis H_0. If we decide that θ lies in Θ_0, then we do not reject H_0.

Example 1.4.1 *(Testing if a Coin is Fair). Let*

$$X_1, ..., X_n \sim \text{Bernoulli}(p)$$

*be n independent coin flips. Suppose we want to test if the coin is fair. Let H_0 denote the hypothesis that the coin is fair and let H_1 denote the hypothesis that the coin is not fair. H_0 is called the **null hypothesis** and H_1 is called the **alternative hypothesis**. We can write the hypotheses as*

$$H_0 : p = 1/2 \quad \textit{versus} \quad H_1 : p \neq 1/2.$$

It seems reasonable to reject H_0 if $T = |\widehat{p}_n - (1/2)|$ is large. When we discuss hypothesis testing in detail, we will be more precise about how large T should be to reject H_0.

Remark 1.4.2 *Statistical inference is covered in many texts. Elementary texts include DeGroot and Schervish (2002) and Larsen and Marx (1986). At the intermediate level Larry recommends Casella and Berger (2002), Bickel and Doksum (2000), and Rice (1995). At the advanced level, Cox and Hinkley (2000), Lehmann and Casella (1998), Lehmann (1986), and van der Vaart (1998).*

简单假设，复合假设

Definition 1.4.3 *If Θ_i $(i = 0, 1)$ contains just a single value of θ, then H_i is called a simple hypothesis. If the set Θ_i $(i = 0, 1)$ contains more than one value of θ, then H_i is called a composite hypothesis.*

单边假设，双边假设

Definition 1.4.4 *Moreover, one-sided null hypotheses are of the form $H_0 : \theta \leq \theta_0$ or $H_0 : \theta \geq \theta_0$. The corresponding one-sided alternative hypotheses being $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$. When the null hypothesis is simple, the alternative hypothesis is usually two-sided.*

当问题需要确定答案时才开始假设检验

Remark 1.4.5 *Warning! There is a tendency to use hypothesis testing methods even when they are not appropriate. Often, estimation and confidence intervals are better tools. Use hypothesis testing only when you want to test a well-defined hypothesis.*

检验统计量，拒绝域

Definition 1.4.6 *Let $X_1, \ldots, X_n$ be a random sample from the population X. Let $Z = Z(X_1, ..., X_n)$ be a statistic and R be a subset of $\mathbb{R}$. Suppose that we will reject H_0 if $Z \in R$. Then we call Z a* ***test statistic*** *and R the rejection region of the test. If $Z \in R$ we reject the null hypothesis, otherwise, we do not reject the null hypothesis:*

$$\begin{aligned} Z &\in R \Rightarrow \textit{reject } H_0 \\ Z &\notin R \Rightarrow \textit{retain (do not reject) } H_0. \end{aligned}$$

Remark 1.4.7 *Here, we use "rejection" instead of "acception". Because we* ***can not use data to prove something****. But it is reasonable to* ***use data to disproof something****.*

第一类错误（拒真），第二类错误（受伪），显著性水平

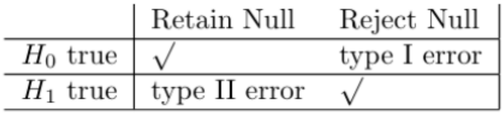
No matter how we do the test, mistakes can not always be avoided. An erroneous decision to reject a true null hypothesis is called a **type I error** (a **false positive conclusion**). An erroneous decision not to reject a false null hypothesis is called a **type II error** (a **false negative conclusion**).

这里我们只对犯第一类错误加以限制，只考虑显著性检验（significance test）

The objective of a statistical test of H_0 is not to explicitly determine whether or not H_0 is true but rather to determine if its validity is consistent with the resultant data. Hence, with this objective it seems reasonable that H_0 should only be rejected if the resultant data are very unlikely. The classical way of accomplishing this is to fix a level of significance α and then require that the test have the property that the probability of a type I error occurring can never be greater than α, i.e.,

$$P(Z \in R) \leq \alpha.$$

(**Type II error is not considered in this note**.)



	Retain Null	Reject Null
H_0 true	√	type I error
H_1 true	type II error	√

Figure 1.5: Summary of outcomes of hypothesis testing.

The procedure of testing hypothesis is as follows:

(1) Construct the hypotheses H_0 and H_1 from the problem. (We propose the null hypothesis H_0.) 提出假设

(2) Choose a suitable test statistic such that its sampling distribution **does not contain unknown parameters**. 构造小概率事件，创建拒绝域

(3) Given level of significant α, find the corresponding rejection region (based on H_1?).

(4) Calculate the value of the statistic from observed sample values, determine whether it is in the reject region (reject H_0) or not (not to reject H_0).

1.4.2 Examples for Simple Hypothesis

Example 1.4.8 *(Z Test) Let us consider the population satisfying $X \sim N(\mu, \sigma^2)$, where σ^2 is known but μ is unknown. To carry out a test of the following hypothesis at the significance level of significant α:*

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0,$$

we have chosen the test statistic $Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$. If H_0 is true, then

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}},$$

has standard normal distribution. Here μ_0 comes from the hypothesis and σ^2 is known, so Z does not contain unknown parameters. Let us choose the significant level to be α. Recall that the pdf is even. Then we have that

$$\alpha = P(|Z| > z_{\alpha/2}) = P(|Z| \in R).$$

So a rejection region of significance level α is $(-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, +\infty)$. $\overline{X}$与μ_0误差很大时，有理由怀疑原假设的正确性，构造小概率事件，创建拒绝域

In this example, if we take $\alpha = 0.05$, then $z_{0.025} = 1.96$. And $n = 9, \sigma = 0.015, \overline{x} = 0.511$. It follows that

$$|z| = \left| \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \right| = 2.2 > 1.96.$$

Then we reject H_0.

Remark 1.4.9 *Consider the test: $H_0 : \mu \neq \mu_0, \quad H_1 : \mu = \mu_0$. Assume that H_0 is true. If we use $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$, then $Z \sim N(0, 1)$. But this random variable involves μ, which is unknown and can not appear in the expression of R. One may consider $Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$ instead. It does not contain unknown parameters. But, its distribution is unknown since its mean involving the unknown μ. We are not able to handle this test in a similar way as in above example. So, people need to construct suitable hypothesis H_0 and H_1 for a concrete problem.*

Example 1.4.10 *(T Test) Suppose that the two independent populations X and Y satisfy that $X \sim N(\mu_1, \sigma^2)$ and $Y \sim N(\mu_2, \sigma^2)$, where σ is unknown (we require that they have same variance). Let us carry out a test of the following hypothesis at level α:*

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2,$$

Let $X_1, ..., X_n$ be a random sample from X, and $Y_1, ..., Y_m$ be a random sample from Y. Take

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n-1)S_{n,1}^{*2} + (m-1)S_{n,2}^{*2}}{m+n-2}(\frac{1}{n} + \frac{1}{m})}}.$$

Here $S_{n,1}^{*2}$ and $S_{n,2}^{*2}$ are the unbiased sample variances of X and Y , respectively. If H_0 is true, then $T \sim t(n+m-2)$. Similar as previous, we have

$$P(|T| \geq t_{\alpha/2}(n+m-2)) = \alpha.$$

The rejection region for T is

$$(-\infty, -t_{\alpha/2}(n+m-2)] \cup [t_{\alpha/2}(n+m-2), +\infty).$$

Example 1.4.11 *(χ^2 Test) Suppose that the population satisfies $X \sim N(\mu, \sigma^2)$, where neither μ nor σ^2 is known. Let us carry out a test of the following hypothesis at level α:*

$$H_0 : \sigma^2 = \sigma_0^2, \quad H_1 : \sigma^2 \neq \sigma_0^2,$$

Let $X_1, ..., X_n$ be a random sample from X. In this example, we should consider the test statistic

$$\chi^2 = \frac{(n-1)S_n^{*2}}{\sigma_0^2}$$

When H_0 is true, we have $\chi^2 \sim \chi^2(n-1)$. Assume that the rejection region has the form

$$P(\{\chi^2 \leq c_1\} \cup \{\chi^2 \geq c_2\}) = \alpha.$$

For convenience, let

$$P(\chi^2 \leq c_1) = P(\chi^2 \geq c_2) = \frac{\alpha}{2}.$$

Then $c_1 = \chi^2_{1-\alpha/2}(n-1), c_2 = \chi^2_{\alpha/2}(n-1)$. The rejection region is

$$(-\infty, \chi^2_{1-\alpha/2}(n-1)] \cup [\chi^2_{\alpha/2}(n-1), +\infty).$$

Remark 1.4.12 *If, in above example, the mean μ is known, then one can use the test statistic $\chi^2 = \frac{\sum(X_i-\mu)^2}{\sigma_0^2}$, which has distribution $\chi^2(n)$. The rejection region can be $(-\infty, \chi^2_{1-\alpha/2}(n)] \cup [\chi^2_{\alpha/2}(n), +\infty)$.*

Example 1.4.13 *(F Test) Let X and Y be two independent population such that $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, where μ_1 and μ_2 are unknown parameters. Let us carry out a test of the following hypothesis at level α:*

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad H_1 : \sigma_1^2 \neq \sigma_2^2,$$

Let $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_m$ be random samples from X and Y, with unbiased sample variance S_X^{*2} and S_Y^{*2}, respectively. If the hypothesis H_0 is true, then the statistic

$$F = \frac{S_X^{*2}}{S_Y^{*2}},$$

has distribution $F(n-1, m-1)$. Similar as previous, we may take use of the facts that

$$P(F \leq F_{1-\alpha/2}(n-1, m-1)) = P(F \geq F_{\alpha/2}(n-1, m-1)) = \frac{\alpha}{2}.$$

So the rejection region for F is

$$(-\infty, F_{1-\alpha/2}(n-1, m-1)] \cup [F_{\alpha/2}(n-1, m-1), +\infty).$$

1.4.3 A Two-step Test Example

Example 1.4.14 *We compare the therapeutic effects of two somnifacients (sleeping pills). We separate 20 patients into two groups with each one 10 persons. The extended sleep time after taking medication is normally distributed. The data are*

$A:$	5.5	4.6	3.4	1.9	1.6	1.1	0.8	0.1	-0.1	4.4
$B:$	3.7	3.4	2.0	2.0	0.8	0.7	0	-0.1	-0.2	-1.6

Then is there any ***significant*** *difference for the therapeutic effects between the two somnifacients at the level* $\alpha = 0.05$*?*

Solution. Suppose that $X \sim N(\mu_1, \sigma_1^2)$ *and* $Y \sim N(\mu_2, \sigma_2^2)$*, and we compute that*

$$\overline{X} = 2.33, \quad \overline{Y} = 0.75, \quad S_X^{*2} = 4.01, \quad S_Y^{*2} = 3.2.$$

We first test the null hypothesis $\sigma_1^2 = \sigma_2^2$ *to write as*

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \textit{versus} \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

The test statistic $F = S_X^{*2}/S_Y^{*2} = 1.25$*. We see that*

$$F_{0.025}(9,9) = 5.35, \quad F_{1-0.025}(9,9) = \frac{1}{F_{0.025}(9,9)} = \frac{1}{5.35} = 0.187.$$

Since $F = 1.25 \in (0.187, 5.35)$ *, we accept the null hypothesis* $\sigma_1^2 = \sigma_2^2$*.*

Under the condition that $\sigma_1^2 = \sigma_2^2$*, we then test the null hypothesis* $\mu_1 = \mu_2$*,*

$$H_0 : \mu_1 = \mu_2 \quad \textit{versus} \quad H_1 : \mu_1 \neq \mu_2.$$

The test statistic is

$$T = \frac{\overline{X} - \overline{Y}}{S_w^* \sqrt{\frac{1}{n} + \frac{1}{m}}}, \quad \textit{with } S_w^* = \sqrt{\frac{(n-1)S_X^{*2} + (m-1)S_Y^{*2}}{m+n-2}}.$$

By computation we see

$$s_w^* = \sqrt{\frac{(9)(4.01) + (9)(3.2)}{18}} = 1.899 \quad \textit{and } t = \frac{2.33 - 0.75}{1.899\sqrt{\frac{1}{10} + \frac{1}{10}}} = 1.86.$$

Since $|t| = 1.86 < 2.101 = t_{0.025}(18)$*, we still accept the null hypothesis* $H_0 : \mu_1 = \mu_2$*. There is NO* ***significant*** *difference for the therapeutic effects between the two somnifacients.*

Remark 1.4.15 *One can also do not test* $H_0 : \sigma_1^2 = \sigma_2^2$ *at first; but directly replace* σ_1^2, σ_2^2 *with* S_X^{*2}, S_Y^{*2} *in a normal distribution instead (see Sec. 1.3.4).*

Remark 1.4.16 *One can see that the test statistic* 1.86 *is very close to the quantile* 2.101 *while still* $|t| = 1.86 < 2.101 = t_{0.025}(18)$*. In the next section, we will introduce the concept for the p-value to further discuss the "quality" of a hypothesis test.*

1.4.4 The Wald Test

The test is named after Abraham Wald (1902–1950), who was a very influential mathematical statistician. Wald died in a plane crash in India in 1950. Let θ be a scalar parameter, let $\widehat{\theta}$ be an estimate of θ and let $\widehat{\sigma}$ be the estimated standard error of $\widehat{\theta}$.

