Chapter Three

Hypothesis Testing

3.1 Introduction

The final phase of analyzing data is to make a decision concerning a set of choices or options. Should I invest in stocks or bonds? Should a new product be marketed? Are my products being produced at the precise specifications? For each of the above questions one could perform an experiment, collect and summarize the data, and then make a decision. In most situations the decisions are all or nothing. I decide to market the new product or to not market it. I decide to buy stocks or I buy bonds.

One would hope that the data from the experiment would provide conclusive support for only one option and reject the others. Unfortunately, most experiments are subject to uncertainty, which means that decisions are sometimes correct and sometimes incorrect. A simple example provides insight into this process of decision-making. Suppose you are flipping a coin and must decide whether the coin is two-headed, fair, or unfair. If you flip ten heads in a row, the data support all three beliefs about the coin, although the favored choice would appear to be the two-headed coin. As soon as you flip a tail the data reveals that one choice-- the coin is two-headed--is clearly incorrect. Unfortunately the data may not clearly reveal whether the coin is fair or unfair. As the proportion of heads fluctuates around one-half the data appears to support the belief that the coin is fair or almost fair. That is, it is more likely that the probability of flipping a head is close to 0.50 or 0.49 rather than 0.01. As this example demonstrates, the data may or may not provide conclusive support for different sets of beliefs.

The title of this chapter suggests that choices of options can be specified as a set of hypotheses, which are then tested. There are differences between Bayesians and Classical statisticians in specifying and interpreting a hypothesis. In order to highlight the differences it is useful to ask the following questions of a Bayesian and a Classical statistician:

Q1: What is the role of prior information in hypothesis testing?

Q2: How does the existence of costs from making wrong decisions affect hypothesis testing?

Q3: What is the role of the sample size in reducing the probability of making a mistake?

The answers to these questions will show that Classical Hypothesis Testing does not adequately handle prior information, the costs from mistakes and exhibits a conflicting behavior over which types of mistakes are more important when the sample size is increased.

In the next section the basic tools of Hypothesis Testing are presented along with some examples. The Classical approach is presented and the Bayesian viewpoint is provided to point out problems (and distortions) from using the Classical approach.

3.2 Basic Tools of Hypothesis Testing

A simple way to introduce the components of hypothesis testing is to borrow the judicial analogy from Leamer (1978). Suppose you are a judge with two options; set a person free or send a person to jail. There are two competing hypotheses: the individual is either innocent (H_0) or guilty (H_1). After evaluating the evidence you must make a decision, which implies that you either accept H_0 or H_1. There are four possible outcomes, which are shown in the table below.

	Decision	
	Set Free	Send to Jail
Hypothesis	(Accept H_0)	(Accept $H_{1)}$
H_0: Innocent	Correct Decision	Incorrect Decision (Type I error)
H_1: Guilty	Correct Decision (Type II error)	Incorrect Decision

H_0 is referred to as the null, (or favored hypothesis) while H_1 is the alternative hypothesis. Two of the four outcomes result in making an incorrect decision. Statisticians creatively label these as Type I and Type II errors. The Type I error is sending an innocent person to jail. The Type II error is setting a guilty person free. As a judge you would like the probabilities of making a Type I or Type II error to equal zero. The Type I and Type II errors are defined as conditional probabilities:

α **= P(Type I error) = P(Reject $H_0 \backslash H_0$ true)**

β **= P(Type II error) = P(Reject $H_1 \backslash H_1$ true).**

In statistical experiments α and β are set prior to observing the data. In the above judicial example the court system determines the values for α and β by determining the type of evidence that can be presented to the court. Since the trial may yield conflicting evidence and testimony, you hope that the probability of Type I and Type II errors are small. In the United States, an individual is presumed innocent until proven guilty. This suggests that the probability of making a Type I error is less than the probability of making a Type II error.

We now attempt to answer question one by employing Bayes Rule. Define $P(H_0)$ and $P(H_1)$ as the respective probabilities that either hypothesis is true. Let the evidence from the trial represent the outcome from a probability distribution. Employing Bayes Rule we specify a posterior probability for H_0 given the evidence.

$$\mathbf{P(H_0/evidence) = [P(evidence/H_0) \times P(H_0)]/P(evidence)}$$

A similar posterior probability is defined for H_1. The ratio of the two posterior probabilities yields

$$\mathbf{P(H_0/evidence)/P(H_1/evidence) = [P(evidence/H_0)/P(evidence/H_1)] \times [P(H_0)/P(H_1)]}$$

The left-hand side of this expression is the posterior odds ratio. A ratio greater than one means that the null hypothesis is favored over the alternative hypothesis. The right hand side of the expression has two terms in brackets. The second bracket is the prior odds ratio. If both H_0 and H_1 are equally likely then the ratio is equal to one. The first term in the bracket is referred to as the Bayes Factor. These conditional probabilities are the probabilities of observing a specific amount of evidence given H_0 (or H_1) is true. If the data prefers H_0 then the Bayes factor is greater than one.

