Probability and Statistics

Kristel Van Steen, PhD^2

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

CHAPTER 6: HYPOTHESIS TESTING

1 Terminology and Notation

1.1 Tests of Hypotheses

1.2 Size and Power of Tests

1.3 Examples

2 One-sided and Two-sided Tests

2.1 Introduction

2.2 Case(a) Alternative is one-sided

2.3 Case (b) Two-sided Alternative

2.4 Two Approaches to Hypothesis Testing

3 Connection between Hypothesis testing and CI’s

3.1 Two faces of the same coin

3.2 The concept of a p-value

3.3 Three approaches for hypothesis testing

4 One-sample problems

4.1 Testing hypotheses about σ^2

4.3 Testing hypotheses about μ

5 Two-Sample Problems

5.1 Testing equality of normal means

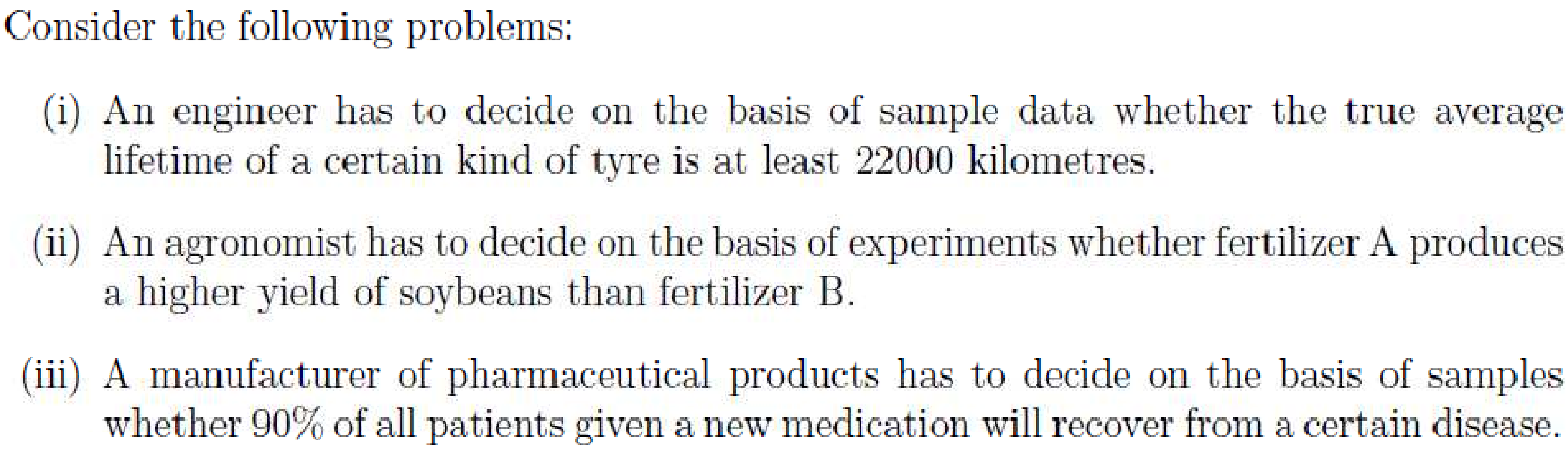
5.2 Testing equality of sample variances

6 Course concluding remarks

1 Terminology and Notation

1.1 Tests of Hypotheses

Introduction



Consider the following problems:

(i) An engineer has to decide on the basis of sample data whether the true average lifetime of a certain kind of tyre is at least 22000 kilometres.

(ii) An agronomist has to decide on the basis of experiments whether fertilizer A produces a higher yield of soybeans than fertilizer B.

(iii) A manufacturer of pharmaceutical products has to decide on the basis of samples whether 90% of all patients given a new medication will recover from a certain disease.

These problems can be translated into the language of **statistical tests of hypotheses.**

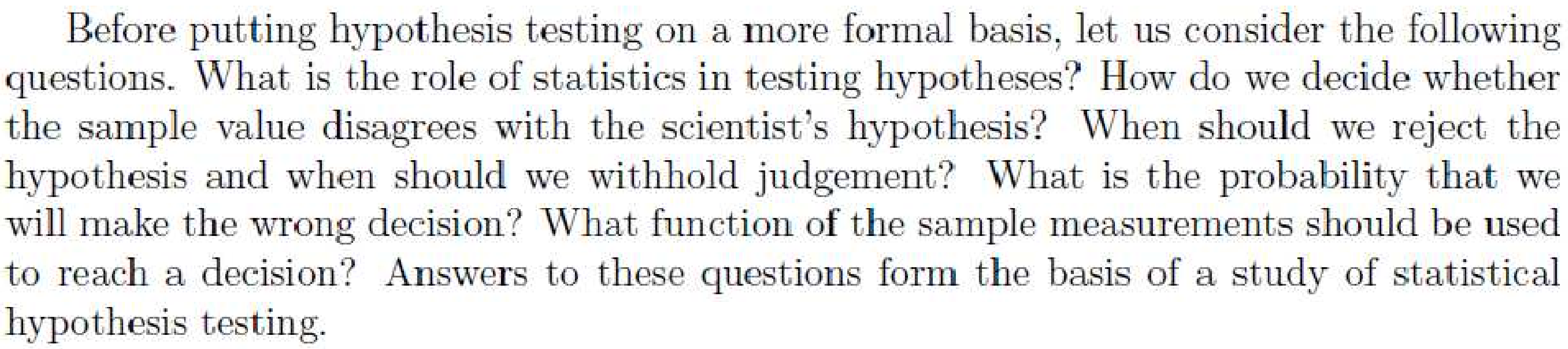
(i) The engineer has to test the assertion that if the lifetime of the tyre has pdf. $f(x) = \alpha e^{-\alpha x},\ x > 0$, then the expected lifetime, $1/\alpha$, is at least 22000.

(ii) The agronomist has to decide whether $\mu_A > \mu_B$ where μ_A, μ_B are the means of 2 normal distributions.

(iii) The manufacturer has to decide whether p, the parameter of a binomial distribution is equal to .9.

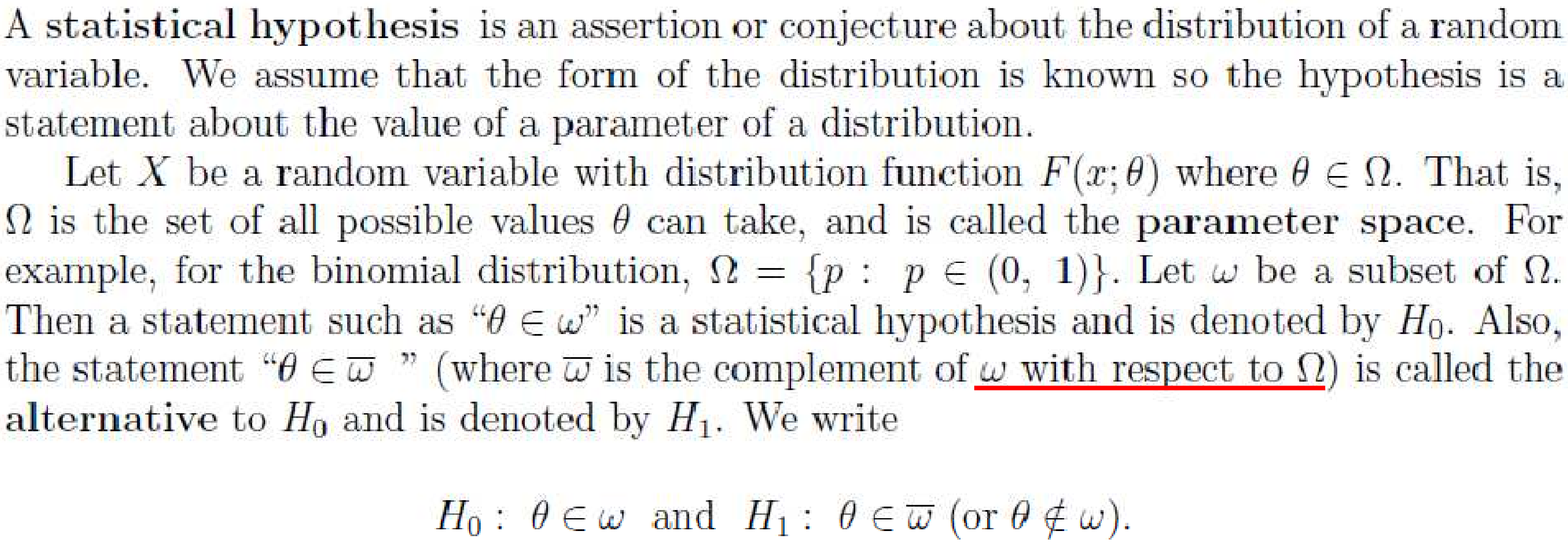
In each case, it is assumed that the stated distribution correctly describes the experimental conditions, and that the hypothesis concerns the **parameter(s)** of that distribution. [A more general kind of hypothesis testing problem is where the **form** of the distribution is unknown.]

