# A Statistics Primer
# Math 6070, Spring 2006

Davar Khoshnevisan
University of Utah

January 30, 2006

## Contents



## 1 Statistical Models

It is convenient to have an abstract framework for discussing statistical theory. The general problem is that there exists an unknown *parameter* $\theta_0$, which we wish to find out about. To have something concrete in mind, consider for example a population with the $N(\theta_0\,, 1)$ distribution, where $\theta_0$ is an unknown constant. If we do not have any *a priori* information about $\theta_0$ then it stands to reason that we consider every distribution of the form $N(\theta\,, 1)$, as $\theta$ ranges over $\mathbf{R}$, and then use data to make inference about the real, unknown $\theta_0$.

The general framework is this: We have a *parameter space* $\Theta$ and the real $\theta_0$ is in $\Theta$, but we do not its value. For every $\theta \in \Theta$, let $\mathrm{P}_\theta$ denote the underlying probability, which is computed by assuming that $\theta_0 = \theta$. Similarly define $\mathrm{E}_\theta$, $\mathrm{Var}_\theta$, $\mathrm{Cov}_\theta$, etc. Then, the idea is to take a sample—typically an independent sample—$\boldsymbol{X} = (X_1\,, \ldots, X_n)$—from $\mathrm{P}_{\theta_0}$. If the true (unknown) $\theta_0$ were equal to some (known) $\theta_1 \in \Theta$, then one would expect $\boldsymbol{X}$ to behave like an independent sample from $\mathrm{P}_{\theta_1}$. If so, then we declare that $\theta_0$ might well be $\theta_0$. Else, we reject the notion that $\theta_0 = \theta_1$. The remainder of these notes make this technique precise in more special settings.

## 2 Classical Parametric Inference

The typical problem of classical statistics is the following: Given a family of probability densities $\{f_\theta\}_{\theta\in\Theta}$ how can we decide whether or not ours is $f_\theta$? More precisely, we have an unknown density $f_{\theta_0}$; we wish to estimate it by choosing one from the family $\{f_\theta\}_{\theta\in\Theta}$ of densities available to us. [Alternatively, you could replace $f_\theta$ by a mass function $p_\theta$.] Here, $\Theta$ is the "parameter space," and $\theta_0$ is the unknown "parameter."

To estimate $\theta_0$ one typically considers an independent sample $X_1, \ldots, X_n$ from the true distribution with density $f_{\theta_0}$, and constructs an estimator $\hat{\theta}$.

**Example 1** Let $\Theta := \mathbf{R}$, and $f_\theta$ the $N(\theta\,, 1)$ density. The standard approach is to estimate $\theta_0$ with

$$\hat{\theta} := \frac{X_1 + \cdots + X_n}{n}. \tag{1}$$

There are many reasons why $\hat{\theta}$ is a good estimate of $\theta$.

1. [Unbiasedness] Evidently,

$$\mathrm{E}_\theta \hat{\theta} = \theta, \qquad \text{for all } \theta \in \Theta. \tag{2}$$

   This is called *unbiasedness*. In general, a random variable $T$ is said to be an *unbiased* estimator of $\theta$ if $\mathrm{E}_\theta T = \theta$ for all $\theta \in \Theta$.

2. [Consistency] By the law of large numbers, for all $\theta \in \Theta$,

$$\hat{\theta} \xrightarrow{\mathrm{P}_\theta} \theta \quad \text{ as } n \to \infty. \tag{3}$$

   This is called *consistency*. In general, a random variable $T$ is said to be a *consistent* estimator of $\theta_0$ if for all $\theta \in \Theta$, $T \xrightarrow{\mathrm{P}_\theta} \theta$ as the sample size tends to infinity.

3. [MLE] The *maximum likelihood estimate* of $\theta_0$—in all cases–is an estimator that maximizes $\theta \mapsto f_\theta(X_1 \ldots, X_n)$ for an independent sample $(X_1\,, \ldots, X_n)$, where $f_\theta$ here represents the joint density function of $n$ i.i.d. random variables each with density $N(\theta\,, 1)$. In the present example.

$$f_\theta(X_1\,, \ldots, X_n) = \frac{1}{(2\pi)^{n/2}} \exp\left( -\frac{1}{2} \sum_{j=1}^n (X_j - \theta)^2 \right). \tag{4}$$

   To find a MLE, it is easier to maximize the *log likelihood*,

$$L(\theta) := \ln f_\theta(X_1\,, \ldots, X_n), \tag{5}$$

   which is the same as minimizing $h(\theta) := \sum_{j=1}^n (X_j - \theta)^2$ over all $\theta$. But $h'(\theta) = -2\sum_{j=1}^n (X_j - \theta)$ and $h''(\theta) = 2n > 0$. Therefore, the MLE is uniquely $\hat{\theta}$.