For a Bayesian the posterior odds ratio depends on the probability distribution that generates the data and the prior beliefs about H_0 and H_1. For example, if $P(H_0)$ is very close to one, then the data will have to strongly favor H_1 in order for the posterior odds ratio to favor H_1.

For the Classical statistician, the posterior odds ratio is meaningless. The prior probabilities for each hypothesis are either zero or one. The reason for this is that Classical inference assumes that the data is being generated from only one hypothesis. Hence, H_0 is either true or false, which means that $P(H_0)$ equals one or zero. In terms of the judicial example, the individual is either guilty or innocent. Either the person committed the crime or did not commit the crime. While the evidence may be inconclusive, it makes no sense to a Classical statistician to set $P(H_0)$ equal to a value other than zero or one. It does, however, make sense for the Classical statistician to talk about the probability of observing a specific set of data given a specific hypothesis. That is, it's appropriate to derive the probability for the event of flipping five heads in a row given the coin is fair, but inappropriate to establish a probability of whether or not the coin is fair. Either the coin is fair (or it is not) and the probability the coin is fair is one (and zero if it is not).

Since Classical inference excludes prior information, the data is allowed to determine the preferred hypothesis. The Bayesian will let the data determine the favored hypothesis only if the prior odds ratio is equal to one. It is a mistake to assume that letting the data determine the favored hypothesis is objective, while resorting to prior information is subjective. Both the Bayesian and the Classical statistician must specify the underlying probability distribution that generates the data. For example, the common assumption of a normal sampling distribution is merely a convenient and useful assumption, but one that is unlikely to be true. Thus, some subjectivity enters into the process from specifying the underlying sampling distribution.

To summarize, the answer to question one is that Bayesians apply Bayes Rule and use prior information to determine the favored hypothesis while Classical inference excludes the use of prior information.

3.3 A Numerical Example

In order to answer the second and third questions it is helpful to present a numerical example. Suppose that a machine in an assembly line is not very reliable. When the machine is working properly the probability of producing an unacceptable product is 0.25. When the machine is working improperly the probability is 0.50. If the machine is working improperly it can be adjusted for a nominal fee. At the beginning of the workday you must decide whether or not the machine is working properly. There are two competing hypotheses:

H_0: $\pi = 0.25$
H_1: $\pi = 0.50$

Assume that the number of unacceptable parts from a given sample will follow a binomial distribution. The probability distribution will depend on π and the sample size. Table 1 summarizes the probability distribution for each hypothesis for a sample size of five. For example, the probability of one unacceptable product is 0.2372 given H_0 is true, and 0.0312 given H_1 is true. The problem is to decide whether or not the machine is working properly. If we use the posterior odds ratio and assume that the prior odds ratio is equal to one, ($P(H_0) = P(H_1) = 1/2$), then the values of the "Bayes Factor" will equal the posterior odds ratio. In other words, the favored hypothesis will be the one with the higher conditional probability. As one can see from Table 3.3.1, H_0 is favored for values zero and one and H_1 is favored for the other values (2,3,4, and 5). As the number of unacceptable products increases we tend to favor H_1 over H_0. However, even in the case where the number of unacceptable products is five, there is still a positive probability(0.0010) that the machine is working properly.

Table 3.3.1
(n=5)

$X = x_i$	0	1	2	3	4	5
H_0: $\pi = 0.25$	.2372	.3955	.2637	.0879	.0146	.0010
H_1: $\pi = 0.50$	.0312	.1562	.3125	.3125	.1562	.0312

$$P(X=x) = C^n_x \ \pi^x (1 - \pi)^{n-x}$$

As already noted, the problem is to decide whether or not the machine is working properly. That is, do you decide to accept or reject H_0? To help you make a decision a random sample of five products is taken and the number of unacceptable products is counted. Since you must make an all or nothing decision regarding the acceptance of H_0, you choose a cutoff point. The cutoff point determines whether or not you accept H_0.

Let X denote the random variable for the number of unacceptable products. Suppose you decide to accept H_0 if the number of unacceptable products is at most three. The Type I and Type II errors are calculated as follows:

$$\begin{aligned}\alpha = P(\text{reject } H_0 \backslash H_0 \text{ true}) &= P(X > 3 \backslash H_0 \text{ true}) \\ &= 0.0146 + 0.0010 \\ &= 0.0156\end{aligned}$$

$$\beta = P(\text{reject } H_1 \backslash H_1 \text{ true}) = P(X \leq 3 \backslash H_1 \text{ true})$$
$$= 0.0312 + 0.1562 + 0.3125 + 0.3125$$
$$= 0.8124$$

With the above decision rule, the probability of making a Type I error is very small relative to the probability of making a Type II error. Choosing a cutoff value equal to three makes it very unlikely that I will decide that the machine is not working properly when in fact it is working properly. On the other hand it is very likely that I will conclude that the machine is working properly when in fact it is not.

An obvious question at this point is the selection of three as the cutoff value. Wasn't this choice arbitrary? The answer is yes. If we choose other cutoff values we can change the probabilities of the Type I and Type II errors. Table 3.3.2 reports the probabilities for the two errors when we change the cutoff value. As the table clearly shows, changing the cutoff value results in a tradeoff between α and β. Increasing the cutoff value reduces α but increases β. Similarly, lowering the cutoff value will increase α and reduce β. There is no cutoff value that sets both α and β equal to zero.