Definition 1.4.17 *(The Wald Test) Consider testing*

$$H_0 : \theta = \theta_0 \quad \textit{versus} \quad H_1 : \theta \neq \theta_0.$$

Assume that $\widehat{\theta}$ is asymptotically normal:

$$\frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}} \xrightarrow{d} N(0,1).$$

The size α Wald test is: reject H_0 when $|W| > z_{\alpha/2}$ where

$$W = \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}}.$$

Theorem 1.4.18 *Asymptotically, the Wald test has size α, that is,*

$$P(|W| > z_{\alpha/2}) \to \alpha,$$

as $n \to \infty$.

Proof. Under $\theta = \theta_0$, we have $\frac{\widehat{\theta}-\theta_0}{\widehat{\sigma}} \xrightarrow{d} N(0,1)$. Hence, the probability of rejecting when the null $\theta = \theta_0$ is true is

$$\begin{aligned} P(|W| > z_{\alpha/2}) &= P\left(\left|\frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}}\right| > z_{\alpha/2}\right) \\ &\to P(|Z| > z_{\alpha/2}) = \alpha, \end{aligned}$$

where $Z \sim N(0,1)$. ■

Remark 1.4.19 *An alternative version of the Wald test statistic is $W = (\widehat{\theta} - \theta_0)/\sigma_0$ where σ_0 is the standard error computed at $\theta = \theta_0$. Both versions of the test are valid.*

Example 1.4.20 *(Comparing Two Prediction Algorithms). We test a prediction algorithm on a test set of size m and we test a second prediction algorithm on a second test set of size n. Let X be the number of incorrect predictions for algorithm 1 and let Y be the number of incorrect predictions for algorithm 2. Then $X \sim$Binomial(m, p_1) and $Y \sim$Binomial(n, p_2). To test the null hypothesis that $p_1 = p_2$ write*

$$H_0 : \delta = 0 \quad \textit{versus} \quad H_1 : \delta \neq 0.$$

where $\delta = p_1 - p_2$. The MLE is $\widehat{\delta} = \widehat{p}_1 - \widehat{p}_2$ with estimated standard error

$$\widehat{\sigma} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}.$$

The size α Wald test is to reject H_0 when $|W| > z_{\alpha/2}$ where

$$W = \frac{\widehat{\delta} - 0}{\widehat{\sigma}} = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}}.$$

The power of this test will be largest when p_1 is far from p_2 and when the sample sizes are large.

Example 1.4.21 *(Comparing Two Means). Let $X_1, ..., X_m$ and $Y_1, ..., Y_n$ be two independent samples from populations with means μ_1 and μ_2, respectively. Let's test the null hypothesis that $\mu_1 = \mu_2$. Write this as $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ where $\delta = \mu_1 - \mu_2$. Recall that the nonparametric plug-in estimate of δ is $\widehat{\delta} = \overline{X} - \overline{Y}$ with estimated standard error*

$$\widehat{\sigma} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}},$$

where s_1^2 and s_2^2 are the sample variances. The size α Wald test rejects H_0 when $|W| > z_{\alpha/2}$ where

$$W = \frac{\widehat{\delta} - 0}{\widehat{\sigma}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}.$$

1.4.5 Examples for Composite Hypothesis

For such types of problems, if the null and alternative hypotheses are already given, the subsequent computational procedure is standard technique. However, at least for me, the difficulty is to design the null and alternative hypotheses.

Example 1.4.22 *The population of all verbal GRE scores are known to have a standard deviation of* 8.5. *A university hopes to receive applicants with a verbal GRE scores over* 210. *This year, the mean verbal GRE scores for the* 42 *applicants was* 212.79. *Conduct a test at the level of significance* 0.05 *to see whether this new mean is significantly greater than the desired mean of* 210.

The test of hypothesis can be as follows:

$$H_0 : \mu \leq 210 \quad \textit{versus} \quad H_1 : \mu > 210.$$

The test statistic is

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

and its value is

$$z = \frac{212.79 - 210}{8.5/\sqrt{42}} = 2.1272.$$

The rejection region constructed based on H_1 is

$$[z_{0.05}, +\infty) = [1.64, +\infty).$$

We see that z is located in rejection region so that we reject the null hypothesis H_0 and states that the verbal GRE scores of the applicants is ***significantly greater than*** 210.

Remark 1.4.23 *In my view, rejection is persuasive so that we'd better put the conclusion we hope to verify in the alternative hypothesis.*

Example 1.4.24 *A manufacture claims that taking a new technique the lifespan of their light bulbs can be extended much over* 1000 *hours. We can take*

$$H_0 : \mu \leq 1000 \quad \textit{versus} \quad H_1 : \mu > 1000. \tag{1.7}$$

For another problem, a manufacture claims that the lifespan of their light bulbs can be not less than 1000 *hours. We can take*

$$H_0 : \mu \geq 1000 \quad \textit{versus} \quad H_1 : \mu < 1000. \tag{1.8}$$

The sample mean is 1005 *for* $n = 16$ *number of random samples. The standard deviation is* 40 *hours. The test statistic is*

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{1005 - 1000}{40/4} = 0.5.$$

The rejection region for problem (1.7) is $[1.64, +\infty)$. *We see that the statistic is not in rejection region so that we accept the null* $\mu \leq 1000$ *and then we claim that the lifespan is not much over* 1000 *hours. For the second problem (1.8) the rejection region is* $(-\infty, -1.64]$. *We see that the statistic is not in rejection region so that we accept the null* $\mu \geq 1000$ *and then we claim that the lifespan is not less than* 1000 *hours. Here is the charm of statistics. One may arrive at completely opposite conclusions for the same problem depending on how you choose the hypothesis, the number of samples, the quality of samples, the significance level, etc. For this problem, their claims indeed are different while they seems no difference. Let's see the following example for the appropriate way to deal with such situation.*

Example 1.4.25 *There are two ways A and B to make the same type of products with standard deviations of tensile strength* 6 *kg and* 8 *kg, respectively. Now* 12 *and* 16 *numbers of products are random selected from A and B with respective sample means* 34 *and* 40 *kg. The question is if the products by A have less tensile strength than the products by B?*
Solution. First, let

$$H_0 : \mu_1 \leq \mu_2 \quad versus \quad H_1 : \mu_1 > \mu_2.$$

The test statistic is

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} = \frac{34 - 40}{\sqrt{\frac{6^2}{12} + \frac{8^2}{16}}} = -2.2678.$$

The rejection region is $[1.64, \infty)$. *Thus we accept* $H_0 : \mu_1 \leq \mu_2$. *Second, let*

$$H_0 : \mu_1 \geq \mu_2 \quad versus \quad H_1 : \mu_1 < \mu_2.$$

The rejection region is $(-\infty, -1.64]$. *Thus we reject* $H_0 : \mu_1 \geq \mu_2$. *Both results agree to conclude that the products by A have less tensile strength than the products by B.*

Remark 1.4.26 *Rejection to the null hypothesis is persuasive whereas acception is not persuisive. In practice, we should keep on hypothesis test until rejection. Sometimes we may have the contradictory situation, which means we accept* $H_0 : \mu_1 \leq \mu_2$ *and we also accept* $H_0 : \mu_1 \geq \mu_2$. *At this time, we should increase the significance level* α *and keep on hypothesis test until we arrive at the consistent conclusion.*

1.4.6 p-Values

Reporting "reject H_0" or "retain H_0" is not very informative. Instead, we could ask, for every α, whether the test rejects at that level. Generally, if the test rejects at level α it will also reject at level $\alpha' > \alpha$. Hence, there is a smallest α at which the test rejects and we call this number the p-value. See Figure 1.6.

Definition 1.4.27 *Suppose that for every* $\alpha \in (0, 1)$ *we have a size* α *test with rejection region* R_α *(the size of* R_α*, in general, is monotonically increasing w.r.t.* α*). Then,*

$$\text{p-value} = \inf \Big\{ \alpha : T(X^n) \in R_\alpha \Big\}.$$

That is, the ***p-value is the smallest level at which we can reject*** H_0.

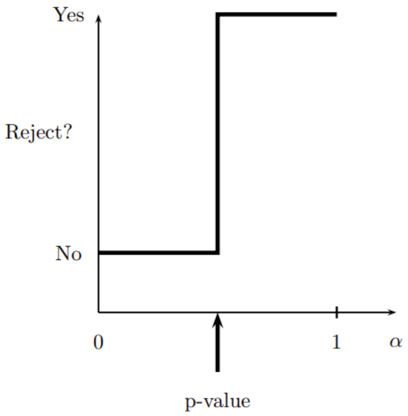
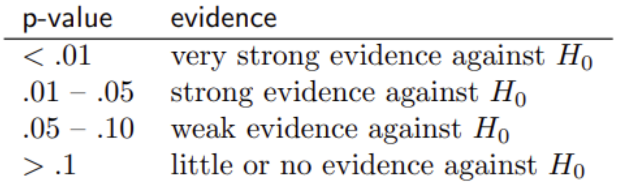


Figure 1.6: p-values explained. For each α we can ask: does our test reject H_0 at level α? The p-value is the smallest α at which we do reject H_0. If the evidence against H_0 is strong, the p-value will be small.

Informally, the p-value is a measure of the evidence against H_0: the smaller the p-value, the stronger the evidence against H_0. Typically, researchers use the evidence scale as shown in Figure 1.7.



p-value	evidence
$< .01$	very strong evidence against H_0
$.01 - .05$	strong evidence against H_0
$.05 - .10$	weak evidence against H_0
$> .1$	little or no evidence against H_0

Figure 1.7: Table for p-values versus evidence.

Remark 1.4.28 *Warning!* ***A large p-value is not strong evidence in favor of*** H_0. *A large p-value can occur for two reasons: (i)* H_0 *is true or (ii)* H_0 *is false but the test has low power.*

Remark 1.4.29 *Warning! Do not confuse the p-value with* $P(H_0|Data)$. *The p-value is not the probability that the null hypothesis is true.*

Theorem 1.4.30 *Let* $w = (\widehat{\theta} - \theta_0)/\widehat{\sigma}$ *denote the observed value of the Wald statistic* W. *The p-value is given by*

$$\text{p-value} = P(|W| > |w|) \approx P(|Z| > |w|) = 2\Phi(-|w|),$$

where $Z \sim N(0, 1)$.

To understand this last theorem, look at Figure 1.8.

Here is an important property of p-values.

Theorem 1.4.31 *If the test statistic has a continuous distribution, then under* $H_0 : \theta = \theta_0$, *the p-value has a Uniform* $(0, 1)$ *distribution. Therefore, if we reject* H_0 *when the p-value is less than* α, *the probability of a type I error is* α.

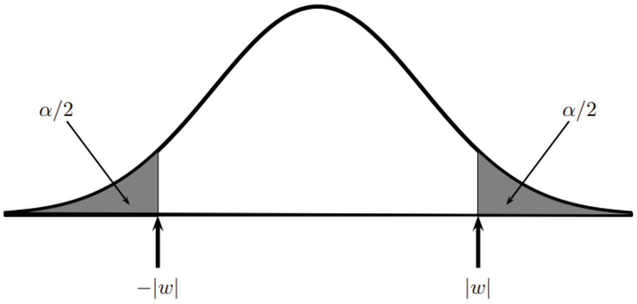


Figure 1.8: The p-value is the smallest α at which you would reject H_0. To find the p-value for the Wald test, we find α such that $|w|$ and $-|w|$ are just at the boundary of the rejection region. Here, w is the observed value of the Wald statistic: $w = (\widehat{\theta} - \theta_0)/\widehat{\sigma}$. This implies that the p-value is the tail area $P(|Z| > |w|)$ where $Z \sim N(0,1)$.

Remark 1.4.32 *In other words, if H_0 is true, the p-value is like a random draw from a Unif$(0,1)$ distribution. If H_1 is true, the distribution of the p-value will tend to concentrate closer to* 0.