The statistics $\hat\theta$ has other optimality features too. See for instance Example 8 (page 6) below.

**Example 2** Suppose $\Theta := \mathbf{R}\times(0\,,\infty)$. Then, we can write $\theta\in\Theta$ as $\theta=(\mu\,,\sigma^2)$ where $\mu\in\mathbf{R}$ and $\sigma>0$. Suppose $f_\theta$ is the $N(\mu\,,\sigma^2)$ density. Then the usual estimator for the true parameter $\theta_0=(\mu_0\,,\sigma_0^2)$ is $\hat\theta := (\hat\mu\,,\hat\sigma^2)$, where

$$\begin{aligned}\hat\mu &:= \frac{1}{n}\sum_{j=1}^n X_j,\\ \hat\sigma^2 &:= \frac{1}{n}\sum_{j=1}^n (X_j-\hat\mu)^2.\end{aligned} \tag{6}$$

[As before, $X_1,\ldots,X_n$ is an independent sample.] As in the previous example, $\hat\theta$ is the unique MLE, and is consistent. However, it is not unbiased. Indeed,

$$\mathrm{E}_\theta\hat\theta = \begin{pmatrix}\mu\\ \left[1-\frac{1}{n}\right]^2\sigma^2\end{pmatrix}, \qquad \text{for all } \theta=(\mu\,,\sigma^2)\in\Theta. \tag{7}$$

So $\hat\theta$ is "biased," although it is *asymptotically unbiased*; i.e., $\mathrm{E}_\theta\hat\theta\to\theta$ as $n\to\infty$.

**Example 3** Suppose $\Theta=(0\,,\infty)$, and $f_\theta$ is the uniform-$(0\,,\theta)$ density for all $\theta\in\Theta$. Given an independent sample $X_1,\ldots,X_n$, we consider

$$\hat\theta := \max_{1\le j\le n} X_j. \tag{8}$$

The distribution of $\hat\theta$ is easily computed, viz.,

$$\mathrm{P}_\theta\left\{\hat\theta\le a\right\} = \left[\mathrm{P}_\theta\{X_1\le a\}\right]^n = (a/\theta_0)^n, \qquad 0\le a\le\theta_0. \tag{9}$$

This gives the density $f_{\hat\theta}(a)=n\theta_0^{-n}a^{n-1}$ for $0\le a\le\theta_0$. Consequently,

$$\mathrm{E}_\theta\hat\theta = \theta_0^{-n}\int_0^{\theta_0} na^n\,da = \frac{n\theta_0}{n+1}. \tag{10}$$

Therefore: (i) $\hat\theta$ is biased; but (ii) it is asymptotically unbiased. Next we show that $\hat\theta$ is consistent. Note that $\hat\theta\le\theta_0$, by force. So it is enough to show that with high probability $\hat\theta$ is not too much smaller than $\theta_0$. Fix $\epsilon>0$, and note that

$$\mathrm{P}_\theta\left\{\hat\theta\le(1-\epsilon)\theta_0\right\} = \int_0^{(1-\epsilon)\theta_0} n\theta_0^{-n}a^{n-1}\,da = (1-\epsilon)^n. \tag{11}$$

Thus,

$$\mathrm{P}_\theta\left\{\left|\frac{\hat\theta}{\theta_0}-1\right|>\epsilon\right\} \le 1-(1-\epsilon)^n\to 0. \tag{12}$$

That is, $\hat{\theta}$ is consistent, as asserted earlier. To complete the example let us compute the MLE for $\theta_0$. Evidently,

$$f_\theta(X_1, \ldots, X_n) = \frac{1}{\theta^n} \mathbf{I}\{\theta > \hat{\theta}\}, \tag{13}$$

where $\mathbf{I}\{A\}$ is the indicator of $A$. So to find the MLE we observe that $\mathbf{I}\{A\} \leq 1$, so that $f_\theta(X_1, \ldots, X_n) \leq 1/\hat{\theta}^n$. The MLE is $\hat{\theta}$ uniquely.