In many ways, the formal procedure for hypothesis testing is similar to the scientific method. The scientist formulates a theory, and then tests this theory against observation. In our context, the scientist poses a theory concerning the value of a parameter. He then samples the population and compares observation with theory. If the observations disagree strongly enough with the theory the scientist would probably reject his hypothesis. If not,



Before putting hypothesis testing on a more formal basis, let us consider the following questions. What is the role of statistics in testing hypotheses? How do we decide whether the sample value disagrees with the scientist's hypothesis? When should we reject the hypothesis and when should we withhold judgement? What is the probability that we will make the wrong decision? What function of the sample measurements should be used to reach a decision? Answers to these questions form the basis of a study of statistical hypothesis testing.

Terminology and notations



A **statistical hypothesis** is an assertion or conjecture about the distribution of a random variable. We assume that the form of the distribution is known so the hypothesis is a statement about the value of a parameter of a distribution.

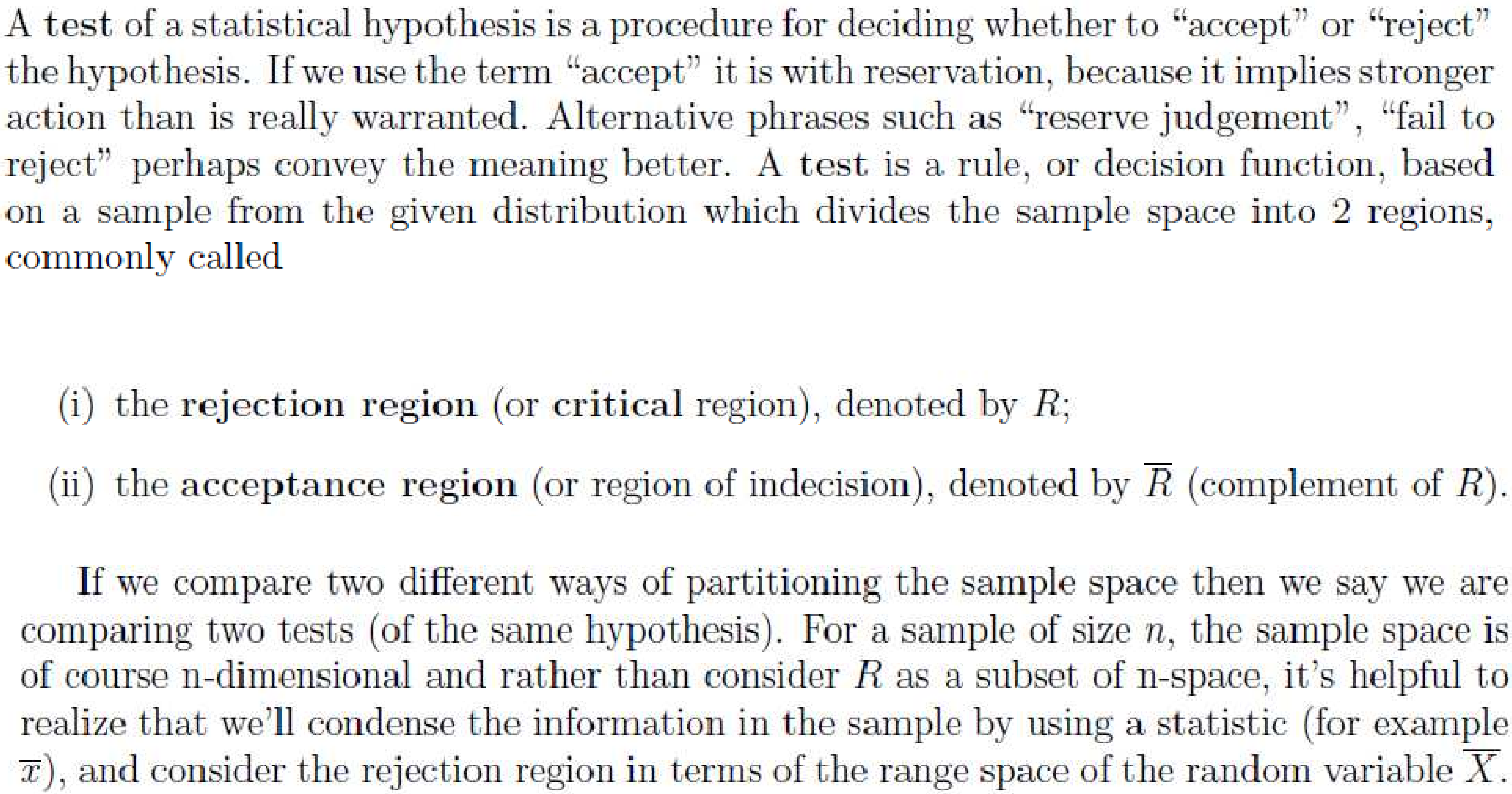
Let X be a random variable with distribution function $F(x;\theta)$ where $\theta \in \Omega$. That is, Ω is the set of all possible values θ can take, and is called the **parameter space**. For example, for the binomial distribution, $\Omega = \{p : \ p \in (0,\ 1)\}$. Let ω be a subset of Ω. Then a statement such as "$\theta \in \omega$" is a statistical hypothesis and is denoted by H_0. Also, the statement "$\theta \in \overline{\omega}$ " (where $\overline{\omega}$ is the complement of ω with respect to Ω) is called the **alternative** to H_0 and is denoted by H_1. We write

$$H_0 : \ \theta \in \omega \quad \text{and} \quad H_1 : \ \theta \in \overline{\omega} \ (\text{or } \theta \notin \omega).$$

Often hypotheses arise in the form of a claim that a new product, technique, etc. is better than the existing one. In this context, H is a statement that nullifies the claim (or represents the *status quo*) and is sometimes called **a null hypothesis**, but we will refer to it as **the hypothesis**.

If ω contains only one point, that is, if $\omega = \{\theta : \ \theta = \theta_0\}$ then H_0 is called a **simple hypothesis**. We may write $H_0 : \ \theta = \theta_0$. Otherwise it is called **composite**. The same applies to alternatives.

Tests of hypotheses



A **test** of a statistical hypothesis is a procedure for deciding whether to "accept" or "reject" the hypothesis. If we use the term "accept" it is with reservation, because it implies stronger action than is really warranted. Alternative phrases such as "reserve judgement", "fail to reject" perhaps convey the meaning better. A **test** is a rule, or decision function, based on a sample from the given distribution which divides the sample space into 2 regions, commonly called

(i) the **rejection region** (or **critical** region), denoted by R;

(ii) the **acceptance region** (or region of indecision), denoted by $\overline{R}$ (complement of R).