Example 1.4.33 *Two groups of cholesterol data with respective means of* 216.19 *and* 195.27. *Each of the groups has* 16 *persons. The estimated standard deviations are* 20.0 *and* 9.6, *respectively. We ask if the means are different.*

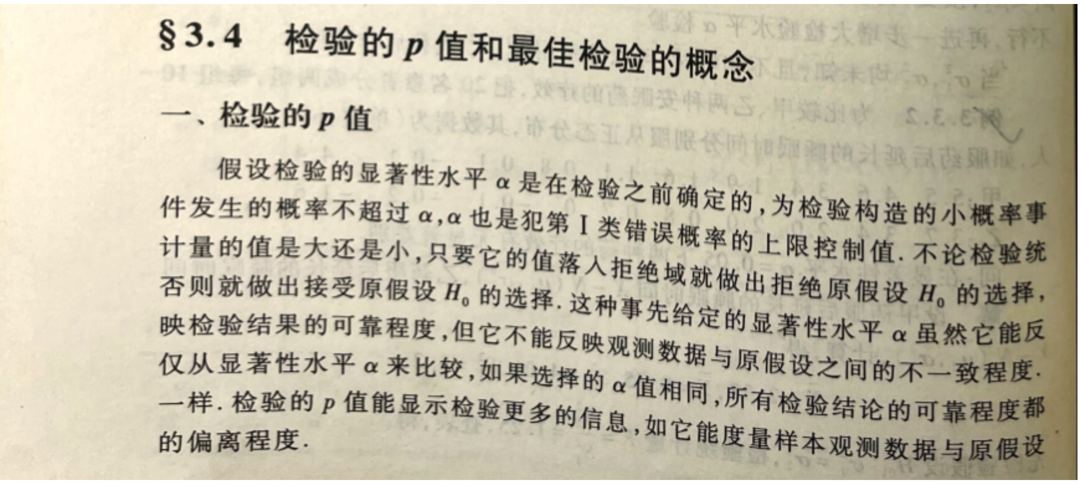
Solution. The Wald statistic is

$$W = \frac{\widehat{\delta} - 0}{\widehat{\sigma}} = \frac{(\overline{X} - \overline{Y}) - 0}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}} = \frac{216.19 - 195.27}{\sqrt{\frac{20^2}{16} + \frac{9.6^2}{16}}} = 3.78.$$

To compute the p-value, let $Z \sim N(0,1)$ denote a standard normal random variable. Then,

$$\text{p-value} = P(|Z| > 3.78) = 2P(Z < -3.78) = 0.0002,$$

which is very strong evidence against the null hypothesis.

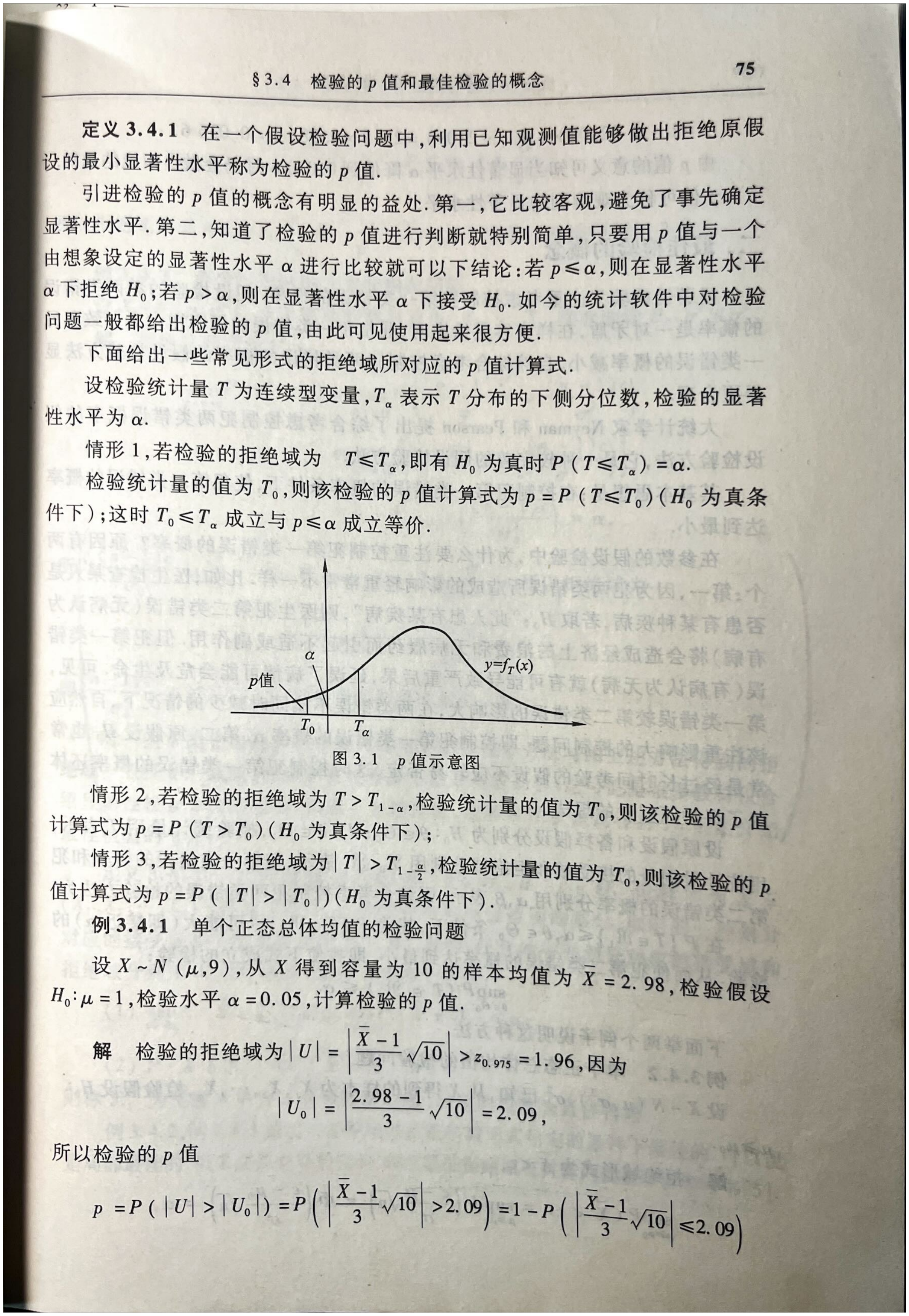


§3.4 检验的 p 值和最佳检验的概念

一、检验的 p 值

假设检验的显著性水平 α 是在检验之前确定的，为检验构造的小概率事
件发生的概率不超过 α，α 也是犯第 I 类错误概率的上限控制值. 不论检验统
计量的值是大还是小，只要它的值落入拒绝域就做出拒绝原假设 H_0 的选择，
否则就做出接受原假设 H_0 的选择. 这种事先给定的显著性水平 α 虽然它能反
映检验结果的可靠程度，但它不能反映观测数据与原假设之间的不一致程度.
仅从显著性水平 α 来比较，如果选择的 α 值相同，所有检验结论的可靠程度都
一样. 检验的 p 值能显示检验更多的信息，如它能度量样本观测数据与原假设
的偏离程度.

Figure 1.9:



定义 3.4.1 在一个假设检验问题中，利用已知观测值能够做出拒绝原假设的最小显著性水平称为检验的p值.

引进检验的p值的概念有明显的益处. 第一，它比较客观，避免了事先确定显著性水平. 第二，知道了检验的p值进行判断就特别简单，只要用p值与一个由想象设定的显著性水平α进行比较就可以下结论：若$p \leqslant \alpha$，则在显著性水平α下拒绝H_0；若$p > \alpha$，则在显著性水平α下接受H_0. 如今的统计软件中对检验问题一般都给出检验的p值，由此可见使用起来很方便.

下面给出一些常见形式的拒绝域所对应的p值计算式.

设检验统计量T为连续型变量，T_α表示T分布的下侧分位数，检验的显著性水平为α.

情形 1，若检验的拒绝域为 $T \leqslant T_\alpha$，即有H_0为真时$P(T \leqslant T_\alpha) = \alpha$.

检验统计量的值为T_0，则该检验的p值计算式为$p = P(T \leqslant T_0)$（H_0为真条件下）；这时$T_0 \leqslant T_\alpha$成立与$p \leqslant \alpha$成立等价.

图 3.1 p值示意图

情形 2，若检验的拒绝域为$T > T_{1-\alpha}$，检验统计量的值为T_0，则该检验的p值计算式为$p = P(T > T_0)$（H_0为真条件下）；

情形 3，若检验的拒绝域为$|T| > T_{1-\frac{\alpha}{2}}$，检验统计量的值为$T_0$，则该检验的$p$值计算式为$p = P(|T| > |T_0|)$（$H_0$为真条件下）.

例 3.4.1 单个正态总体均值的检验问题

设$X \sim N(\mu, 9)$，从X得到容量为 10 的样本均值为$\overline{X} = 2.98$，检验假设$H_0: \mu = 1$，检验水平$\alpha = 0.05$，计算检验的p值.

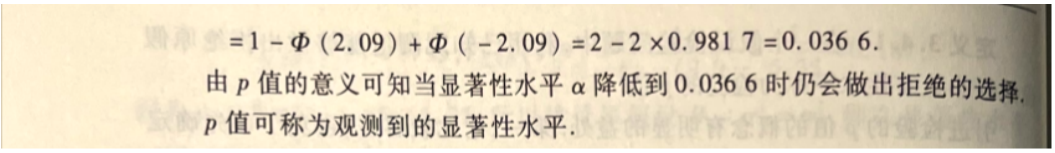
解 检验的拒绝域为$|U| = \left|\frac{\overline{X}-1}{3}\sqrt{10}\right| > z_{0.975} = 1.96$，因为

$$|U_0| = \left|\frac{2.98-1}{3}\sqrt{10}\right| = 2.09,$$

所以检验的p值

$$p = P(|U| > |U_0|) = P\left(\left|\frac{\overline{X}-1}{3}\sqrt{10}\right| > 2.09\right) = 1 - P\left(\left|\frac{\overline{X}-1}{3}\sqrt{10}\right| \leqslant 2.09\right)$$

Figure 1.10:



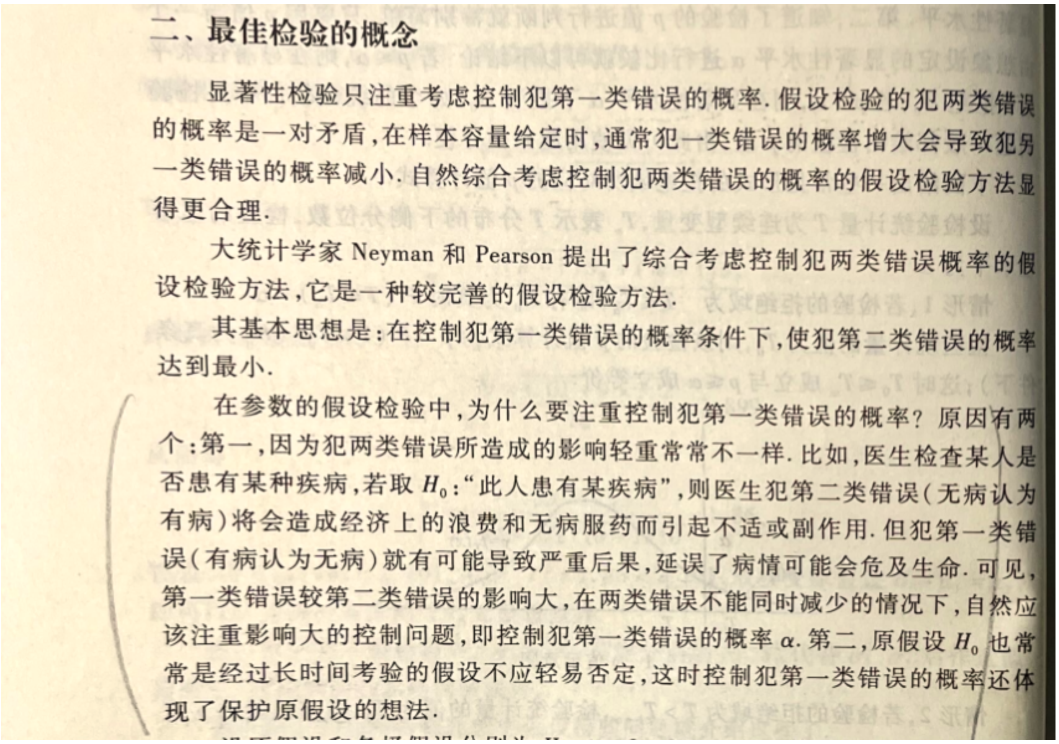
$= 1 - \Phi(2.09) + \Phi(-2.09) = 2 - 2 \times 0.981\,7 = 0.036\,6.$

由 p 值的意义可知当显著性水平 α 降低到 0.036 6 时仍会做出拒绝的选择. p 值可称为观测到的显著性水平.