Table 3.3.2

<u>**Decision Rule:**</u> Accept H_0 if $X \leq C^*$,
Otherwise accept H_1 if $X > C^*$
C^* = cutoff value

C*	α	β
0	0.7627	0.0312
1	0.3672	0.1874
2	0.1035	0.4999
3	0.0156	0.8124
4	0.0010	0.9686
5	0.0000	1.0000

How does one choose the appropriate cutoff value? It would seem that further information is required. That is, if committing a Type I error is very costly shouldn't we set α equal to a very small number? Classical hypothesis testing tends to view H_0 as the favored hypothesis. Rejection of H_0 is considered only when it is very unlikely that the data was generated by the probability distribution assumed under H_0. Essentially this means that the cutoff value is chosen so that α is very small. Unfortunately the choice of setting α to small number is made without regard to the sample size and without explicitly specifying the costs associated from a Type I and Type II error. A Bayesian would select a cutoff value so that the expected costs from making a mistake are minimized. An example of the Bayesian approach is shown in the appendix.

Classical hypothesis testing sets α to a small number. The traditional values chosen for α are 1%, 5% and 10%. These values are chosen regardless of the values of β. In our example suppose we set the cutoff value at three. Referring to Table 2, α will be equal to 0.0156 and β will equal 0.8124. As this example demonstrates, you are more likely to make a Type II error rather than a Type I error.

Returning to answer question two we can make the following statements. Classical hypothesis testing, by choosing a fixed value for α, ignores the costs associated with the Type I and Type II errors and will choose a cutoff value that does not minimize expected costs. A Bayesian sets the cutoff value so that expected costs are minimized. As the appendix to this chapter indicates we are better off setting β to a small number, (rather than α), since the cost of making a Type II error is more expensive.

In hypothesis testing we can always take a larger sample and reduce the variance of the sampling distribution. A sample size equal to five yields very little information because the underlying probability distributions overlap each other. As we increase the sample size, there will be less of an overlap between the two distributions. Table 3.3.3 reveals the probability distributions for H_0 and H_1 when the sample size equals ten.

Table 3.3.3
(n=10)

$X = x_i$	0	1	2	3	4	5	6	7	8	9	10
H_0: π =0.25	0.0563	0.1877	0.2816	0.2503	0.1460	0.0584	0.0162	0.0031	0.0004	0.0000	0.0000
H_1: π =0.50	0.0010	0.0098	0.0439	0.1172	0.2051	0.2461	0.2051	0.1172	0.0439	0.0098	0.0010

For larger sample sizes Classical hypothesis testing keeps α fixed in setting the cutoff value. In the previous problem we set α equal to .0156. We can choose a similar value for α by choosing a new cutoff value at five.

Decision Rule: Accept H_0 if $X \leq 5$,
Otherwise reject H_0.

$$
\begin{aligned}
\alpha &= P(\text{reject } H_0/ H_0 \text{ true}) \\
&= P(X > 5/ H_0 \text{ true}) \\
&= 0.0162 + 0.0031 + 0.0004 \\
&= 0.0197
\end{aligned}
$$

$$
\begin{aligned}
\beta &= P(\text{reject } H_1/ H_1 \text{ true}) \\
&= P(\text{accept } H_0/ H_1 \text{ true}) \\
&= P(X \leq 5/ H_1 \text{ true}) \\
&= 0.0010 + 0.0098 + 0.0439 + 0.1172 + 0.2051 + 0.2461 \\
&= 0.6231
\end{aligned}
$$

Notice that by keeping α fixed at around 2%, β is reduced by increasing the sample size. As the sample size is increased, holding α fixed reduces the probability of a Type II error since less of the distribution will fall below the cutoff value. The reason for the declining value of β is due to the sampling variance declining with the sample size. (Recall that for the binomial distribution the sampling variance of the proportion is $\pi(1 - \pi)/n$.)

If we hold α at a fixed level of 2% we can eventually reduce β to almost zero as the sample size increases to larger number. Notice that keeping α fixed as the sample size grows implies that at some point β will be less than α. Originally we set α equal to a small number because we regarded the Type I error as the more serious mistake. However, as we increase the sample size, β decreases and at some point becomes smaller than α. By setting β less than α it seems that we now regard the Type II error as the more serious mistake. Is H_1 now the favored hypothesis? If we still want to keep H_0 as the favored hypothesis we must reduce α as the sample size grows. In other words, a larger sample size should lead us to reduce both α and β.

The Bayesian sets the cutoff value to minimize expected costs. If competing hypotheses are equally likely to be true and the costs are equal, then we can write expected cost in the following form:

$E(C) = (L/2)(\alpha + \beta)$.

The above function is minimized by minimizing ($\alpha + \beta$). The formal solution for choosing α and β is beyond the scope of this book. However, suppose we choose the cutoff value at the midpoint between the two competing hypotheses, (0.25 + 0.50)/2. For a sample size of 100 the cutoff value is 37.5 [(0.25 + 0.50)/2 * 100]. As the sample size grows to a larger number both α and β decrease and in the limit approach zero. Rather than hold α fixed, the Bayesian approach leads to a reduction in both α and β.