If we compare two different ways of partitioning the sample space then we say we are comparing two tests (of the same hypothesis). For a sample of size n, the sample space is of course n-dimensional and rather than consider R as a subset of n-space, it's helpful to realize that we'll condense the information in the sample by using a statistic (for example $\overline{x}$), and consider the rejection region in terms of the range space of the random variable $\overline{X}$.

1.2 Size and Power of Tests

There are two types of errors that can occur. If we reject H when it is true, we commit a **Type I** error. If we fail to reject H when it is false, we commit a **Type II** error. You may like to think of this in tabular form.

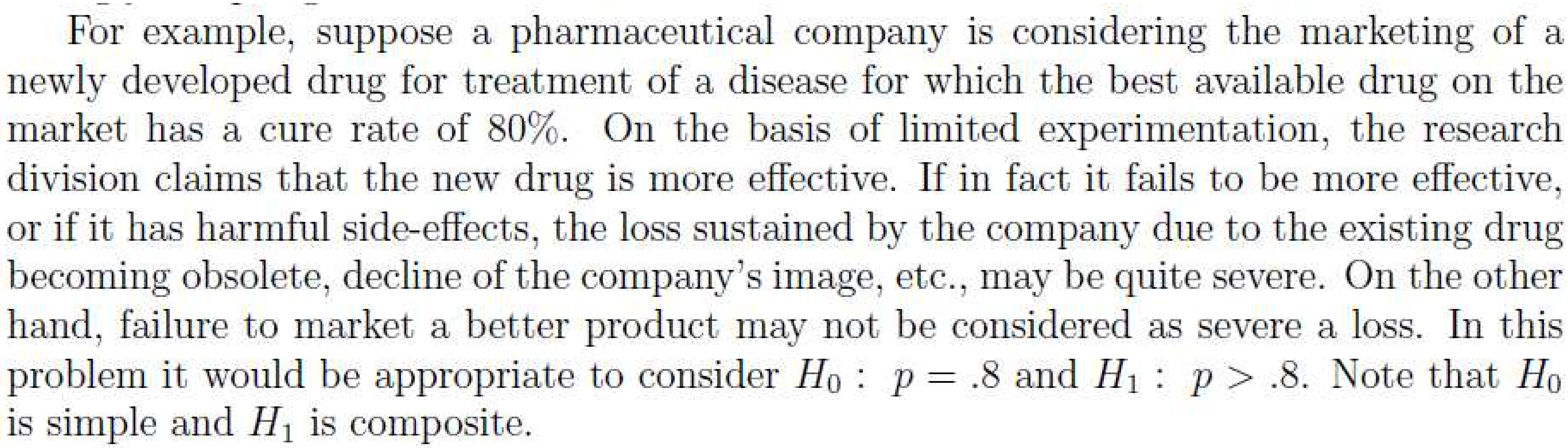
		Our decision	
		do not reject H_0	reject H_0
Actual	H_0 is true	correct decision	Type I error
situation	H_0 is not true	Type II error	correct decision

Probabilities associated with the two incorrect decisions are denoted by

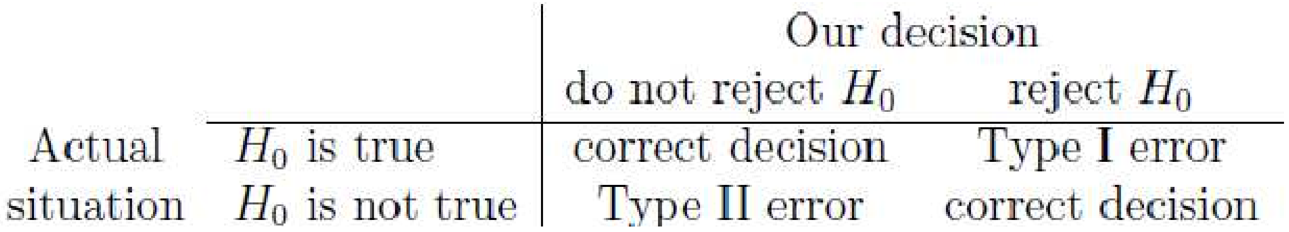
$$\begin{aligned}\alpha &= P(H_0 \text{ is rejected when it is true}) = \mathrm{P}(\text{Type I error})\\ \beta &= P(H_0 \text{ is not rejected when it is false}) = \mathrm{P}(\text{Type II error})\end{aligned}$$

The probability α is sometimes referred to as the **size** of the critical region or the **significance level** of the test, and the probability $1-\beta$ as the **power** of the test.

Are the roles of the hypothesis and alternative hypothesis symmetric?

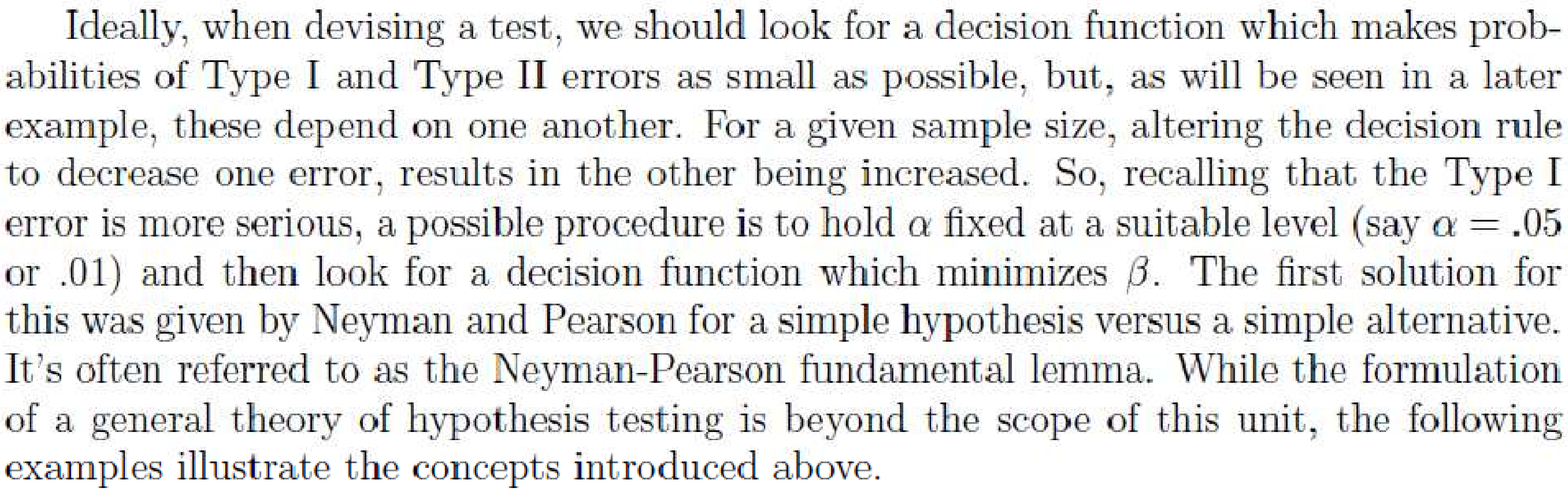


For example, suppose a pharmaceutical company is considering the marketing of a newly developed drug for treatment of a disease for which the best available drug on the market has a cure rate of 80%. On the basis of limited experimentation, the research division claims that the new drug is more effective. If in fact it fails to be more effective, or if it has harmful side-effects, the loss sustained by the company due to the existing drug becoming obsolete, decline of the company's image, etc., may be quite severe. On the other hand, failure to market a better product may not be considered as severe a loss. In this problem it would be appropriate to consider $H_0: \ p = .8$ and $H_1: \ p > .8$. Note that H_0 is simple and H_1 is composite.