Figure 1.11:

1.4.7 Concept for the Best Test

下述为什么第一类显著性检验相比于犯第二类错误更重要



二、最佳检验的概念

显著性检验只注重考虑控制犯第一类错误的概率. 假设检验的犯两类错误的概率是一对矛盾,在样本容量给定时,通常犯一类错误的概率增大会导致犯另一类错误的概率减小. 自然综合考虑控制犯两类错误的概率的假设检验方法显得更合理.

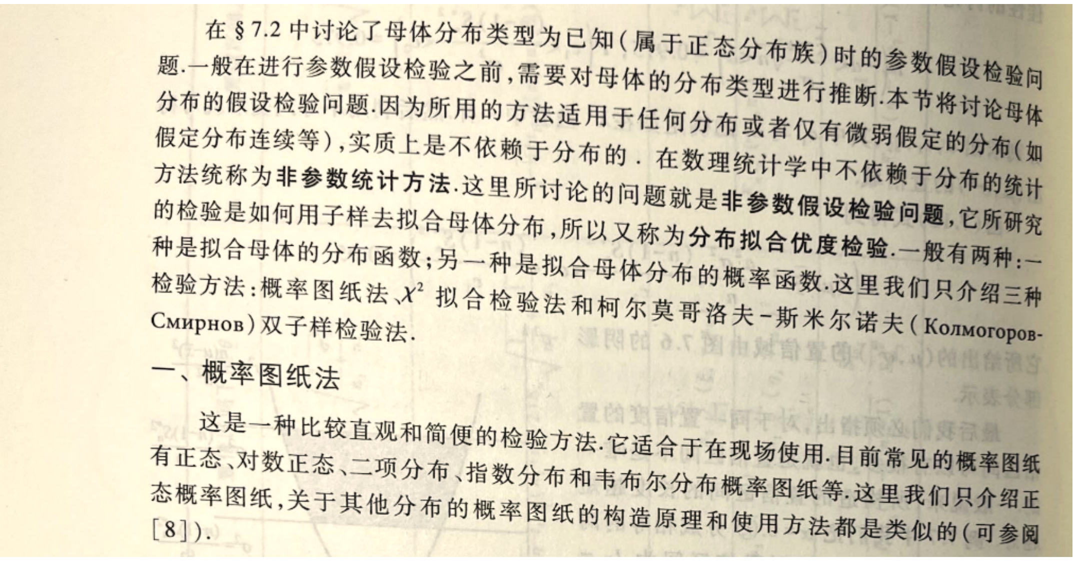
大统计学家 Neyman 和 Pearson 提出了综合考虑控制犯两类错误概率的假设检验方法,它是一种较完善的假设检验方法.

其基本思想是:在控制犯第一类错误的概率条件下,使犯第二类错误的概率达到最小.

在参数的假设检验中,为什么要注重控制犯第一类错误的概率? 原因有两个:第一,因为犯两类错误所造成的影响轻重常常不一样. 比如,医生检查某人是否患有某种疾病,若取 H_0:"此人患有某疾病",则医生犯第二类错误(无病认为有病)将会造成经济上的浪费和无病服药而引起不适或副作用. 但犯第一类错误(有病认为无病)就有可能导致严重后果,延误了病情可能会危及生命. 可见,第一类错误较第二类错误的影响大,在两类错误不能同时减少的情况下,自然应该注重影响大的控制问题,即控制犯第一类错误的概率 α. 第二,原假设 H_0 也常常是经过长时间考验的假设不应轻易否定,这时控制犯第一类错误的概率还体现了保护原假设的想法.

Figure 1.12:

1.4.8 Nonparametric Hypothesis Test (Pearson's χ^2 Test)



在 § 7.2 中讨论了母体分布类型为已知(属于正态分布族)时的参数假设检验问题.一般在进行参数假设检验之前,需要对母体的分布类型进行推断.本节将讨论母体分布的假设检验问题.因为所用的方法适用于任何分布或者仅有微弱假定的分布(如假定分布连续等),实质上是不依赖于分布的. 在数理统计学中不依赖于分布的统计方法统称为**非参数统计方法**.这里所讨论的问题就是**非参数假设检验问题**,它所研究的检验是如何用子样去拟合母体分布,所以又称为**分布拟合优度检验**.一般有两种:一种是拟合母体的分布函数;另一种是拟合母体分布的概率函数.这里我们只介绍三种检验方法:概率图纸法、χ^2 拟合检验法和柯尔莫哥洛夫–斯米尔诺夫(Колмогоров-Смирнов)双子样检验法.

一、概率图纸法

这是一种比较直观和简便的检验方法.它适合于在现场使用.目前常见的概率图纸有正态、对数正态、二项分布、指数分布和韦布尔分布概率图纸等.这里我们只介绍正态概率图纸,关于其他分布的概率图纸的构造原理和使用方法都是类似的(可参阅[8]).

Figure 1.13:

We have ignored the probability fitting method, （概率图纸法） in which one can neither obtain high accuracy nor control the probability of making errors. In this section, we only focus on the Pearson's χ^2 test, in which one can control the probability of making the type I error like aforementioned significance tests.

Test for Multinomial Data and Discrete Distributions

Pearson's χ^2 test is used for multinomial data. Recall that if $X = (X_1, ..., X_k)$ has a multinomial (n, p) distribution, then the MLE of p is $\widehat{p} = (\widehat{p}_1, ..., \widehat{p}_k) = (X_1/n, ..., X_k/n)$.

Let $p_0 = (p_{01}, ..., p_{0k})$ be some fixed vector and suppose we want to test

$$H_0 : p = p_0 \ \text{ versus } H_1 : p \neq p_0.$$

Definition 1.4.34 ***Pearson's*** χ^2 ***statistic*** *is*

$$T = \sum_{j=1}^{k} \frac{(X_j - np_{0j})^2}{np_{0j}} = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j},$$

where $E_j = E(X_j) = np_{0j}$ *is the expected value of* X_j *under* H_0.

Theorem 1.4.35 *Under* H_0, $T \xrightarrow{d} \chi^2(k-1)$. *Hence the test: reject* H_0 *if* $T > \chi^2_\alpha(k-1)$ *has asymptotic level* α. *The p-value is* $P(\chi^2(k-1) > t)$ *where t is the observed value of the test statistic.*

10.18 Example (Mendel's peas). Mendel bred peas with round yellow seeds and wrinkled green seeds. There are four types of progeny: round yellow, wrinkled yellow, round green, and wrinkled green. The number of each type is multinomial with probability $p = (p_1, p_2, p_3, p_4)$. His theory of inheritance predicts that p is equal to

$$p_0 \equiv \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16} \right).$$

In $n = 556$ trials he observed $X = (315, 101, 108, 32)$. We will test $H_0 : p = p_0$ versus $H_1 : p \neq p_0$. Since, $np_{01} = 312.75, np_{02} = np_{03} = 104.25$, and $np_{04} = 34.75$, the test statistic is

$$\begin{aligned} \chi^2 &= \frac{(315 - 312.75)^2}{312.75} + \frac{(101 - 104.25)^2}{104.25} \\ &\quad + \frac{(108 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} = 0.47. \end{aligned}$$

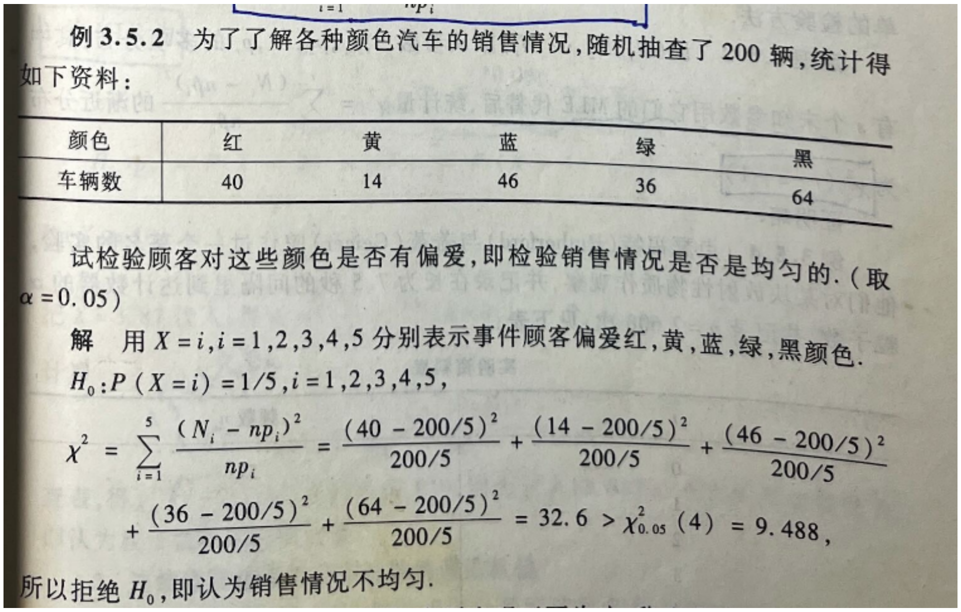
The $\alpha = .05$ value for a χ^2_3 is 7.815. Since 0.47 is not larger than 7.815 we do not reject the null. The p-value is

$$\text{p-value} = \mathbb{P}(\chi^2_3 > .47) = .93$$

which is not evidence against H_0. Hence, the data do not contradict Mendel's theory.[3]■

In the previous example, one could argue that hypothesis testing is not the right tool. Hypothesis testing is useful to see if there is evidence to reject H_0. This is appropriate when H_0 corresponds to the status quo. It is not useful for proving that H_0 is true. Failure to reject H_0 might occur because H_0 is true, but it might occur just because the test has low power. Perhaps a confidence set for the distance between p and p_0 might be more useful in this example.

Figure 1.14:



例 3.5.2 为了了解各种颜色汽车的销售情况，随机抽查了 200 辆，统计得如下资料：

颜色	红	黄	蓝	绿	黑
车辆数	40	14	46	36	64

试检验顾客对这些颜色是否有偏爱，即检验销售情况是否是均匀的.（取 $\alpha = 0.05$）

解 用 $X = i, i = 1,2,3,4,5$ 分别表示事件顾客偏爱红，黄，蓝，绿，黑颜色.

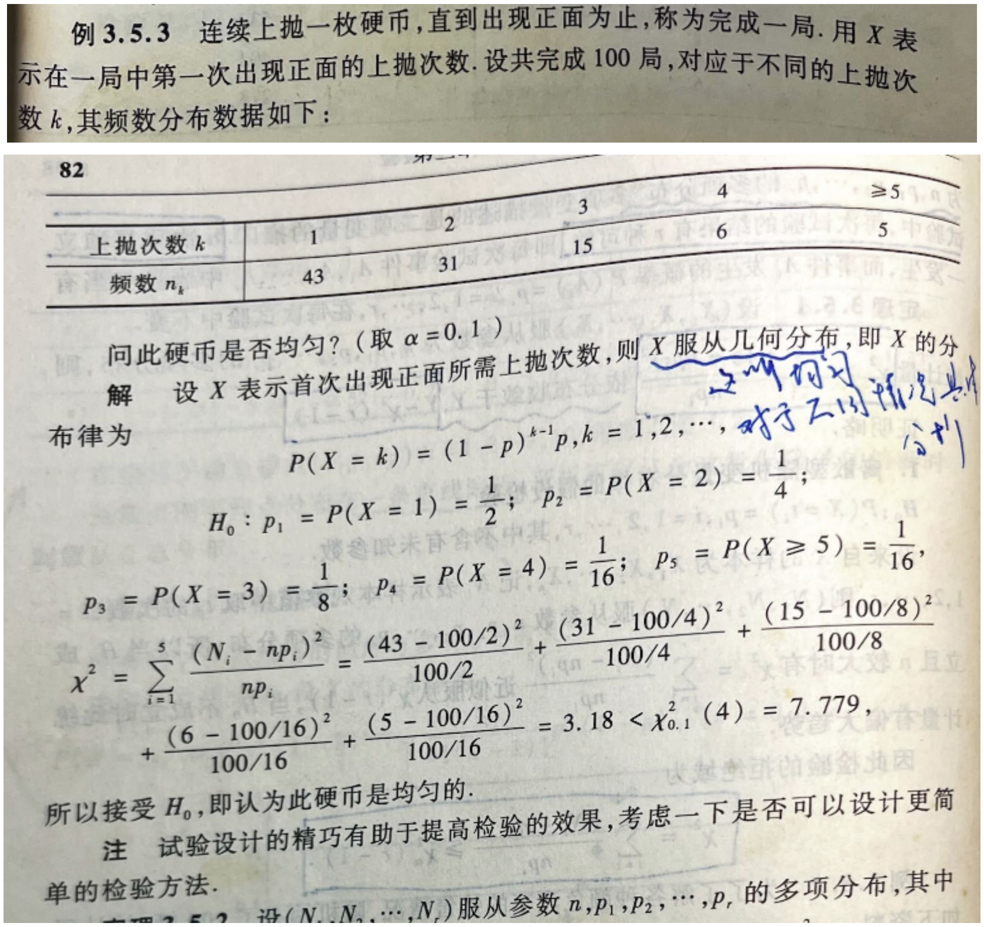
$H_0: P(X = i) = 1/5, i = 1,2,3,4,5$,

$$\chi^2 = \sum_{i=1}^{5} \frac{(N_i - np_i)^2}{np_i} = \frac{(40 - 200/5)^2}{200/5} + \frac{(14 - 200/5)^2}{200/5} + \frac{(46 - 200/5)^2}{200/5}$$

$$+ \frac{(36 - 200/5)^2}{200/5} + \frac{(64 - 200/5)^2}{200/5} = 32.6 > \chi^2_{0.05}(4) = 9.488,$$

所以拒绝 H_0，即认为销售情况不均匀.