One can consider a variant of $\hat{\theta}$, here, that is unbiased and consistent, but only "approximately" MLE for large $n$. Namely, we can consider the statistic $\tilde{\theta} := (n+1)\max_{1\leq j\leq n} X_j/n = (1+\frac{1}{n})\max_{1\leq j\leq n} X_j$.

# 3 The Information Inequality

Let us concentrate on the case where every $\theta \in \Theta$ is one-dimensional, and hence so is $\theta_0$.

Let $\boldsymbol{X} := (X_1, \ldots, X_n)$ be a random vector with joint density $f_\theta(\boldsymbol{x})$. The *Fisher information* of the family $\{f_\theta\}_{\theta\in\Theta}$ is defined as the function $I(\theta)$, where

$$I(\theta) := \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f_\theta(\boldsymbol{X})\right)^2\right], \qquad (\theta \in \Theta), \tag{14}$$

provided that the expectation exists and is finite. If $\boldsymbol{X}$ is discrete we define $I(\theta)$ in the same way, but replace $f_\theta$ by the joint mass function $p_\theta$.

In the continuous case, for example, the Fisher information is computed as follows:

$$\begin{aligned} I(\theta) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\ln f_\theta(\boldsymbol{x})\right)^2 f_\theta(\boldsymbol{x})\,d\boldsymbol{x} \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{1}{f_\theta(\boldsymbol{x})}\left(\frac{\partial}{\partial\theta}f_\theta(\boldsymbol{x})\right)^2 dx. \end{aligned} \tag{15}$$

So in fact $I(\theta)$ is always defined, but could be any number in $[0, \infty]$.

**Example 4** In the case of independent $N(\theta, 1)$'s,

$$\ln f_\theta(\boldsymbol{x}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{j=1}^n (x_j - \theta)^2. \tag{16}$$

The $\theta$-derivative is $\sum_{j=1}^n (x_j - \theta)$. Therefore,

$$I(\theta) = \mathrm{E}_\theta\left[\left(\sum_{j=1}^n X_j - n\theta\right)^2\right] = \mathrm{Var}_\theta\left(\sum_{j=1}^n X_j\right) = n. \tag{17}$$

[Here it does not depend on $\theta$.]

**Example 5** Suppose $X_1, \ldots, X_n \sim \text{Poisson}(\theta)$ are independent, where $\theta \in \Theta := (0\,,\infty)$. [Remember that "$Y \sim D$" means that "$Y$ is distributed as $D$."] Now we have the joint mass function $p_\theta(\boldsymbol{x})$ instead of densities. Then,

$$\ln p_\theta(\boldsymbol{x}) = -n\theta + \ln\theta \sum_{j=1}^n x_j - \sum_{j=1}^n \ln(x_j!). \tag{18}$$

Differentiate to obtain

$$\frac{\partial}{\partial\theta} \ln p_\theta(\boldsymbol{x}) = -n + \frac{1}{\theta}\sum_{j=1}^n x_j. \tag{19}$$

Therefore,

$$I(\theta) = \frac{1}{\theta^2} \mathrm{E}\left[\left(\sum_{j=1}^n X_j - n\theta\right)^2\right] = \frac{\mathrm{Var}(\sum_{j=1}^n X_j)}{\theta^2} = \frac{n}{\theta}. \tag{20}$$

The following is due to Fréchét originally, and was rediscovered independently, and later on, by Crámer and Rao.

**Theorem 6 (The Information Inequality)** *Suppose $T$ is a non-random function of $n$ variables. Then, under "mild regularity conditions,"*

$$\mathrm{Var}_\theta(T(\boldsymbol{X})) \geq \frac{[h'(\theta)]^2}{I(\theta)}, \tag{21}$$

*for all $\theta$, where $h(\theta) := \mathrm{E}_\theta[T(\boldsymbol{X})]$.*

The regularity conditions are indeed mild; they guarantee that certain integrals and derivatives commute. See (24) and (27) below.