		Our decision	
		do not reject H_0	reject H_0
Actual	H_0 is true	correct decision	Type I error
situation	H_0 is not true	Type II error	correct decision

Considerations when constructing a test



Ideally, when devising a test, we should look for a decision function which makes probabilities of Type I and Type II errors as small as possible, but, as will be seen in a later example, these depend on one another. For a given sample size, altering the decision rule to decrease one error, results in the other being increased. So, recalling that the Type I error is more serious, a possible procedure is to hold α fixed at a suitable level (say $\alpha = .05$ or .01) and then look for a decision function which minimizes β. The first solution for this was given by Neyman and Pearson for a simple hypothesis versus a simple alternative. It's often referred to as the Neyman-Pearson fundamental lemma. While the formulation of a general theory of hypothesis testing is beyond the scope of this unit, the following examples illustrate the concepts introduced above.

1.3 Examples

Example .1

Suppose that random variable X has a normal distribution with mean μ and variance 4. Test the hypothesis that $\mu = 1$ against the alternative that $\mu = 2$, based on a sample of size 25.

Solution: An unbiased estimate of μ is $\overline{X}$ and we know that $\overline{X}$ is distributed normally with mean μ and variance σ^2/n which in this example is $4/25$. We note that values of $\overline{x}$ close to 1 support H whereas values of $\overline{x}$ close to 2 support A. We could make up a decision rule as follows:

$$\text{If } \overline{x} > 1.6 \text{ claim that } \mu = 2,$$

$$\text{If } \overline{x} \leq 1.6 \text{ claim that } \mu = 1.$$

The diagram in Figure fig.CRUpperTail shows the sample space of $\overline{x}$ partitioned into

(i) the critical region, R$= \{\overline{x} : \ \overline{x} > 1.6\}$

(ii) the acceptance region, $\overline{R} = \{\overline{x} : \ \overline{x} \leq 1.6\}$

Here, 1.6 is the critical value of $\overline{x}$.

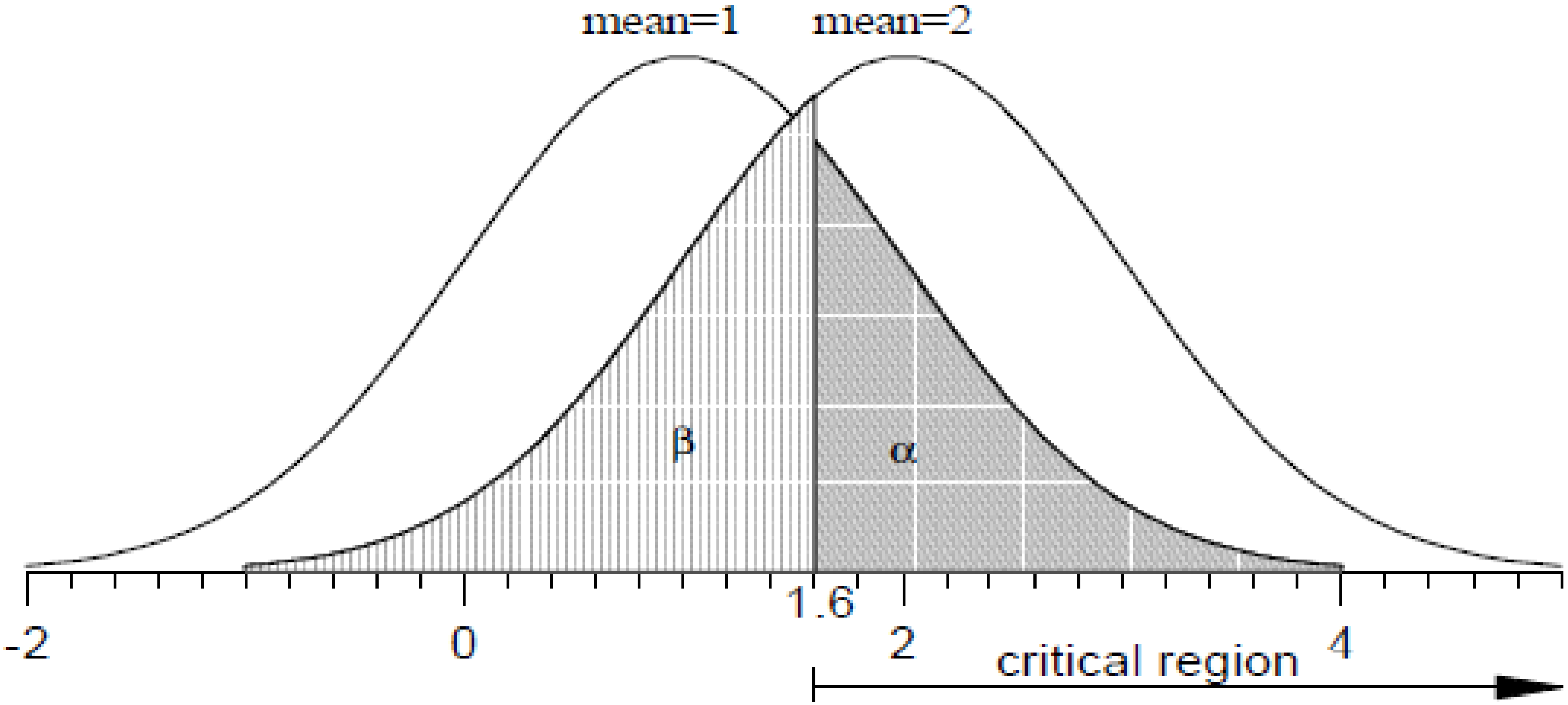


Fig: CRUpperTail

We will find the probability of Type I and Type II error,

$P(\overline{X} > 1.6|\mu = 1, \sigma = \frac{2}{5}) = .0668.$ (`pnorm(q=1.6,mean=1,sd=0.4,lower.tail=F)`)

This is

$$\text{P}(H_0 \text{ is rejected}|H_0 \text{ is true}) = \text{P}(\text{Type I error}) = \alpha$$

Also

$$\begin{aligned}\beta = \text{P}(\text{Type II error}) &= \text{P}(H_0 \text{ is not rejected}|H_0 \text{ is false}) \\ &= P(\overline{X} \leq 1.6|\mu = 2, \sigma = \frac{2}{5}) \\ &= .1587 \qquad (\texttt{pnorm(q=1.6,mean=2,sd=0.4,lower.tail=T)})\end{aligned}$$

To see how the decision rule could be altered so that $\alpha = .05$, let the critical value be c. We require

$$\begin{aligned} P(\overline{X} > c|\mu = 1, \sigma = \frac{2}{5}) &= 0.05 \\ \Rightarrow c &= 1.658 \quad (\texttt{qnorm(p=0.05,mean=1,sd=0.4,lower.tail=T)}) \end{aligned}$$