Figure 1.15:



例 3.5.3 连续上抛一枚硬币，直到出现正面为止，称为完成一局. 用 X 表示在一局中第一次出现正面的上抛次数. 设共完成 100 局，对应于不同的上抛次数 k，其频数分布数据如下：

82

上抛次数 k	1	2	3	4	≥5
频数 n_k	43	31	15	6	5

问此硬币是否均匀？（取 $\alpha = 0.1$）

解 设 X 表示首次出现正面所需上抛次数，则 X 服从几何分布，即 X 的分布律为

这明确 对于不同情况具体分析

$$P(X = k) = (1 - p)^{k-1}p, k = 1,2,\cdots,$$

$$H_0: p_1 = P(X = 1) = \frac{1}{2}; \quad p_2 = P(X = 2) = \frac{1}{4};$$

$$p_3 = P(X = 3) = \frac{1}{8}; \quad p_4 = P(X = 4) = \frac{1}{16}; \quad p_5 = P(X \geqslant 5) = \frac{1}{16},$$

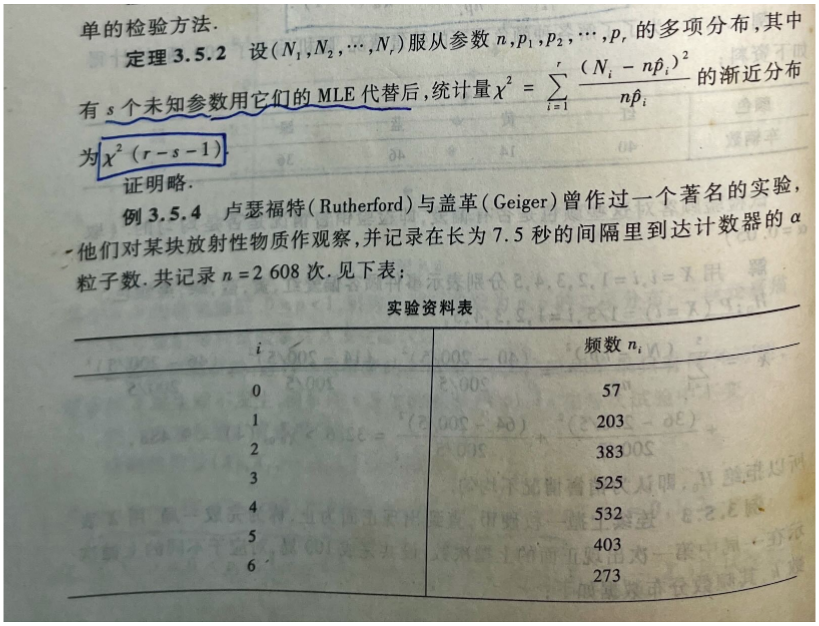
$$\chi^2 = \sum_{i=1}^{5} \frac{(N_i - np_i)^2}{np_i} = \frac{(43 - 100/2)^2}{100/2} + \frac{(31 - 100/4)^2}{100/4} + \frac{(15 - 100/8)^2}{100/8}$$

$$+ \frac{(6 - 100/16)^2}{100/16} + \frac{(5 - 100/16)^2}{100/16} = 3.18 < \chi^2_{0.1}(4) = 7.779,$$

所以接受 H_0，即认为此硬币是均匀的.

注 试验设计的精巧有助于提高检验的效果，考虑一下是否可以设计更简单的检验方法.

Figure 1.16:



单的检验方法.

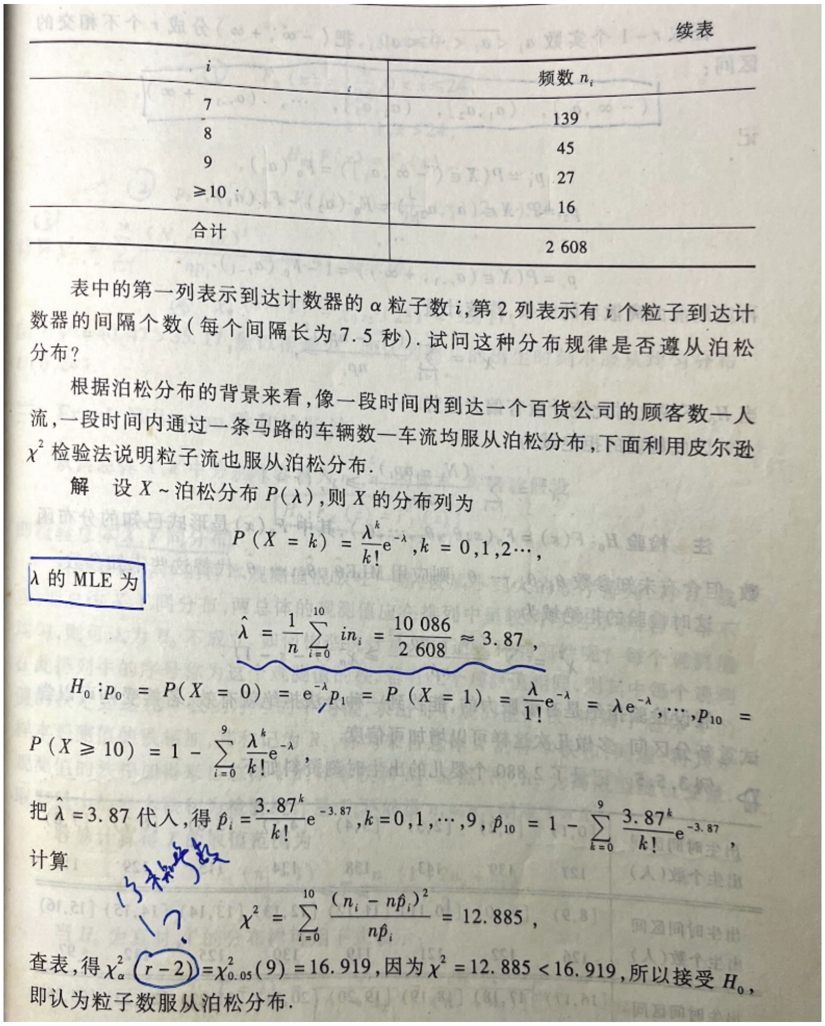
定理 3.5.2 设$(N_1,N_2,\cdots,N_r)$服从参数 $n,p_1,p_2,\cdots,p_r$ 的多项分布，其中有 s 个未知参数用它们的 MLE 代替后，统计量$\chi^2=\sum\limits_{i=1}^{r}\frac{(N_i-n\hat{p}_i)^2}{n\hat{p}_i}$ 的渐近分布为$\boxed{\chi^2(r-s-1)}$.

证明略.

例 3.5.4 卢瑟福特(Rutherford)与盖革(Geiger)曾作过一个著名的实验，他们对某块放射性物质作观察，并记录在长为 7.5 秒的间隔里到达计数器的 α 粒子数. 共记录 $n=2\ 608$ 次. 见下表:

实验资料表

i	频数 n_i
0	57
1	203
2	383
3	525
4	532
5	403
6	273



续表

i	频数 n_i
7	139
8	45
9	27
≥10	16
合计	2 608

表中的第一列表示到达计数器的 α 粒子数 i，第 2 列表示有 i 个粒子到达计数器的间隔个数(每个间隔长为 7.5 秒). 试问这种分布规律是否遵从泊松分布?

根据泊松分布的背景来看，像一段时间内到达一个百货公司的顾客数—人流，一段时间内通过一条马路的车辆数—车流均服从泊松分布，下面利用皮尔逊 χ^2 检验法说明粒子流也服从泊松分布.

解 设 $X\sim$泊松分布 $P(\lambda)$，则 X 的分布列为

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},k=0,1,2\cdots,$$

λ 的 MLE 为

$$\hat{\lambda}=\frac{1}{n}\sum_{i=0}^{10}in_i=\frac{10\ 086}{2\ 608}\approx 3.87,$$

$H_0:p_0=P(X=0)=e^{-\lambda},p_1=P(X=1)=\frac{\lambda}{1!}e^{-\lambda}=\lambda e^{-\lambda},\cdots,p_{10}=P(X\geqslant 10)=1-\sum\limits_{i=0}^{9}\frac{\lambda^k}{k!}e^{-\lambda}$,

把 $\hat{\lambda}=3.87$ 代入，得 $\hat{p}_i=\frac{3.87^k}{k!}e^{-3.87},k=0,1,\cdots,9,\hat{p}_{10}=1-\sum\limits_{k=0}^{9}\frac{3.87^k}{k!}e^{-3.87}$，

计算

13未知参数

17?

$$\chi^2=\sum_{i=0}^{10}\frac{(n_i-n\hat{p}_i)^2}{n\hat{p}_i}=12.885,$$

查表，得$\chi^2_\alpha\boxed{(r-2)}=\chi^2_{0.05}(9)=16.919$，因为$\chi^2=12.885<16.919$，所以接受 H_0，即认为粒子数服从泊松分布.

Figure 1.17:

Test for Continuous Distribution Function

Let the population X have the distribution function $F(x)$ (such as normal distribution, exponential distribution, binomial distribution, Poisson distribution, etc). Let us separate the range (R_a) of the random variable X into k disjoint intervals $A_1 = (a_0, a_1], A_2 = (a_1, a_2], \ldots, A_k = (a_{k-1}, a_k]$, where the length of each interval $a_j - a_{j-1}$ $(j = 1, \ldots, k)$ may be different. Here, $A_j A_m = \varnothing$ $(j \neq m, j, m = 1, \ldots, k)$ and $\cup_{j=1}^k A_j = R_a$. Let $x_1, \ldots, x_n$ be n observations of the population X and n_i is the number of observations in the set A_i such that $\sum_{i=1}^k n_i = n$. Hence among the n observations, the frequency for observing a data in A_i is $\frac{n_i}{n}$.

We now test **the null hypothesis** $H_0 : F(x) = F_0(x)$. If H_0 is true, the probability for the random variable $X \in A_i$ is p_i, where

$$p_i = P(A_i) = F_0(a_i) - F_0(a_{i-1}), \quad i = 1, \ldots, k. \tag{1.9}$$

Moreover, the probability, for n_1 observations in A_1, n_2 observations in $A_2, \ldots, n_k$ observations in A_k, is

$$\frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k},$$

which is a **multinomial** distribution. According to the law of large numbers, when H_0 is true, the frequency $\frac{n_i}{n}$ and the probability p_i should not have too much deviations. As a result, Pearson constructed a test statistic

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}, \tag{1.10}$$

which is called Pearson's χ^2 statistic. In the following, we will see that its limit distribution is an asymptotic χ^2 distribution with $k-1$ degrees of freedom.

For our ease, we first discuss the simple situation for $k = 2$. When H_0 is true,

$$P(A_1) = p_1, \quad P(A_2) = p_2,$$

where $p_1 + p_2 = 1$. We also have $n_1 + n_2 = n$. We now examine the quantity

$$\chi^2 = \frac{(n_1 - np_1)^2}{np_1} + \frac{(n_2 - np_2)^2}{np_2}.$$

Let

$$Y_1 = n_1 - np_1, \quad Y_2 = n_2 - np_2,$$

and then we see that

$$Y_1 + Y_2 = n_1 - np_1 + n_2 - np_2 = n - n(p_1 + p_2) = 0.$$

Hence, Y_1 and Y_2 are not independent. Let $Y_2 = -Y_1$, then we find that

$$\chi^2 = \frac{Y_1^2}{np_1} + \frac{Y_2^2}{np_2} = \frac{Y_1^2}{np_1 p_2} = \frac{(n_1 - np_1)^2}{np_1 p_2} = \left(\frac{n_1 - np_1}{\sqrt{np_1(1 - p_1)}} \right)^2.$$

We treat n_1 as a random variable with a binomial distribution $B(n, p_1)$. According to the de Moivre-Laplace Theorem in Section **??**, the random variable $\frac{n_1 - np_1}{\sqrt{np_1(1-p_1)}}$ has asymptotically normal distribution when n is sufficiently large. Thus the Pearson's χ^2 statistic is asymptotically $\chi^2(1)$ distributed with 1 degree of freedom when n is sufficiently large as $k = 2$. For general cases, we have the following theorem.