The proof requires the following form of the Cauchy–Schwarz inequality:

**Lemma 7 (Cauchy–Schwarz Inequality)** *For all rv's $X$ and $Y$,*

$$|\mathrm{Cov}(X\,,Y)|^2 \leq \mathrm{Var}X \cdot \mathrm{Var}Y, \tag{22}$$

*provided that all the terms inside the expectations are integrable.*

**Proof.** Let $X' := (X - \mathrm{E}X)/\sqrt{\mathrm{Var}X}$ and $Y' := (Y - \mathrm{E}Y)/\sqrt{\mathrm{Var}Y}$. Then,

$$\begin{aligned} 0 \leq \mathrm{E}\left[(X'-Y')^2\right] &= \mathrm{E}[(X')^2] + \mathrm{E}[(Y')^2] - 2\mathrm{E}[X'Y'] \\ &= 2\left[1 - \frac{\mathrm{Cov}(X\,,Y)}{\sqrt{\mathrm{Var}X \cdot \mathrm{Var}Y}}\right]. \end{aligned} \tag{23}$$

This proves the result when $\mathrm{Cov}(X\,,Y) \geq 0$. When $\mathrm{Cov}(X\,,Y) < 0$, we consider instead $\mathrm{E}[(X'+Y')^2]$. □

**Proof of the Information Inequality in the Continuous Case.** Note that if $f_\theta$ is nice then

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} f_\theta(\boldsymbol{x})\, d\boldsymbol{x} = \frac{\partial}{\partial\theta}\left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_\theta(\boldsymbol{x})\, d\boldsymbol{x}\right] = 0. \tag{24}$$

This is so simply because $[\cdots] = 1$. Therefore,

$$\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta} \ln f_\theta(\boldsymbol{X})\right] = \int_{-\infty}^{\infty} f_\theta(\boldsymbol{x}) \frac{\partial}{\partial\theta} \ln f_\theta(\boldsymbol{x})\, d\boldsymbol{x} = 0. \tag{25}$$

This proves that

$$I(\theta) = \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta} \ln f_\theta(\boldsymbol{X})\right). \tag{26}$$

Similarly, if things are nice then

$$\begin{aligned}
\mathrm{E}_\theta\left[T(\boldsymbol{X}) \frac{\partial}{\partial\theta} \ln f_\theta(\boldsymbol{X})\right] &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T(\boldsymbol{x}) \frac{\partial}{\partial\theta} f_\theta(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \frac{\partial}{\partial\theta}\left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T(\boldsymbol{x}) f_\theta(\boldsymbol{x})\, d\boldsymbol{x}\right] \\
&= \frac{\partial}{\partial\theta} \mathrm{E}_\theta[T(\boldsymbol{X})] = h'(\theta).
\end{aligned} \tag{27}$$

Combine (24) and (27) to find that

$$\mathrm{Cov}_\theta\left(T(\boldsymbol{X}), \frac{\partial}{\partial\theta} \ln f_\theta(\boldsymbol{X})\right) = h'(\theta). \tag{28}$$

Thanks to Lemma 7,

$$|h'(\theta)|^2 \le \mathrm{Var}_\theta(T(\boldsymbol{X})) \cdot \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta} \ln f_\theta(\boldsymbol{X})\right) = \mathrm{Var}_\theta(T(\boldsymbol{X})) \cdot I(\theta). \tag{29}$$

See (26). This proves the information inequality. $\square$

A useful consequence of the information inequality is that, under mild conditions, any **unbiased** estimator $T(\boldsymbol{X})$ has the property that

$$\mathrm{Var}_\theta(T(\boldsymbol{X})) \ge \frac{1}{I(\theta)}. \tag{30}$$

This leads to the notion of MVU estimators: These are unbiased estimators that have minimum variance. Thanks to (30), if we can find a function $T$ such that $\mathrm{Var}T(\boldsymbol{X}) = 1/I(\theta_0)$, then we have found an MVU estimator of $\theta$.

**Example 8** Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\theta, 1)$'s. Let $T$ be such that $T(\boldsymbol{X})$ is an unbiased estimator of $\theta$. According to Example 4, $I(\theta) = n$, so that $\mathrm{Var}_\theta(T(\boldsymbol{X})) \ge 1/n = \mathrm{Var}_\theta \bar{X}_n$. That is, $\hat{\theta} := (X_1 + \cdots + X_n)/n$ has the smallest variance among all unbiased estimators of $\theta$. This is the "MVU" property. More precisely, any estimator $\hat{\theta}$ is said to be *MVUE* when it is a (often, "the") *minimum variance unbiased estimator* of $\theta_0$.