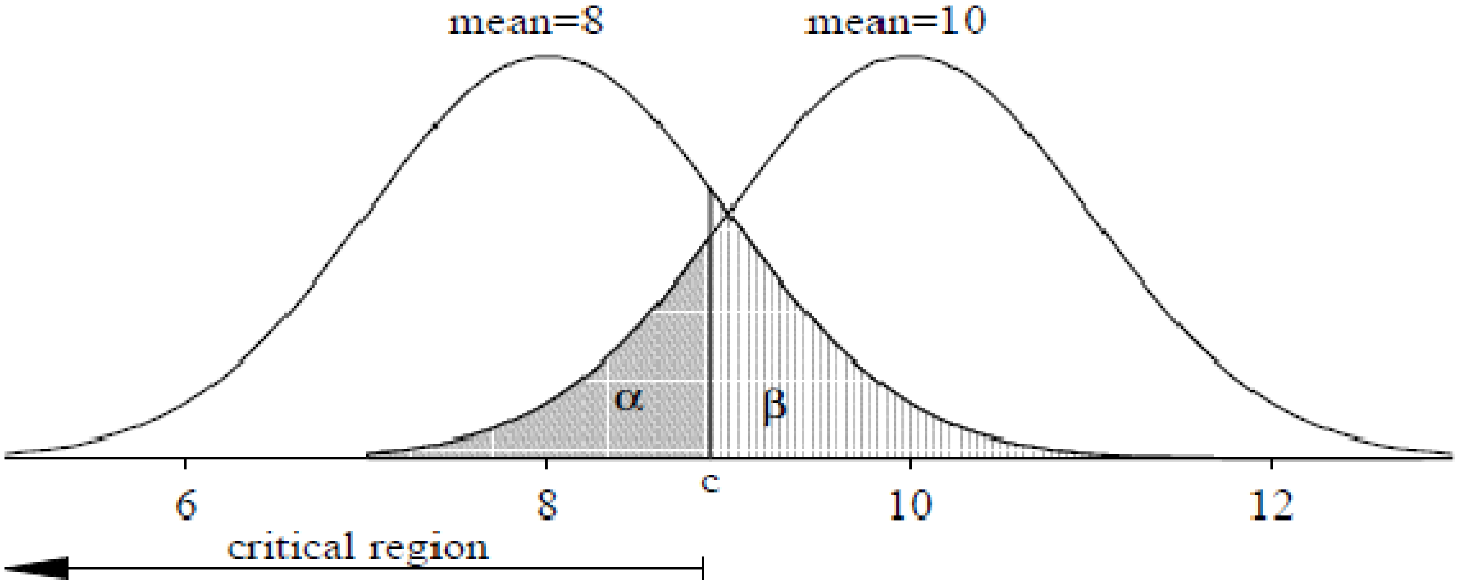
$$P(\overline{X} < c|\mu = 2, \sigma = \frac{2}{5}) = 0.196 \quad (\texttt{pnorm(q=1.658,mean=2,sd=0.4,lower.tail=T)})$$

This value of c gives an α of 0.05 and a β of 0.196 illustrating that as one type of error (α) decreases the other (β) increases.

Example .2

Suppose we have a random sample of size n from a N(μ,4) distribution and wish to test $H_0 : \mu = 10$ against $H_1 : \mu = 8$. The decision rule is to reject H_0 if $\overline{x} < c$. We wish to find n and c so that $\alpha = 0.05$ and $\beta \approx 1$.

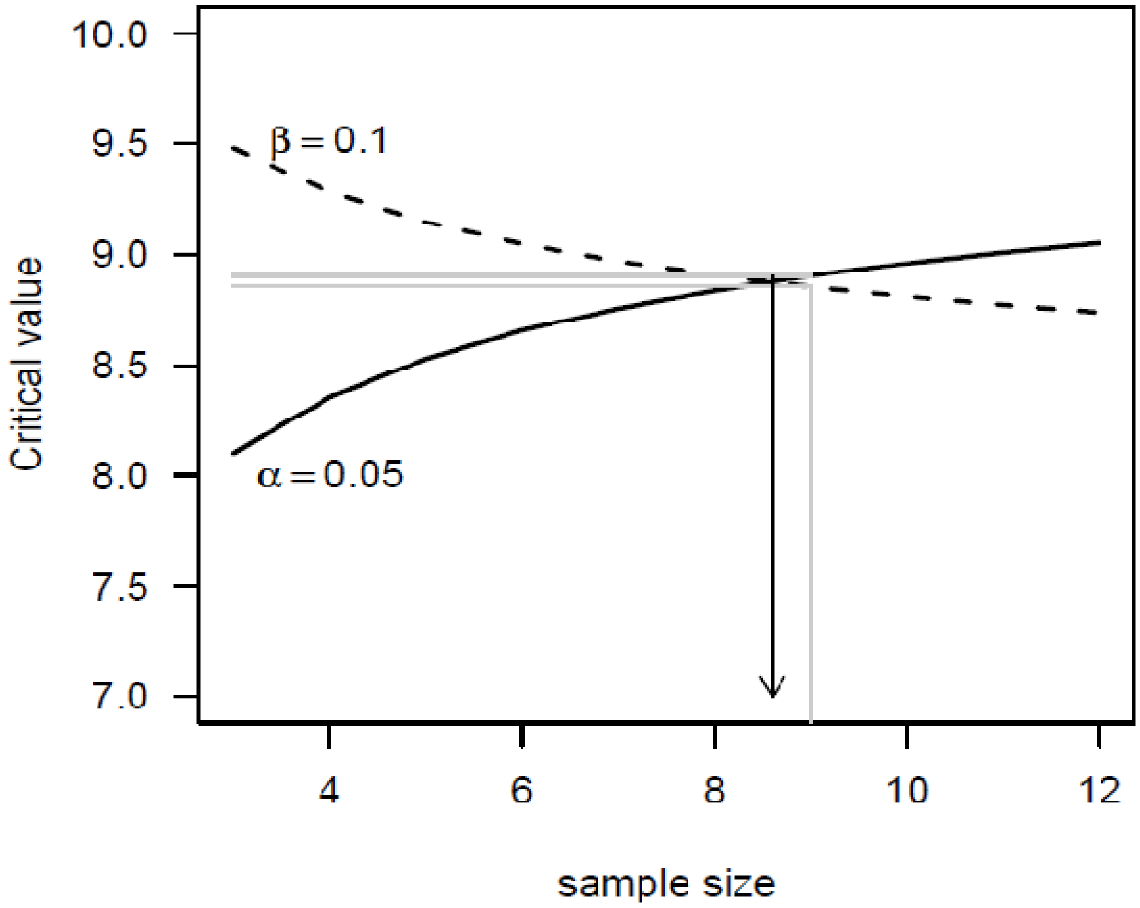
Solution: , the left curve is $f(\overline{x}|H_1)$ and the right curve is $f(\overline{x}|H_0)$. The critical region is $\{\overline{x} : \overline{x} < c\}$, so α is the left shaded area and β is the right shaded area.



Now

$$
\begin{aligned}
\alpha &= 0.05 = P(\overline{X} < c | \mu = 10, \sigma = \frac{2}{\sqrt{n}}) \\
\beta &= 0.1 = P(\overline{X} \geq c | \mu = 8, \sigma = \frac{2}{\sqrt{n}})
\end{aligned}
$$

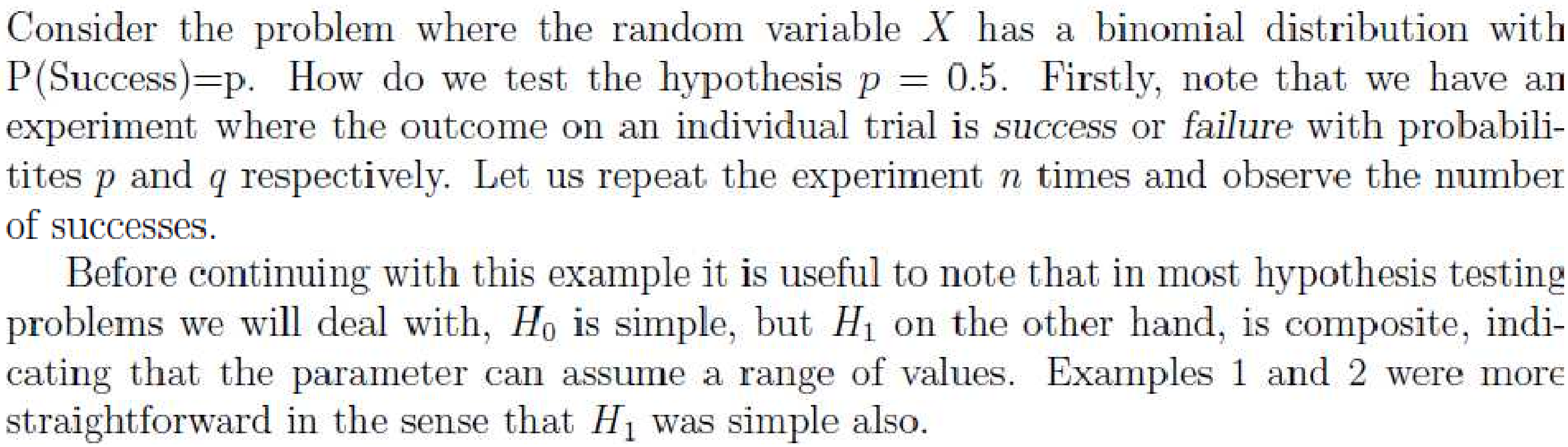
We need to solve these simultaneously for n as shown in Figure



A sample size $n = 9$ and critical value $c = 8.9$ gives $\alpha \approx 0.05$ and $\beta \approx 0.1$.