We now answer question three. What is the role of the sample size in reducing the probability of making a mistake? For Classical inference a larger sample size leads to a reduction only in the Type II error. The probability of a Type I error remains fixed and eventually becomes the more likely mistake relative to the Type II error. For Bayesian inference both α and β are reduced as the sample size grows larger.

The remaining material in this chapter covers the various types of hypothesis testing. For example, one can specify a ***"point"*** hypothesis or a ***"composite"*** hypothesis. The hypothesis can imply a one-tail or two-tail test.

3.4 Point Null Hypothesis vs. Composite Alternative Hypothesis

In the preceding example both H_0, and H_1 were specified as a point hypothesis. Either H_0 was true or H_1 was true. No other values of π were permitted to generate the data.

An alternative hypothesis would be to state that π is greater than 0.25. In this case H_1 is a composite hypothesis. Rejection of H_0 implies the acceptance of a set of values greater than 0.25 rather than a specific value. Since the calculation of β depends on the underlying distribution generating the data, a composite hypothesis will result in β being a function of π. That is, β is calculated for each value of π greater than 0.25.

In this next problem we will specify H_0 as a point hypothesis and specify H_1 as a composite hypothesis. The Type I error will be specified while the value of β will not be calculated.

3.5 The Pizza Problem

While visiting a local pizza parlor you order a large pizza. The menu claims that the diameter of a large pizza is 15 inches. Being a member in good standing of the local consumer activist group, ***"The False Traders"***, you measure the diameter of the pizza and determine that it is less than 15 inches. You approach the manager and claim that since the pizza is less than 15 inches in diameter, the claim on the menu is false.

The manager responds by claiming that the machine is set to cut the pizzas at a 15-inch diameter. Unfortunately, due to factors beyond his control, some of the pizzas are cut less than 15 inches and some are cut more than 15 inches in diameter. The manager then directs you to read the fine print below the description of a large pizza. The fine print reveals that the diameters of large pizzas are normally distributed with a mean of 15 inches and a standard deviation of 1 inch.

You still believe the machine is cutting the pizzas at a diameter less than 15 inches and propose a test. From a sample of four large pizzas, the sample mean of the diameters will be calculated and a decision will be make regarding the null hypothesis which is H_0: $\mu = 15$. The Type I error will be set at 0.05. The above experiment is formally described below:

Hypothesis:

H_0: $\mu = 15$
H_1: $\mu \leq 15$
$\alpha = P(\text{Type I error}) = 0.05$

Sampling Distribution for H_0:

$$\bar{X} \sim N(\mu = 15, \sigma^2(\bar{X}) = 1/n)$$

Decision Rule:

Reject H_0 if $\bar{X} \leq C^*$,
Otherwise accept H_0.

The null hypothesis is a specified as a point hypothesis. It could be specified as a composite hypothesis ($\mu \geq 15$) but you are less concerned about whether or not the manager is making the pizzas too large. Furthermore, if μ is greater than 15 inches, the alternative hypothesis is less likely to be true. Finally, the manager has little incentive in terms of profit maximization to advertise large pizzas at 15 inches and then make them larger.

In order to perform the hypothesis test we need to calculate $\overline{X}$ and compare it to C*. The cutoff value C* is calculated from setting the Type I error equal to 5%. Figure 3.5.1 shows a picture of how C* is determined when the Type I error is set to 0.05. Lowering the Type I error will lower C*. A higher value for the Type I error will imply a higher value for C*. At the 5% level the cutoff value is set 1.64 standard deviations below the mean. Since the standard deviation is equal to 1/2 ($\sigma(\overline{X}) = 1/2$), moving 1.64 standard deviations below the mean, leads to setting C* at 14.18 inches. The calculation is shown below:

$$\alpha = 0.05 = P(\text{reject } H_0 / H_0 \text{ true})$$

$$= P(\overline{X} < C^* \backslash \mu=15, \sigma(\overline{X}) = 1/2)$$

$$= P(\,(\overline{X} - \mu)/\sigma(\overline{X}) < (C^*-\mu)/\sigma(\overline{X}))$$

$$= P(Z < (C^*-\mu)/\sigma(\overline{X}))$$

$$= P(Z < -1.64)$$

The last line reveals the z-value that will make α equal to 0.05. From the last two expressions above we have:

$$-1.64 = (C^* - 15)/0.5$$

or

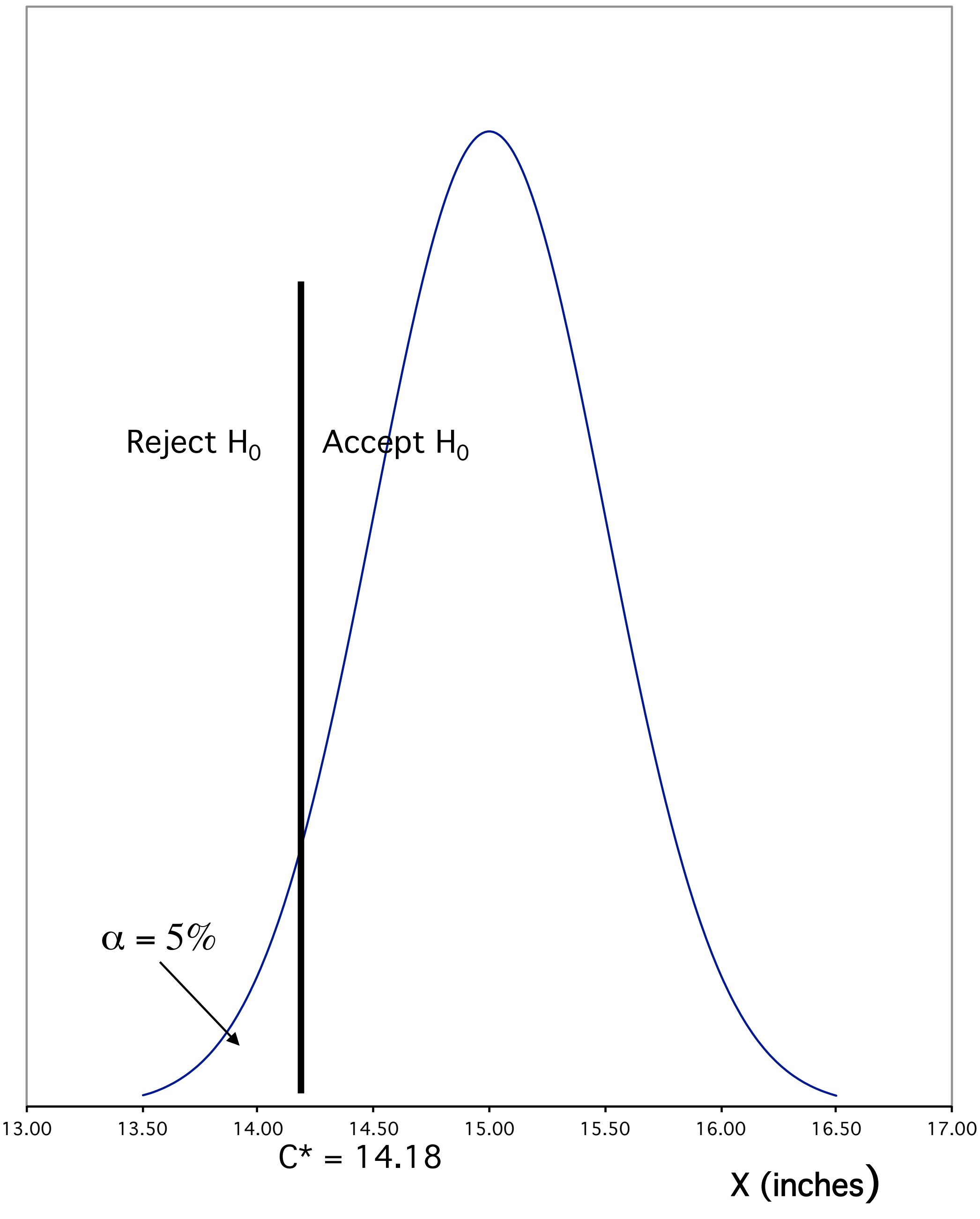
$$\begin{aligned} C^* &= 15 + 1/2(-1.64) \\ &= 15 + -0.82 \\ &= 14.18 \end{aligned}$$

The decision rule can now be written as follows:

Decision Rule:

Reject H_0 if $\overline{X} < 14.18$,
Otherwise accept H_0.

Figure 3.5.1



The experiment is performed and $\overline{X}$ is equal 14.25 inches. The null hypothesis is accepted. An important point to note here is that acceptance of H_0 does not imply, or prove that H_0 is true[1]. Unless one is able to set α and β equal to zero there is always the chance of accepting the wrong hypothesis.

Since we assumed normality in the previous problem there is no way to set the cutoff value so that α equal zero. However, with the normal distribution assumption, sample means outside of three standard deviations from the mean are very unlikely.

The decision rule in the pizza problem was specified in terms of inches, the unit of measurement. The null hypothesis was rejected if $\overline{X}$ was less than 14.18 inches. Another way to state the decision rule is in terms of the standard normal distribution. Since 14.18 inches is 1.64 standard deviations below the population mean, the cutoff value for the standard normal distribution is -1.64.