**Example 9** Suppose $X_1, \ldots, X_n$ are Poisson$(\theta)$, where $\theta > 0$ is an unknown parameter. [The true parameter is some unknown $\theta_0$, so we model it this way.] Because $\mathrm{E}_\theta X_1 = \theta$, the law of large numbers implies that

$$\bar{X}_n := \frac{X_1 + \cdots + X_n}{n} \stackrel{\mathrm{P}_\theta}{\to} \theta. \tag{31}$$

So, $\bar{X}_n$ is a consistent estimator of $\theta_0$. Recall also that $\mathrm{Var}_\theta X_1 = \theta$, so that $\mathrm{Var}_\theta \bar{X}_n = \theta/n$. We claim that $\bar{X}_n$ is a minimum variance unbiased estimator. In order to prove it it suffices to show that $I(\theta) = n/\theta$. But this was shown to be the case already; see Example 5 on page 5.

# 4 A Glance at Confidence Intervals

Choose and fix $\alpha \in (0\,,1)$. A *confidence set $C$ with level* $(1-\alpha)$ is a random set that depends on the sample $\boldsymbol{X}$, and has the property that $\mathrm{P}_\theta\{\theta \in C\} \geq 1-\alpha$ for all $\theta \in \Theta$. If $C$ varies with $n$, and $\lim_{n\to\infty} \mathrm{P}_\theta\{\theta \in C\} \geq 1-\alpha$ for all $\theta \in \Theta$, then we say that $C$ is a confidence interval for $\theta_0$ with *asymptotic level* $(1-\alpha)$.

**Example 10** Consider the model $N(\theta\,,1)$ where $\theta \in \Theta := \mathbf{R}$. Then, it easy to see that

$$\frac{\bar{X}_n - \theta}{1/\sqrt{n}} \sim N(0\,,1) \quad \text{under } \mathrm{P}_\theta. \tag{32}$$

Here, "Under $\mathrm{P}_\theta$" is short-hand for "If $\theta$ were the true parameter, for all $\theta \in \Theta$." Consider the random set

$$C(z) := \left[\bar{X}_n - \frac{z}{\sqrt{n}}\,, \bar{X}_n + \frac{z}{\sqrt{n}}\right], \tag{33}$$

where $z \geq 0$ is fixed. Then,

$$\begin{aligned} \mathrm{P}_\theta\left\{\theta \in C(z)\right\} &= \mathrm{P}_\theta\left\{|\bar{X}_n - \theta| \leq \frac{z}{\sqrt{n}}\right\} \\ &= \mathrm{P}_\theta\left\{\frac{|\bar{X}_n - \theta|}{1/\sqrt{n}} \leq z\right\} \\ &= \mathrm{P}\{|N(0\,,1)| \leq z\} = 2\Phi(z) - 1. \end{aligned} \tag{34}$$

See (32) for the last identity. Choose $z = z_{\alpha/2}$ such that $2\Phi(z_{\alpha/2}) - 1 = 1-\alpha$ to see that $\mathrm{P}_\theta\{\theta \in C(z_{\alpha/2})\} = 1-\alpha$. That is, $C(z_{\alpha/2})$ is a confidence interval for $\theta_0$ with level $1-\alpha$. Note that $z_{\alpha/2}$ is defined by $\Phi(z_{\alpha/2}) = 1-(\alpha/2)$. The numbers $z_{\alpha/2}$ are called "normal quantiles," because $\mathrm{P}\{N(0\,,1) \leq z_{\alpha/2}\} = \Phi(z_{\alpha/2}) = 1-(\alpha/2)$.

**Example 11** Consider the model Binomial$(n\,,p)$, where $n$ is a known integer, but $p \in [0\,,1]$ is an unknown constant. Here, $\Theta = [0\,,1]$, and every $p \in \Theta$ is a parameter. We consider the estimate

$$\hat{p} := \frac{S_n}{n}, \tag{35}$$

where $S_n$ denotes the total number of successes in $n$ independent samples. Evidently, $S_n \sim$ Binomial$(n\,,p)$ under $\mathrm{P}_p$. Therefore, $\mathrm{E}_p\hat{p} = p$ and $\mathrm{Var}_p\hat{p} = p(1-p)/n$.