2 One-sided and Two-sided Tests

2.1 Introduction



Consider the problem where the random variable X has a binomial distribution with P(Success)=p. How do we test the hypothesis $p = 0.5$. Firstly, note that we have an experiment where the outcome on an individual trial is *success* or *failure* with probabilites p and q respectively. Let us repeat the experiment n times and observe the number of successes.

Before continuing with this example it is useful to note that in most hypothesis testing problems we will deal with, H_0 is simple, but H_1 on the other hand, is composite, indicating that the parameter can assume a range of values. Examples 1 and 2 were more straightforward in the sense that H_1 was simple also.

If the range of possible parameter values lies entirely on the one side of the hypothesized value, the aternative is said to be **one-sided**. For example, $H_1 : p > .5$ is one-sided but $H_1 : p \neq .5$ is **two-sided**. In a real-life problem, the decision of whether to make the alternative one-sided or two-sided is not always clear cut. As a general rule-of-thumb, if parameter values in only one direction are physically meaningful, or are the only ones that are possible, the alternative should be one-sided. Otherwise, H_1 should be two-sided. Not all statisticians would agree with this rule.

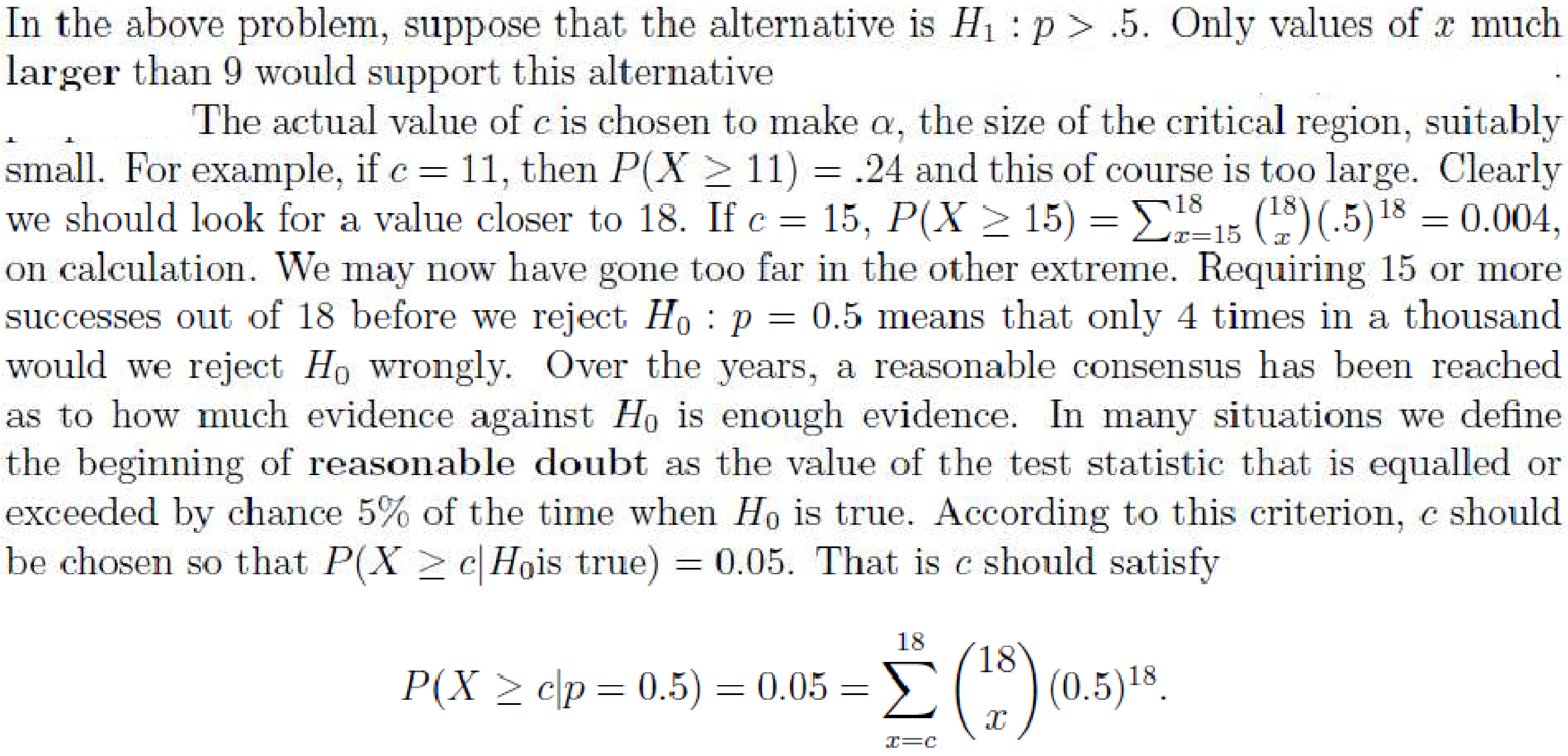
The next question is what test statistic we use to base our decision on.

Recall that, the principle of hypothesis testing is that we will assume H_0 is correct, and our position will change only if the data show **beyond all reasonable doubt** that H_1 is true. The problem then is to define in quantitative terms what reasonable doubt means. Let us suppose that $n = 18$ in our problem above. Then the range space for X is $R_X = \{0, 1, \ldots, 18\}$ and E(X)=np= 9 **if** H_0 is true. If the observed number of successes is close to 9 we would be obliged to think that H was true. On the other hand, if the observed value of X was 0 or 18 we would be fairly sure that H_0 was not true. Now **reasonable doubt** does not have to be as extreme as 18 cases out of 18. Somewhere between x-values of 9 and 18 (or 9 and 0), there is a point, c say, when for all practical purposes the credulity of H_0 ends and reasonable doubt begins. This point is called the **critical value** and it completely determines the decision-making process. We could make up a decision rule

$$\begin{array}{llll} \text{If } x & \geq & c, & \text{reject } H_0 \\ \text{If } x & < & c, & \text{conclude that } H_0 \text{ is probably correct.} \end{array}$$

In this case, $\{x : x \geq c\}$ is the rejection region, R referred to in §2.2.

2.2 Case(a) Alternative is one-sided



In the above problem, suppose that the alternative is $H_1 : p > .5$. Only values of x much larger than 9 would support this alternative

The actual value of c is chosen to make α, the size of the critical region, suitably small. For example, if $c = 11$, then $P(X \geq 11) = .24$ and this of course is too large. Clearly we should look for a value closer to 18. If $c = 15$, $P(X \geq 15) = \sum_{x=15}^{18} \binom{18}{x}(.5)^{18} = 0.004$, on calculation. We may now have gone too far in the other extreme. Requiring 15 or more successes out of 18 before we reject $H_0 : p = 0.5$ means that only 4 times in a thousand would we reject H_0 wrongly. Over the years, a reasonable consensus has been reached as to how much evidence against H_0 is enough evidence. In many situations we define the beginning of **reasonable doubt** as the value of the test statistic that is equalled or exceeded by chance 5% of the time when H_0 is true. According to this criterion, c should be chosen so that $P(X \geq c|H_0\text{is true}) = 0.05$. That is c should satisfy

$$P(X \geq c|p = 0.5) = 0.05 = \sum_{x=c}^{18} \binom{18}{x} (0.5)^{18}.$$

A little trial and error shows that $c = 13$ is the appropriate value. Of course because of the discrete nature of X it will not be possible to obtain an α of exactly 0.05.