We can rewrite the decision rule in the following way:

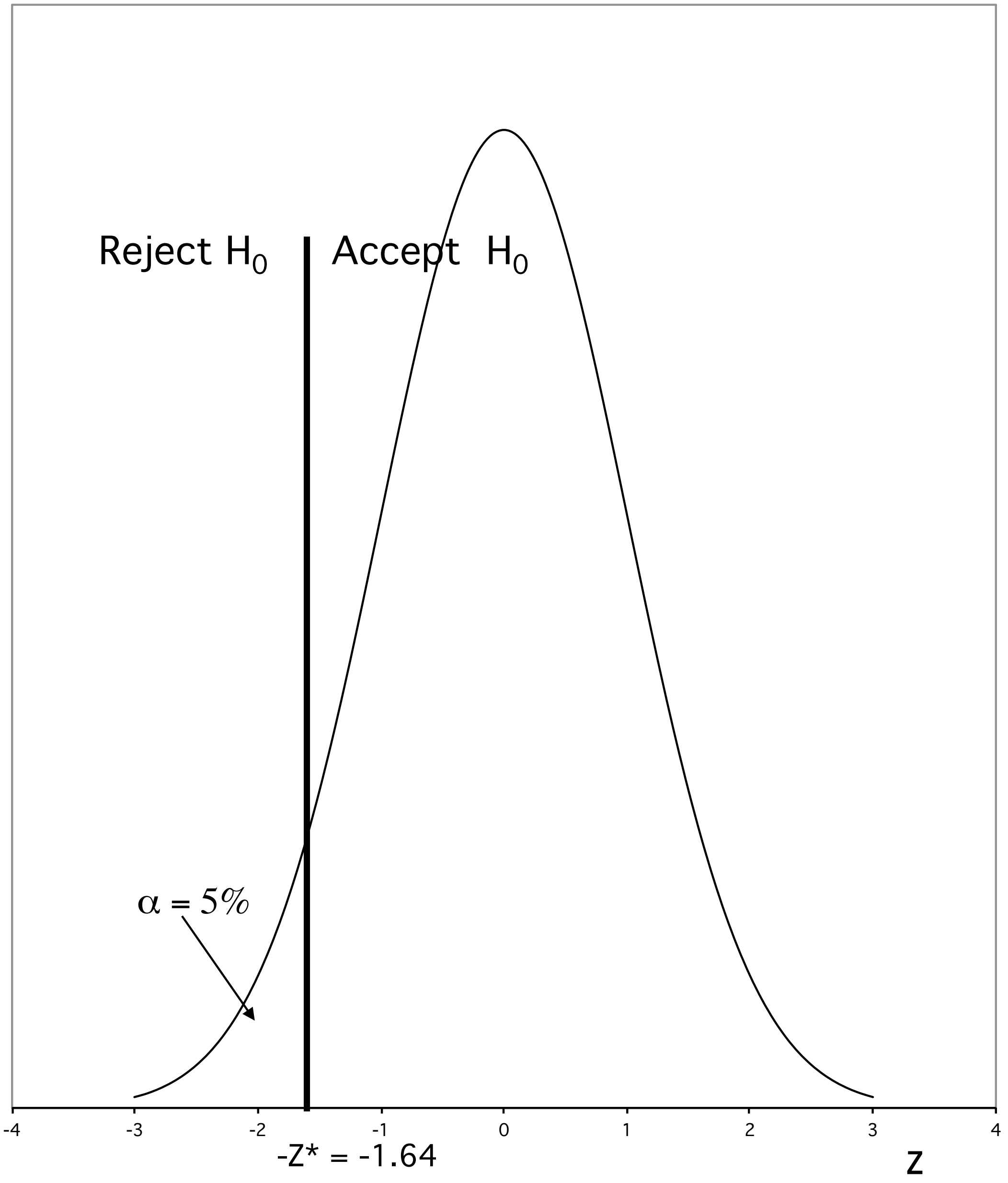
Reject H_0 if $Z_S < -1.64$,
Otherwise accept H_0,

where $Z_S = (\overline{X} - 15)/0.5$

The variable Z_S represents the sample Z-value based on standardizing the sample mean. Figure 3.5.2 shows the picture of the decision along with the probability distribution under H_0.

[1] Rather than state the null hypothesis is accepted, most textbooks prefer to state, "we fail to reject the null hypothesis". Similarly, rejection of the null hypothesis can be stated as "a failure to accept the null hypothesis". I find this distinction amusing. Try this on a loved one. Rather than say "I accept the hypothesis that you love me", state "I fail to reject the hypothesis that you love me", just in case you believe in a small type one error.

Figure 3.5.2



3.6 Types of Hypothesis Tests

The pizza problem tested a point null hypothesis against a composite alternative hypothesis. Since H_1 specified alternative values that were lower, the cutoff value was set below the mean assumed under H_0. The area for α was distributed in the lower-tail of the normal distribution. This type of test is referred to as a lower-tail test. Similarly, an alternative hypothesis can be specified to set up an upper-tail test or a two-tail test. The general forms for these three types of tests are shown below.

Case 1: **Lower-tail test:** **H_0: $\mu = \mu_0$**
H_1: $\mu < \mu_0$

Decision Rule: **Reject H_0 if $Z_S < -Z^*$,**
Accept H_0 if $Z_S \geq -Z^*$.