By the central limit theorem, as $n$ tends to infinity,

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0\,,1), \tag{36}$$

under $\mathrm{P}_p$. (Why?) Equivalently,

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{d} N(0\,,1), \tag{37}$$

under $\mathrm{P}_p$. Also, by the law of large numbers, $\hat{p} \xrightarrow{\mathrm{P}_p} p$. (Why?) Apply the latter two results, via Slutsky's theorem, to find that under $\mathrm{P}_p$,

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \xrightarrow{d} N(0\,,1). \tag{38}$$

Now consider

$$C_n(z) := \left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,, \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]. \tag{39}$$

Then, we have shown that

$$\lim_{n\to\infty} \mathrm{P}_p\left\{p \in C_n(z)\right\} = \mathrm{P}\{|N(0\,,1)| \le z\} = 2\Phi(z) - 1. \tag{40}$$

Therefore, $C_n(z_{\alpha/2})$ is asymptotically a level-$(1-\alpha)$ confidence interval for $p$.

There are many variants of confidence intervals that are also useful. For instance, a *one-sided confidence interval* is a half-infinite random interval that should contain the parameter of interest with a pre-described probability. Similarly, there are one-sided confidence intervals that have a given asymptotic level. Finally, there are higher-dimensional generalizations. For example, there are confidence ellipsoids, confidence bands, etc. All of them are random sets—often with a pre-described geometry—that have exact or asymptotic level $(1-\alpha)$ for a pre-described level $\alpha \in (0\,,1)$.

# 5 A Glance at Testing Statistical Hypotheses

Someone proposes the theory that a certain coin is fair. To test this hypothesis, a statistician can flip the said coin $n$ times, independently. Record the number of heads $S_n$. In any event, we know that $S_n \sim$binomial$(n\,,p)$ for some $p$. Thus, we write the proposed hypothesis as the *null hypothesis*, $H_0:\ p=\frac{1}{2}$, versus the *alternative*, $H_1:\ p\neq\frac{1}{2}$. If the null hypothesis is correct, then $\hat{p}:=S_n/n$ is close to $p=1/2$ with high probability. Fix $\alpha\in(0\,,1)$, and consider the confidence interval $C_n(z_{\alpha/2})$ from Example 11 on page 8. It is more convenient to write $\mathrm{P}_{H_0}$ here instead of $\mathrm{P}_p$. With this in mind, we know then that for large $n$,

$$\mathrm{P}_{H_0}\left\{p\notin C_n(z_{\alpha/2})\right\}\approx\alpha. \tag{41}$$

Here is how we make an inference about $H_0$: If $p\notin C_n(z_{\alpha/2})$, then we reject the null hypothesis $H_0$. Else, we accept $H_0$, but only in the sense that we do not reject it. There are two sources of error in testing statistical hypotheses:

1. Type-I Error: This is the probability of incorrect rejection of $H_0$. In our example, (41) shows that the type-I error is asymptotically $\alpha$.

2. Type-II Error: This is the probability of incorrect acceptance of $H_1$. In our example, type-II error is

$$\beta=\mathrm{P}_{H_1}\left\{p\in C_n(z_{\alpha/2})\right\}, \tag{42}$$

which goes to zero as $n\to\infty$.

A slightly more general parametric testing problem is to decide between $H_0:\ \theta\in\Theta_0$ versus $H_1:\ \theta\in\Theta_1$, where $\Theta_0$ and $\Theta_1$ are subsets of $\Theta$. It need not be the case that $\Theta_0\cup\Theta_1=\Theta$, but it must be that $\Theta_0\cap\Theta_1=\varnothing$. Our answer is typically found by finding a confidence interval (or set, or ...) $C$ of a predescribed asymptotic level $(1-\alpha)$ such that $\mathrm{P}_{H_0}\{\theta\in C\}\approx 1-\alpha$, and hopefully $\mathrm{P}_{H_1}\{\theta\in C\}$ is small. If $C\cap\Theta_0=\varnothing$ then reject $H_0$, else accept $H_1$.