Theorem 1.4.36 *When H_0 is true, i.e., $p_1, \dots, p_k$ are the true probabilities of the population, the Pearson's χ^2 statistic defined by (1.10) has an asymptotic χ^2 distribution with $k-1$ degrees of freedom. Its pdf is given by*

$$f(x) = \begin{cases} \frac{1}{2^{\frac{k-1}{2}}\Gamma(\frac{k-1}{2})} x^{\frac{k-3}{2}} e^{-\frac{x}{2}}, & x > 0, \\ 0, & x \le 0 \end{cases}$$

Proof. The probability, for n_1 observations in A_1, n_2 observations in $A_2, \dots, n_k$ observations in A_k, is

$$P(N_1 = n_1, \dots, N_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k},$$

where $n_1 + \cdots + n_k = n$. The Ch.f. of $(N_1, \dots, N_k)$ is

$$\begin{aligned} \varphi_N(t_1, \dots, t_k) &= Ee^{i\mathbf{t}\cdot\mathbf{N}} = \sum_{\substack{n_1,\cdots,n_k \\ n_1+\cdots+n_k=n}} e^{i\mathbf{t}\cdot\mathbf{n}} \frac{n!}{n_1!n_2!\cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \\ &= \left(\sum_{j=1}^k p_j e^{it_j} \right)^n . \end{aligned}$$

Let

$$Y_i = \frac{n_i - np_i}{\sqrt{np_i}}, \quad i = 1, \dots, k. \tag{1.11}$$

Then we obviously have the relations

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^k Y_i^2,$$

with the constraint

$$\sum_{i=1}^k Y_i \sqrt{p_i} = 0.$$

Thus all random variables $(Y_1, \dots, Y_k)$ are not linearly independent. From the relation between Y_i and N_i in (1.11), the Ch.f. of $(Y_1, \dots, Y_k)$ is

$$\varphi(t_1, \dots, t_k) = \exp\left[-\sum_{j=1}^k it_j \sqrt{np_j} \right] \cdot \left(\sum_{j=1}^k p_j \exp\left(\frac{it_j}{\sqrt{np_j}} \right) \right)^n,$$

where the first term comes from the translation and the second term comes from the scaling multiplication. Take the logarithm on both sides,

$$\ln \varphi(t_1, \dots, t_k) = \left[-i\sqrt{n} \sum_{j=1}^k t_j \sqrt{p_j} \right] + n \ln \left[\sum_{j=1}^k p_j \exp\left(\frac{it_j}{\sqrt{np_j}} \right) \right].$$

We now Taylor expand the exponential and logarithm function at $t_j = 0$,

$$\begin{aligned} \exp\left(\frac{it_j}{\sqrt{np_j}} \right) - 1 &= \frac{it_j}{\sqrt{np_j}} - \frac{t_j^2}{2np_j} + o(\frac{1}{n}), \\ \ln(1+x) &= x - \frac{x^2}{2} + o(x^2). \end{aligned}$$

Then

$$\begin{aligned}
\ln \varphi(t_1,\dots,t_k) &= -i\sqrt{n}\sum_{j=1}^{k} t_j\sqrt{p_j} + n\ln\left[\sum_{j=1}^{k} p_j\left(1+\frac{it_j}{\sqrt{np_j}}-\frac{t_j^2}{2np_j}+o(\frac{1}{n})\right)\right]\\
&= -i\sqrt{n}\sum_{j=1}^{k} t_j\sqrt{p_j} + n\ln\left[1+\frac{i}{\sqrt{n}}\sum_{j=1}^{k} t_j\sqrt{p_j}-\frac{1}{2n}\sum_{j=1}^{k} t_j^2+o(\frac{1}{n})\right]\\
&= -i\sqrt{n}\sum_{j=1}^{k} t_j\sqrt{p_j} + n\left[\frac{i}{\sqrt{n}}\sum_{j=1}^{k} t_j\sqrt{p_j}-\frac{1}{2n}\sum_{j=1}^{k} t_j^2-\frac{1}{2}\left(\frac{i}{\sqrt{n}}\sum_{j=1}^{k} t_j\sqrt{p_j}\right)^2+o(\frac{1}{n})\right]\\
&= -\frac{1}{2}\sum_{j=1}^{k} t_j^2-\frac{1}{2}\left(i\sum_{j=1}^{k} t_j\sqrt{p_j}\right)^2+o(1).
\end{aligned}$$

As $n \to \infty$,

$$\ln \varphi(t_1,\dots,t_k) \to -\frac{1}{2}\sum_{j=1}^{k} t_j^2+\frac{1}{2}\left(\sum_{j=1}^{k} t_j\sqrt{p_j}\right)^2,$$

which is equivalent to

$$\lim_{n\to\infty}\varphi(t_1,\dots,t_k) = \exp\left(-\frac{1}{2}\left[\sum_{j=1}^{k} t_j^2-\left(\sum_{j=1}^{k} t_j\sqrt{p_j}\right)^2\right]\right). \tag{1.12}$$

We now take an orthogonal transformation:

$$\begin{bmatrix} Z_1\\ \vdots\\ Z_{k-1}\\ Z_k\end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & \cdots & a_{1k}\\ \vdots & \ddots & \ddots & \vdots\\ a_{k-1,1} & \cdots & \cdots & a_{k-1,k}\\ \sqrt{p_1} & \sqrt{p_2} & \cdots & \sqrt{p_k}\end{bmatrix}\begin{bmatrix} Y_1\\ \vdots\\ Y_{k-1}\\ Y_k\end{bmatrix},$$

where the transformation matrix $\mathbf{A}$ satisfies $\mathbf{A}^\top\mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}$. Correspondingly, the variables for the Ch.f.s have the linear transformation,

$$\begin{bmatrix} u_1\\ \vdots\\ u_{k-1}\\ u_k\end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & \cdots & a_{1k}\\ \vdots & \ddots & \ddots & \vdots\\ a_{k-1,1} & \cdots & \cdots & a_{k-1,k}\\ \sqrt{p_1} & \sqrt{p_2} & \cdots & \sqrt{p_k}\end{bmatrix}\begin{bmatrix} t_1\\ \vdots\\ t_{k-1}\\ t_k\end{bmatrix},$$

where $(u_1,\cdots,u_k)^\top$ corresponds to $(Z_1,\cdots,Z_k)^\top$ and $(t_1,\cdots,t_k)^\top$ corresponds to $(Y_1,\cdots,Y_k)^\top$. In the equation (1.12), we can compute the enclosed quantity,

$$\sum_{j=1}^{k} t_j^2-\left(\sum_{j=1}^{k} t_j\sqrt{p_j}\right)^2 = \sum_{j=1}^{k} u_j^2 - u_k^2 = \sum_{j=1}^{k-1} u_j^2.$$

Thus, as $n \to \infty$, The Ch.f. of $(Z_1,\cdots,Z_k)$ becomes

$$\lim_{n\to\infty}\varphi(t_1,\dots,t_k) = \exp\left(-\frac{1}{2}\sum_{j=1}^{k-1} u_j^2\right).$$

This means that each of $Z_1, \cdots, Z_{k-1}$ has weak convergence of distribution to pairwisely independent normal distribution $N(0,1)$ (corresponding to the Ch.f. $\varphi(t_j) \to \exp\left(-\frac{1}{2}u_j^2\right)$ for $j = 1, \ldots, k-1$) whereas Z_k has convergence in probability to a constant 0 (corresponding to the Ch.f. $\varphi(t_k) \to 1$). Hence,

$$\chi^2 = \sum_{i=1}^{k} Y_i^2 = \mathbf{Y}^\top \mathbf{Y} = \mathbf{Y}^\top \mathbf{A}^\top \mathbf{A}\mathbf{Y} = \mathbf{Z}^\top \mathbf{Z} = \sum_{i=1}^{k} Z_i^2,$$

has an asymptotic χ^2 distribution with $k-1$ degrees of freedom. ■

Remark 1.4.37 *If in the null hypothesis H_0 only the the type of the distribution of the population is given (such as Gaussian, exponential, Poisson, etc) but with unknown parameters $\theta_1, \ldots, \theta_m$ in the distribution, we cannot directly apply the above theorem to test the null hypothesis. Fisher proved the following theorem to resolve the hypothesis test problem for a distribution with unknown parameters.*

Theorem 1.4.38 *Let $F(x; \theta_1, \ldots, \theta_m)$ be the true distribution of the population, where $\theta_1, \ldots, \theta_m$ are m unknown parameters. Replace $(\theta_1, \ldots, \theta_m)$ in $F(x; \theta_1, \ldots, \theta_m)$ with their maximum likelihood estimates (MLEs) $(\widehat{\theta}_1, \ldots, \widehat{\theta}_m)$, and futher replace $F(x)$ in (1.9) with $F(x; \widehat{\theta}_1, \ldots, \widehat{\theta}_m)$ to obtain*

$$\widehat{p}_i = F(a_i; \widehat{\theta}_1, \ldots, \widehat{\theta}_m) - F(a_{i-1}; \widehat{\theta}_1, \ldots, \widehat{\theta}_m). \tag{1.13}$$

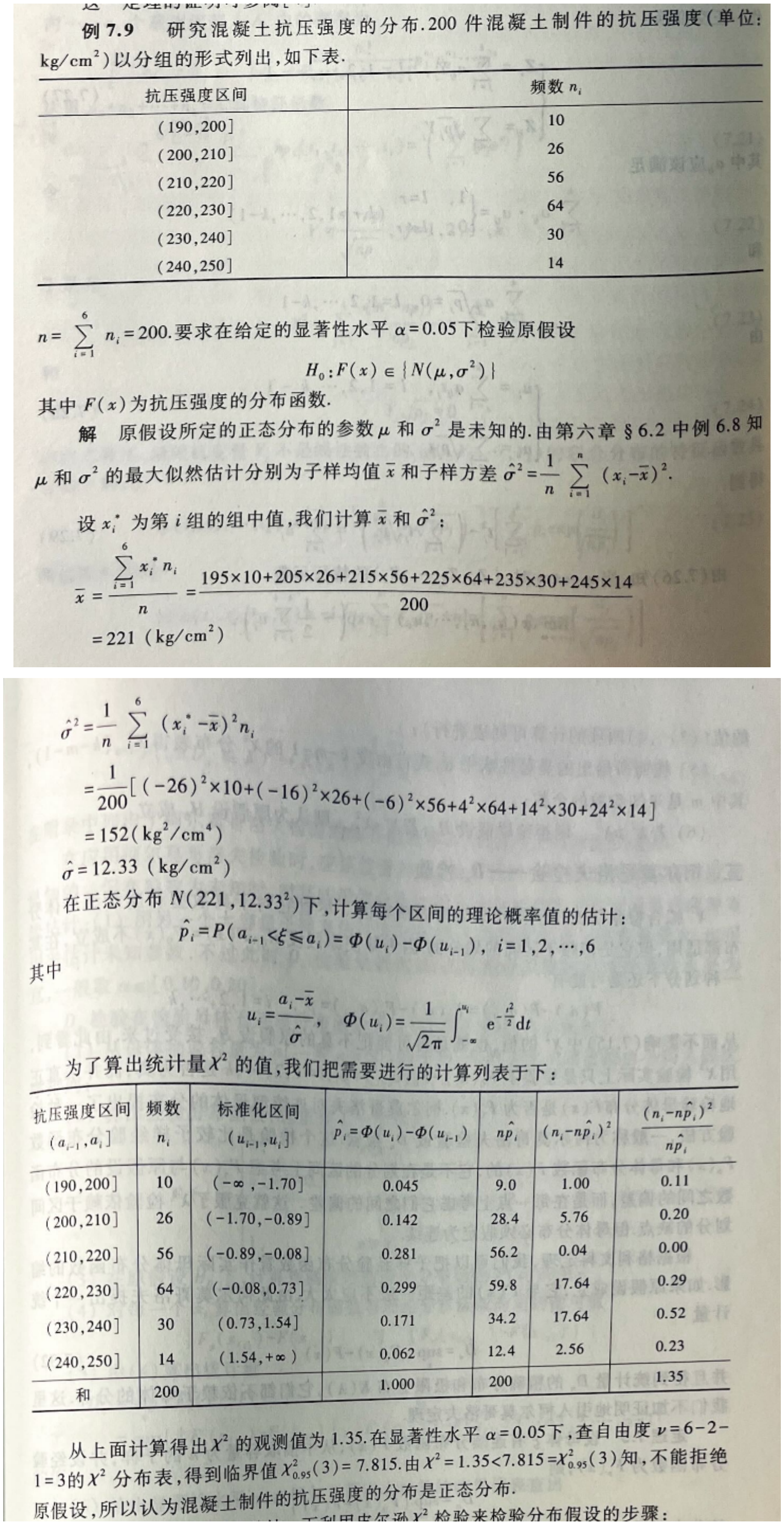
Substitute the above into (1.10) to obtain the test statistic,

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\widehat{p}_i)^2}{n\widehat{p}_i}.$$

Then the statistic χ^2 has an asymptotic χ^2 distribution with $k-m-1$ degrees of freedom as $n \to \infty$.