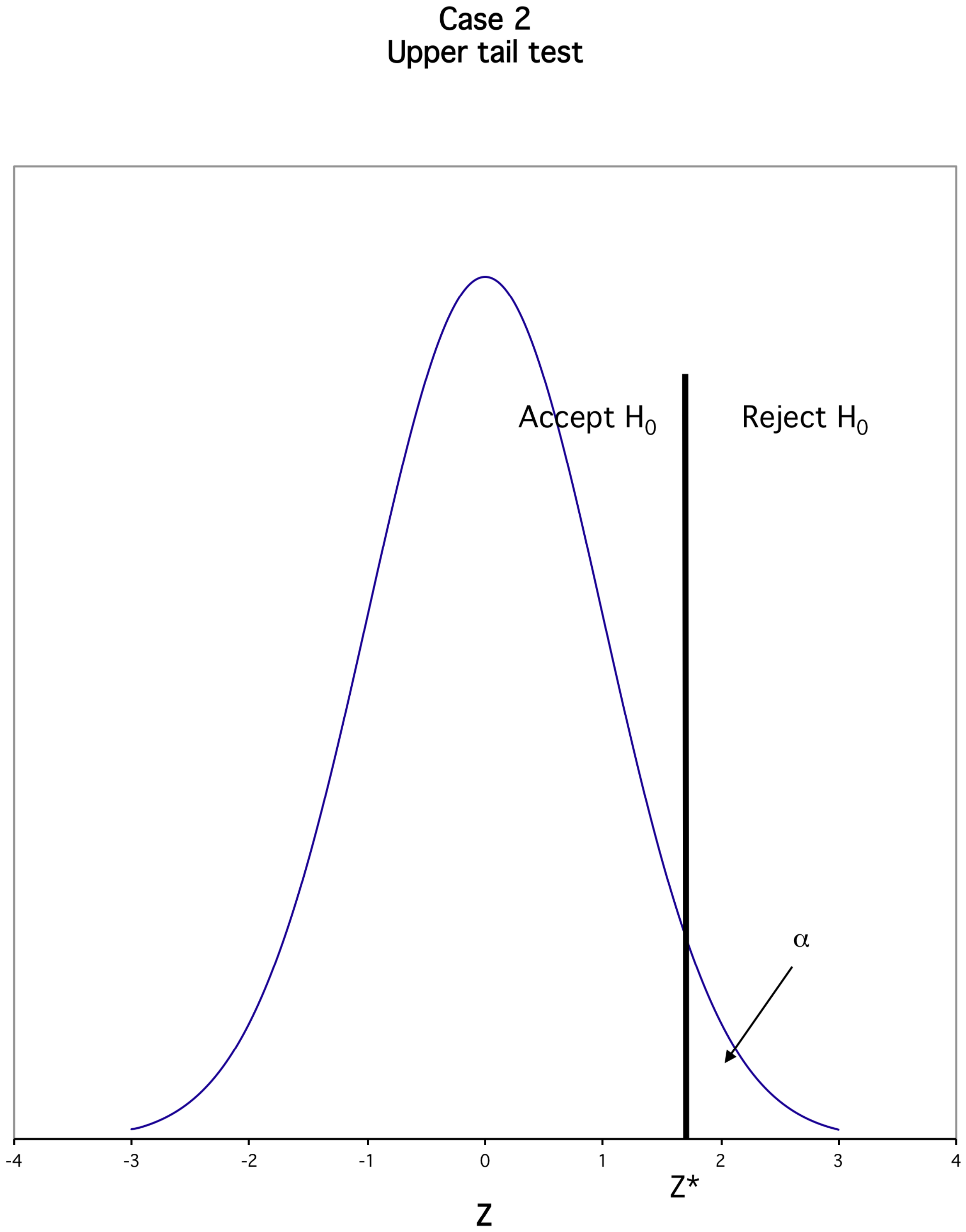
Case 2: **Upper-tail test:** **H_0: $\mu = \mu_0$**
H_1: $\mu > \mu_0$