Defining the critical region in terms of the x-value that is exceeded only 5% of the time when H_0 is true is the most common way to quantify reasonable doubt, but there are others. The figure 1% is frequently used and if the critical value is exceeded only 1% of the time we say there is **strong evidence** against H_0. If the critical value is only exceeded .1% of the time we may say that there is **very strong evidence** against H_0.

So far we have considered a one-sided alternative. Now we'll consider the other case where the alternative is two-sided.

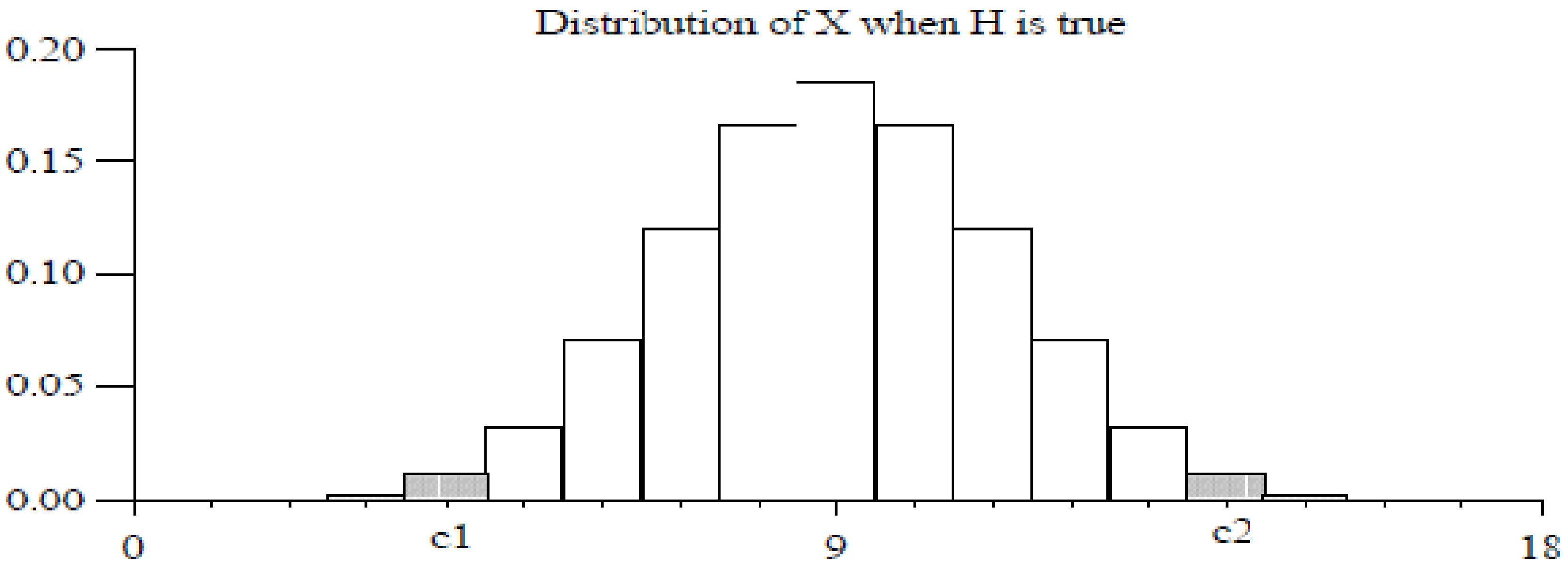
2.3 Case (b) Two-sided Alternative

Consider now the alternative $H_1 : \ p \neq 0.5$. Values of x too large or too small would support this alternative. In this case there are two critical regions (or more correctly, *the* critical region consists of two disjoint sets), one in each 'tail' of the distribution of X. For a 5% critical region, there would be two critical values c_1 and c_2 such that

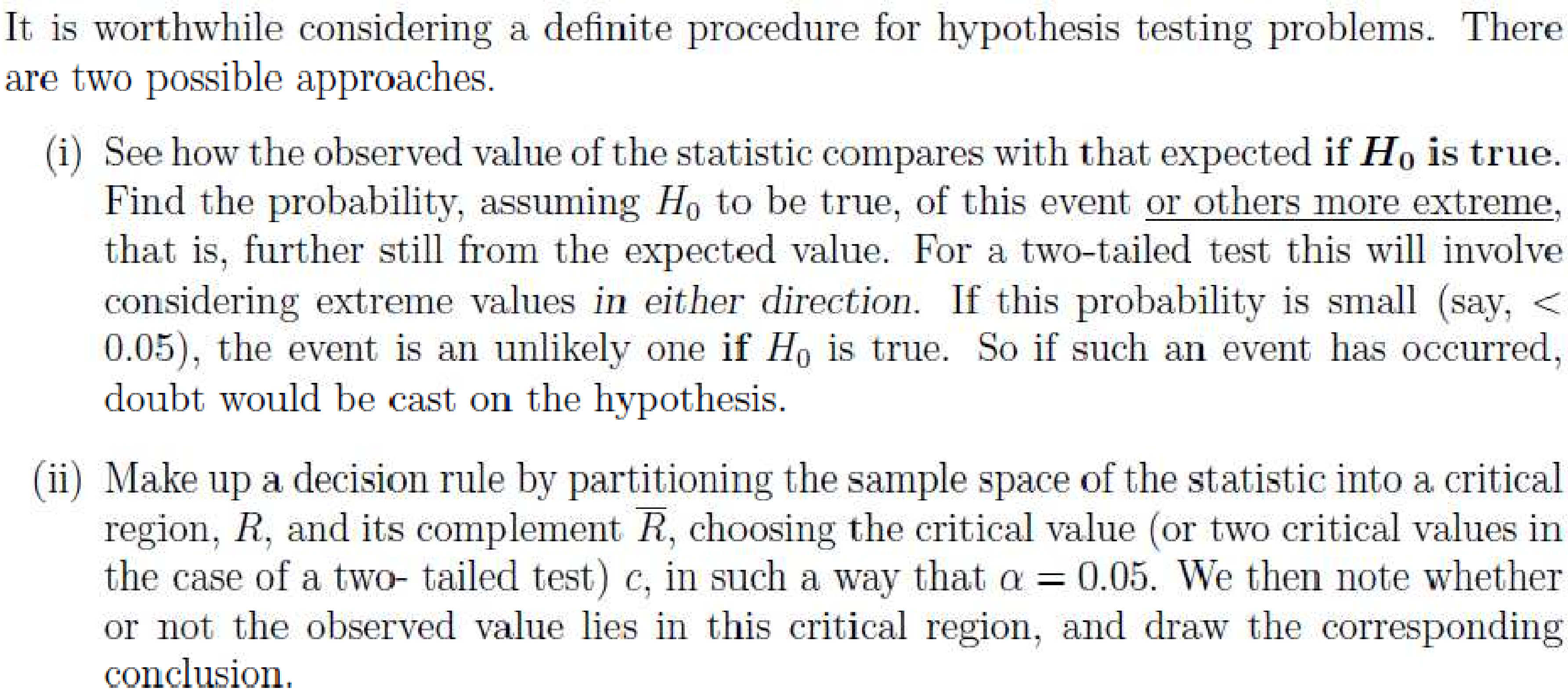
$$P(X \leq c_1 | H_0 \text{ is true}) \approx 0.025 \text{ and } P(X \geq c_2 | H_0 \text{ is true}) \approx 0.025.$$

This can be seen in Figure below, where the graph is of the distribution of X when H_0 is true. (It can be shown that $c_1 = 4$ and $c_2 = 14$ are the critical values in this case.)

Tests with a one-sided critical region are called **one-tailed tests**, whereas those with a two-sided critical region are called **two-tailed tests**.