统计学数学方法 克拉美 1966

Proof. The proof can be found in one of H. Cramér's book. ■



例 7.9 研究混凝土抗压强度的分布.200 件混凝土制件的抗压强度(单位:kg/cm^2)以分组的形式列出,如下表.

抗压强度区间	频数 n_i
(190,200]	10
(200,210]	26
(210,220]	56
(220,230]	64
(230,240]	30
(240,250]	14

$n=\sum\limits_{i=1}^{6} n_i=200$.要求在给定的显著性水平 $\alpha=0.05$下检验原假设

$$H_0:F(x)\in\{N(\mu,\sigma^2)\}$$

其中 $F(x)$为抗压强度的分布函数.

解 原假设所定的正态分布的参数 μ 和 σ^2 是未知的.由第六章 § 6.2 中例 6.8 知 μ 和 σ^2 的最大似然估计分别为子样均值 $\bar{x}$ 和子样方差 $\hat{\sigma}^2=\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2$.

设 x_i^* 为第 i 组的组中值,我们计算 $\bar{x}$ 和 $\hat{\sigma}^2$:

$$\bar{x}=\frac{\sum\limits_{i=1}^{6}x_i^* n_i}{n}=\frac{195\times10+205\times26+215\times56+225\times64+235\times30+245\times14}{200}$$
$$=221\ (\mathrm{kg/cm^2})$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{6}(x_i^*-\bar{x})^2 n_i$$
$$=\frac{1}{200}[(-26)^2\times10+(-16)^2\times26+(-6)^2\times56+4^2\times64+14^2\times30+24^2\times14]$$
$$=152(\mathrm{kg^2/cm^4})$$
$$\hat{\sigma}=12.33\ (\mathrm{kg/cm^2})$$

在正态分布 $N(221,12.33^2)$下,计算每个区间的理论概率值的估计:

$$\hat{p}_i=P(a_{i-1}<\xi\leqslant a_i)=\Phi(u_i)-\Phi(u_{i-1}),\quad i=1,2,\cdots,6$$

其中

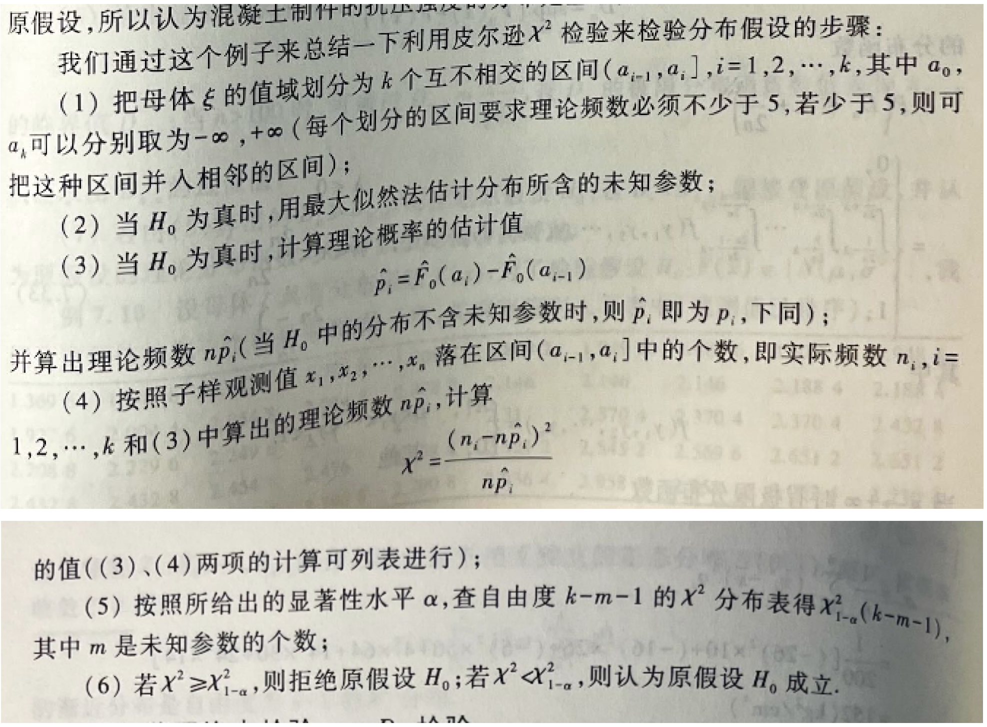
$$u_i=\frac{a_i-\bar{x}}{\hat{\sigma}},\quad \Phi(u_i)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u_i}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t$$

为了算出统计量 χ^2 的值,我们把需要进行的计算列表于下:

抗压强度区间 $(a_{i-1},a_i]$	频数 n_i	标准化区间 $(u_{i-1},u_i]$	$\hat{p}_i=\Phi(u_i)-\Phi(u_{i-1})$	$n\hat{p}_i$	$(n_i-n\hat{p}_i)^2$	$\frac{(n_i-n\hat{p}_i)^2}{n\hat{p}_i}$
(190,200]	10	$(-\infty,-1.70]$	0.045	9.0	1.00	0.11
(200,210]	26	$(-1.70,-0.89]$	0.142	28.4	5.76	0.20
(210,220]	56	$(-0.89,-0.08]$	0.281	56.2	0.04	0.00
(220,230]	64	$(-0.08,0.73]$	0.299	59.8	17.64	0.29
(230,240]	30	$(0.73,1.54]$	0.171	34.2	17.64	0.52
(240,250]	14	$(1.54,+\infty)$	0.062	12.4	2.56	0.23
和	200		1.000	200		1.35

从上面计算得出 χ^2 的观测值为 1.35.在显著性水平 $\alpha=0.05$下,查自由度 $\nu=6-2-1=3$的 χ^2 分布表,得到临界值 $\chi^2_{0.95}(3)=7.815$.由 $\chi^2=1.35<7.815=\chi^2_{0.95}(3)$知,不能拒绝原假设,所以认为混凝土制件的抗压强度的分布是正态分布.

Figure 1.18: This is the 1st example using χ^2 test for a continuous distribution.



原假设,所以认为混凝土制件的抗压强度的

我们通过这个例子来总结一下利用皮尔逊χ^2检验来检验分布假设的步骤:

(1) 把母体ξ的值域划分为k个互不相交的区间$(a_{i-1},a_i]$, $i=1,2,\cdots,k$,其中a_0, a_k可以分别取为$-\infty$, $+\infty$(每个划分的区间要求理论频数必须不少于5,若少于5,则可把这种区间并入相邻的区间);

(2) 当H_0为真时,用最大似然法估计分布所含的未知参数;

(3) 当H_0为真时,计算理论概率的估计值

$$\hat{p}_i=\hat{F}_0(a_i)-\hat{F}_0(a_{i-1})$$

并算出理论频数$n\hat{p}_i$(当H_0中的分布不含未知参数时,则$\hat{p}_i$即为p_i,下同);

(4) 按照子样观测值$x_1,x_2,\cdots,x_n$落在区间$(a_{i-1},a_i]$中的个数,即实际频数n_i, $i=1,2,\cdots,k$和(3)中算出的理论频数$n\hat{p}_i$,计算

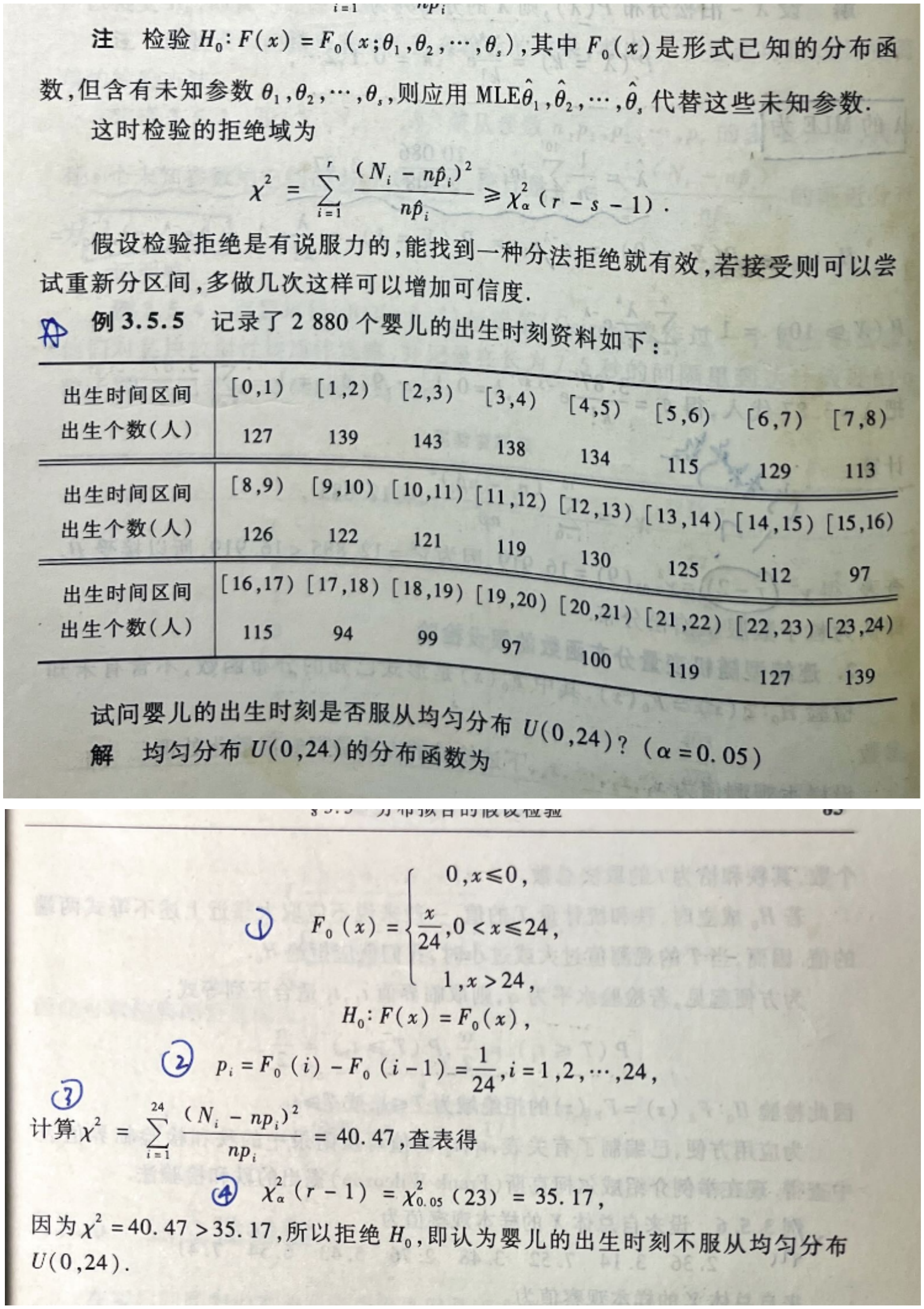
$$\chi^2=\frac{(n_i-n\hat{p}_i)^2}{n\hat{p}_i}$$

的值((3)、(4)两项的计算可列表进行);

(5) 按照所给出的显著性水平α,查自由度$k-m-1$的χ^2分布表得$\chi^2_{1-\alpha}(k-m-1)$,其中m是未知参数的个数;

(6) 若$\chi^2\geqslant\chi^2_{1-\alpha}$,则拒绝原假设$H_0$;若$\chi^2<\chi^2_{1-\alpha}$,则认为原假设$H_0$成立.

Figure 1.19: Summary of χ^2 test for a continuous distribution.



注 检验 $H_0: F(x) = F_0(x;\theta_1,\theta_2,\cdots,\theta_s)$，其中 $F_0(x)$ 是形式已知的分布函数，但含有未知参数 $\theta_1,\theta_2,\cdots,\theta_s$，则应用 MLE$\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_s$ 代替这些未知参数.