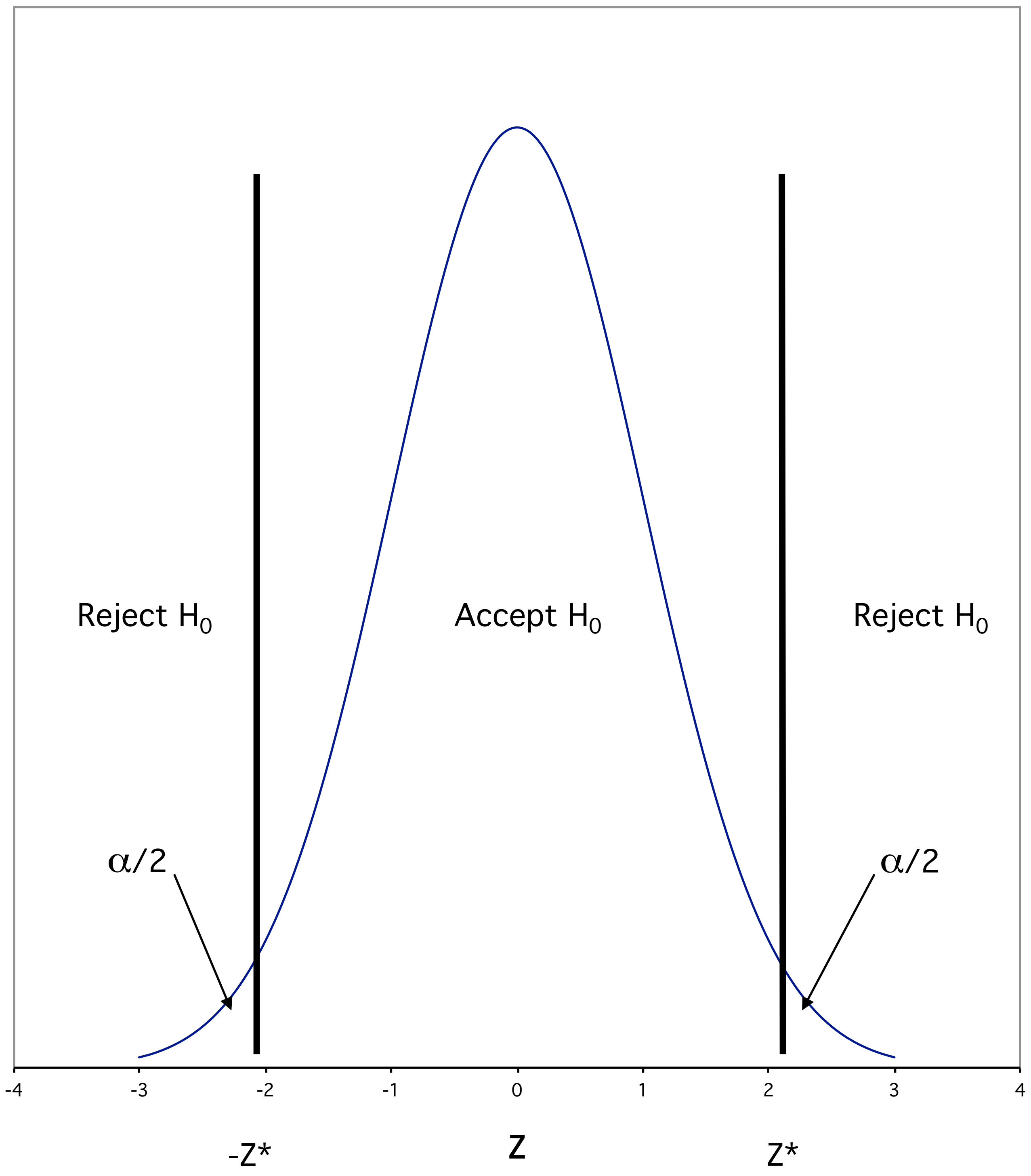
Decision Rule: **Reject H_0 if $Z_S > Z^*$**
Accept H_0 if $Z_S \leq Z^*$

Case 3: **Two-tail test:** **H_0: $\mu = \mu_0$**
H_1: $\mu \neq \mu_0$

Decision Rule: **Reject H_0 if $|Z_S| > Z^*$**
Accept H_0 if $|Z_S| \leq Z^*$

Figure 3.6.1 shows the picture for each case. The lower-tail and upper-tail tests are similar. The cutoff value Z* is determined by allocating the Type I error to the lower or upper portion of the standard normal distribution. The two-tail test is similar to constructing a confidence interval. The difference being that the interval for the decision rule is centered around the population mean as opposed to the sample mean. An extremely large or small value for Z_S will result in H_0 being rejected. Note that the Type I error is split equally on both sides of the distribution.


Case 2
Upper tail test
Accept H_0
Reject H_0
α
-4
-3
-2
-1
0
1
2
3
4
Z*
Z



Case 3
Two tail test

APPENDIX

Below is the derivation of expected costs from using the Bayesian approach to take into account differences in the probabilities and costs of mistakenly accepting the wrong hypothesis.

Definitions

E(C) = expected costs
L_1 = loss from Type I error
L_2 = loss from Type II error

$E(C) = L_1 * P(H_0$ true **and** reject $H_0) + L_2 * P(H_1$ true **and** reject $H_1)$

$= L_1 * P(H_0) * P(\text{reject } H_0 / H_0) + L_2 * P(H_1) * P(\text{reject } H_1 / H_1)$

$= \alpha * L_1 * P((H_0) + \beta * L_1 * P((H_1)$

Suppose $L_1 = L_2 = L$ and $P(H_0) = P(H_1) = 1/2$. We can then reduce the expected cost expression to

$$\mathbf{E(C) = (L/2) * (\alpha + \beta)}$$

which implies that E(C) is minimized when $(\alpha + \beta)$ is minimized.

<u>Numerical Example Using Table 2 Cutoff Values</u>

Let L1 = 10, L2 = 40, and $P(H_0) = P(H_1) = 1/2$

$E(C) \quad = \alpha(10)(1/2) + \beta(40)(1/2)$

$= 5\alpha + 20\beta$

Cutoff Value (X*)	α	β	E(C)	$(\alpha + \beta)$
0	.7627	.0312	4.4375 .7939	
1	.3672	.1874	5.58	.5546
2	.1035	.4999	10.5155	.6034
3	.0156	.8124	16.326 .8280	
4	.0010	.9686	19.377 .9696	
5	0.0	1.0	20.0	1.0