这时检验的拒绝域为

$$\chi^2 = \sum_{i=1}^{r}\frac{(N_i - n\hat{p}_i)^2}{n\hat{p}_i} \geqslant \chi^2_\alpha(r-s-1).$$

假设检验拒绝是有说服力的，能找到一种分法拒绝就有效，若接受则可以尝试重新分区间，多做几次这样可以增加可信度.

例 3.5.5 记录了 2 880 个婴儿的出生时刻资料如下：

出生时间区间	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)
出生个数(人)	127	139	143	138	134	115	129	113
出生时间区间	[8,9)	[9,10)	[10,11)	[11,12)	[12,13)	[13,14)	[14,15)	[15,16)
出生个数(人)	126	122	121	119	130	125	112	97
出生时间区间	[16,17)	[17,18)	[18,19)	[19,20)	[20,21)	[21,22)	[22,23)	[23,24)
出生个数(人)	115	94	99	97	100	119	127	139

试问婴儿的出生时刻是否服从均匀分布 $U(0,24)$？$(\alpha = 0.05)$

解 均匀分布 $U(0,24)$ 的分布函数为

$$① \quad F_0(x) = \begin{cases} 0, x\leqslant 0, \\ \dfrac{x}{24}, 0 < x \leqslant 24, \\ 1, x > 24, \end{cases}$$

$$H_0: F(x) = F_0(x),$$

$$② \quad p_i = F_0(i) - F_0(i-1) = \frac{1}{24}, i = 1,2,\cdots,24,$$

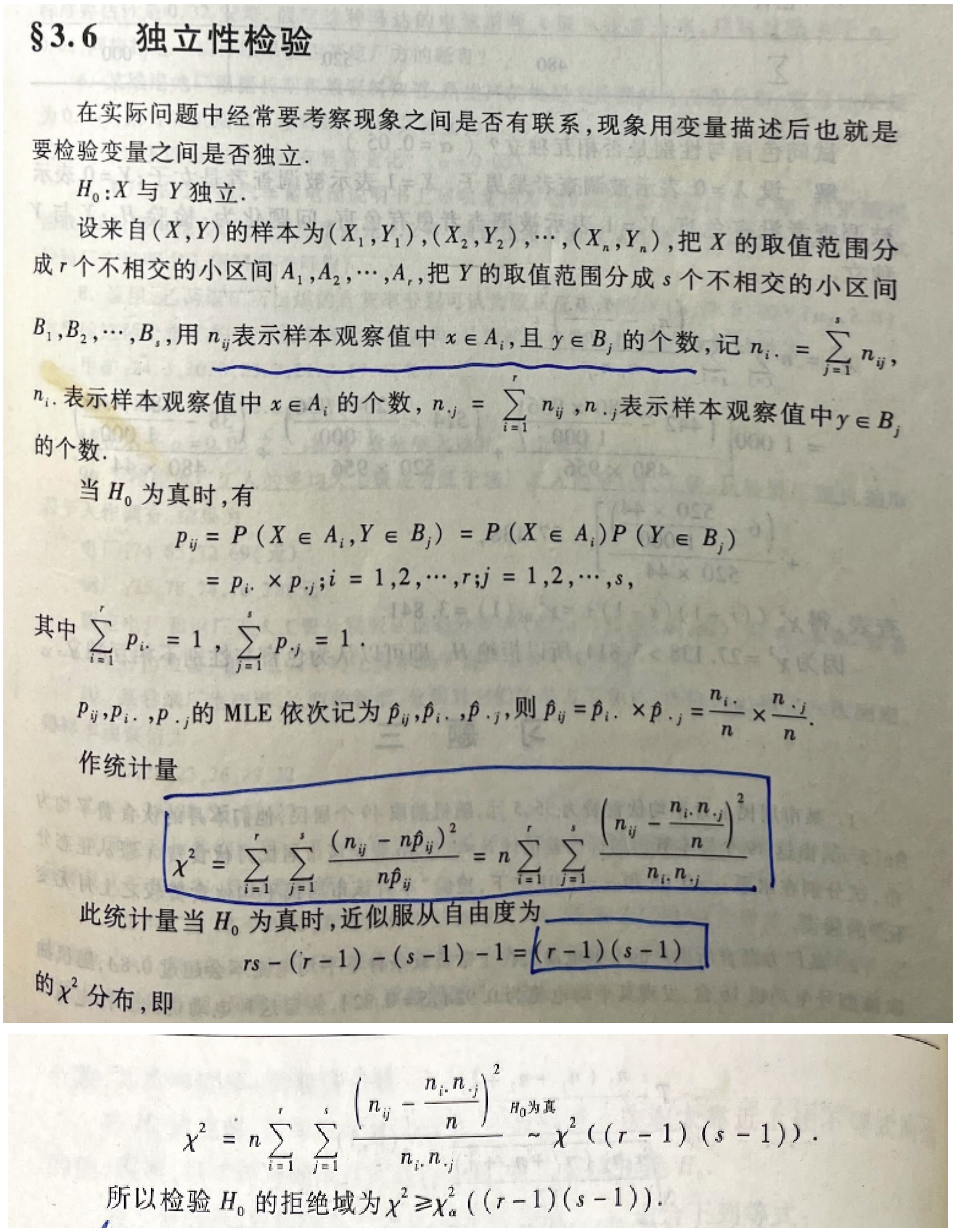
③

计算 $\chi^2 = \sum_{i=1}^{24}\frac{(N_i - np_i)^2}{np_i} = 40.47$，查表得

$$④ \quad \chi^2_\alpha(r-1) = \chi^2_{0.05}(23) = 35.17,$$

因为 $\chi^2 = 40.47 > 35.17$，所以拒绝 H_0，即认为婴儿的出生时刻不服从均匀分布 $U(0,24)$.

Figure 1.20: This is the 2nd example using χ^2 test for a continuous distribution.



§3.6 独立性检验

在实际问题中经常要考察现象之间是否有联系，现象用变量描述后也就是要检验变量之间是否独立．

H_0：X 与 Y 独立．

设来自(X,Y)的样本为$(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)$，把 X 的取值范围分成 r 个不相交的小区间 $A_1,A_2,\cdots,A_r$，把 Y 的取值范围分成 s 个不相交的小区间 $B_1,B_2,\cdots,B_s$，用 n_{ij}表示样本观察值中 $x\in A_i$，且 $y\in B_j$ 的个数，记 $n_{i\cdot}=\sum\limits_{j=1}^{s}n_{ij}$，$n_{i\cdot}$表示样本观察值中 $x\in A_i$ 的个数，$n_{\cdot j}=\sum\limits_{i=1}^{r}n_{ij}$，$n_{\cdot j}$表示样本观察值中$y\in B_j$ 的个数．

当 H_0 为真时，有

$$\begin{aligned} p_{ij} &= P(X\in A_i,Y\in B_j)=P(X\in A_i)P(Y\in B_j)\\ &= p_{i\cdot}\times p_{\cdot j};\ i=1,2,\cdots,r;\ j=1,2,\cdots,s, \end{aligned}$$

其中 $\sum\limits_{i=1}^{r}p_{i\cdot}=1$，$\sum\limits_{j=1}^{s}p_{\cdot j}=1$．

$p_{ij},p_{i\cdot},p_{\cdot j}$的 MLE 依次记为 $\hat{p}_{ij},\hat{p}_{i\cdot},\hat{p}_{\cdot j}$，则 $\hat{p}_{ij}=\hat{p}_{i\cdot}\times\hat{p}_{\cdot j}=\frac{n_{i\cdot}}{n}\times\frac{n_{\cdot j}}{n}$．

作统计量

$$\boxed{\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n\hat{p}_{ij})^2}{n\hat{p}_{ij}}=n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij}-\frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{n_{i\cdot}n_{\cdot j}}}$$

此统计量当 H_0 为真时，近似服从自由度为

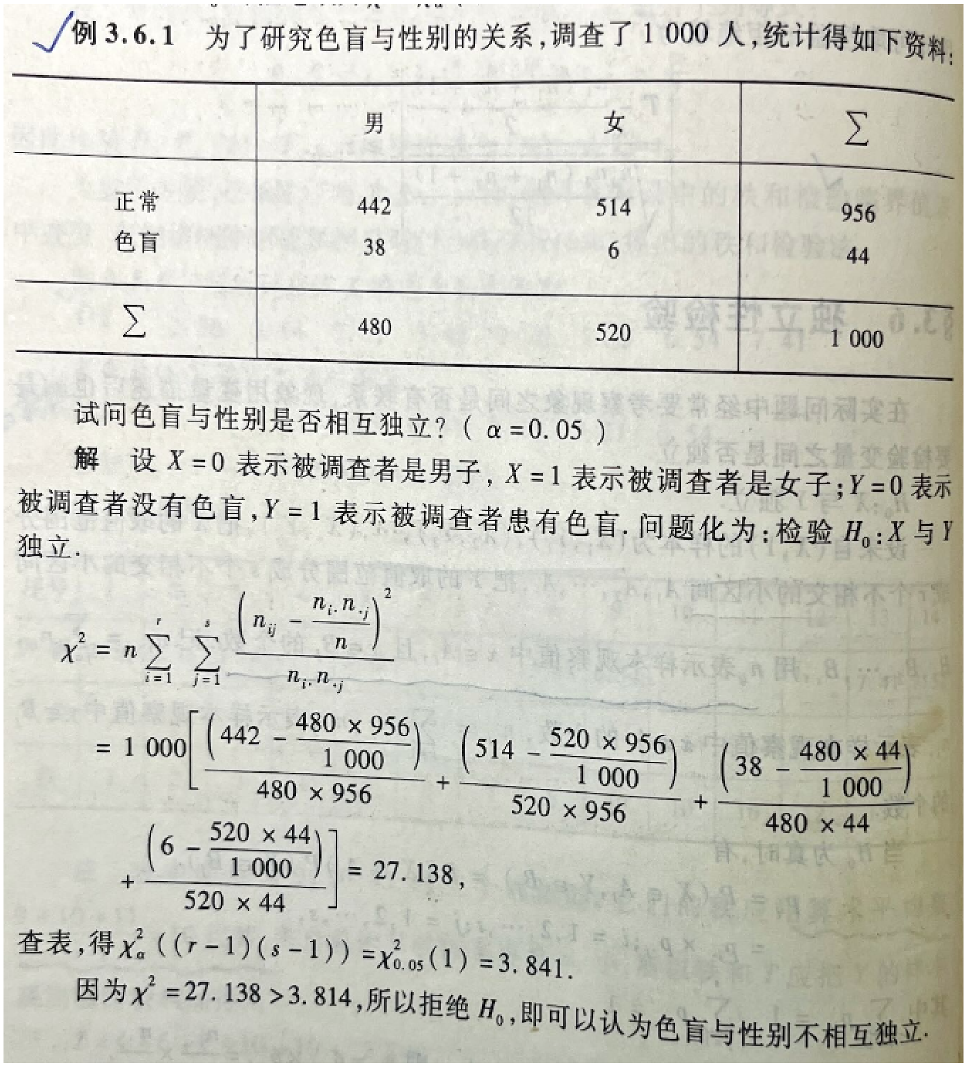
$$rs-(r-1)-(s-1)-1=\boxed{(r-1)(s-1)}$$

的 χ^2 分布，即

$$\chi^2=n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij}-\frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{n_{i\cdot}n_{\cdot j}}\overset{H_0\text{为真}}{\sim}\chi^2((r-1)(s-1)).$$

所以检验 H_0 的拒绝域为 $\chi^2\geqslant\chi^2_\alpha((r-1)(s-1))$．

Figure 1.21: Test for Independence.



例 3.6.1 为了研究色盲与性别的关系，调查了 1 000 人，统计得如下资料：

	男	女	$\sum$
正常 色盲	442 38	514 6	956 44
$\sum$	480	520	1 000

试问色盲与性别是否相互独立？（ $\alpha=0.05$ ）

解 设 $X=0$ 表示被调查者是男子，$X=1$ 表示被调查者是女子；$Y=0$ 表示被调查者没有色盲，$Y=1$ 表示被调查者患有色盲. 问题化为：检验 H_0：X 与 Y 独立.

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij}-\frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{n_{i\cdot}n_{\cdot j}}$$

$$= 1\,000\left[\frac{\left(442-\frac{480\times956}{1\,000}\right)}{480\times956}+\frac{\left(514-\frac{520\times956}{1\,000}\right)}{520\times956}+\frac{\left(38-\frac{480\times44}{1\,000}\right)}{480\times44}+\frac{\left(6-\frac{520\times44}{1\,000}\right)}{520\times44}\right]=27.138,$$

查表，得 $\chi^2_\alpha((r-1)(s-1))=\chi^2_{0.05}(1)=3.841$.

因为 $\chi^2=27.138>3.814$，所以拒绝 H_0，即可以认为色盲与性别不相互独立.

Figure 1.22: The example for the test for Independence

Bibliography