Distribution of X when H is true
0.20
0.15
0.10
0.05
0.00
0
c1
9
c2
18

2.4 Two Approaches to Hypothesis Testing



It is worthwhile considering a definite procedure for hypothesis testing problems. There are two possible approaches.

(i) See how the observed value of the statistic compares with that expected **if $\boldsymbol{H_0}$ is true**. Find the probability, assuming H_0 to be true, of this event or others more extreme, that is, further still from the expected value. For a two-tailed test this will involve considering extreme values *in either direction*. If this probability is small (say, $<$ 0.05), the event is an unlikely one **if** H_0 is true. So if such an event has occurred, doubt would be cast on the hypothesis.

(ii) Make up a decision rule by partitioning the sample space of the statistic into a critical region, R, and its complement $\overline{R}$, choosing the critical value (or two critical values in the case of a two- tailed test) c, in such a way that $\alpha = 0.05$. We then note whether or not the observed value lies in this critical region, and draw the corresponding conclusion.

Example 3

Suppose we want to know whether a given die is biased towards 5 or 6 or whether it is "true". To examine this problem the die is tossed 9000 times and it is observed that on 3164 occasions the outcome was 5 or 6.

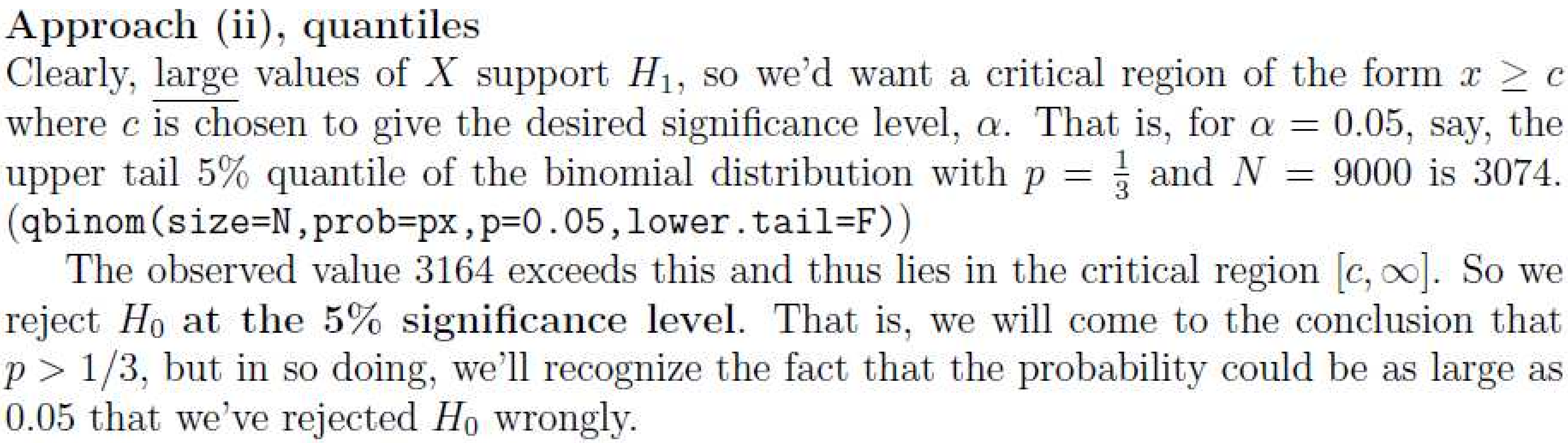
Solution: Let X be the number of successes (5's or 6's) in 9000 trials. Then if $p = P(S)$, X is distributed bin(9000,p). As is usual in hypothesis testing problems, we set up H_0 as the hypothesis we wish to "disprove". In this case, it is that the die is "true", that is, $p = 1/3$. If H_0 is not true, the alternative we wish to claim is that the die is biased towards 5 or 6, that is $p > 1/3$. In practice, one decides on this alternative before the experiment is carried out. We will consider the 2 approaches mentioned above.

Approach (i), probabilities

If $p = 1/3$ and $N = 9000$ then $E(X) = np = 3000$ and $\text{Var}(X) = npq = 2000$. The observed number of successes, 3164, was greater than expected if H_0 were true. So, assuming $p = 1/3$, the probability of the observed event together with others more extreme (that is, further still from expectation) is

$$P_B(X \geq 3164 | p = 1/3) = 0.0001 \quad \text{(pbinom(q=3164,size=9000,prob=1/3,lower.tail=F)}$$

This probability is small, so the event $X \geq 3164$ is an unlikely one if the assumption we've made ($p = 1/3$) is correct. Only about 1 times in 10000 would we expect such an occurrence. Hence, if such an event did occur, we'd doubt the hypothesis and conclude that there is evidence that $p > 1/3$.



Approach (ii), quantiles

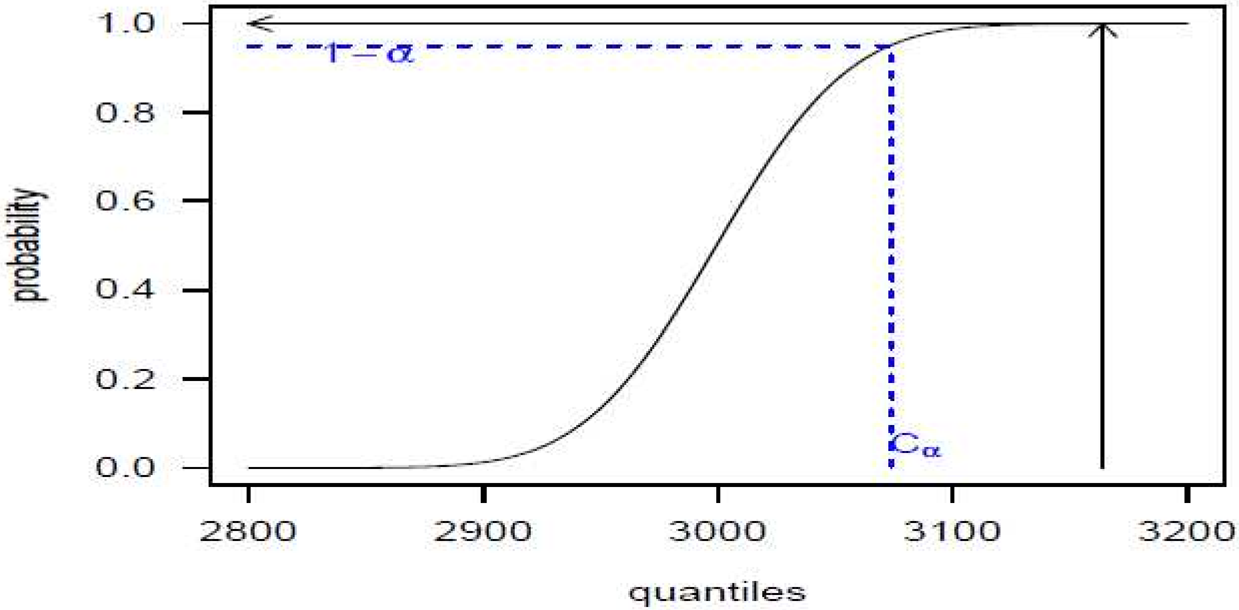
Clearly, large values of X support H_1, so we'd want a critical region of the form $x \geq c$ where c is chosen to give the desired significance level, α. That is, for $\alpha = 0.05$, say, the upper tail 5% quantile of the binomial distribution with $p = \frac{1}{3}$ and $N = 9000$ is 3074. (`qbinom(size=N,prob=px,p=0.05,lower.tail=F)`)

The observed value 3164 exceeds this and thus lies in the critical region $[c, \infty]$. So we reject H_0 **at the 5% significance level**. That is, we will come to the conclusion that $p > 1/3$, but in so doing, we'll recognize the fact that the probability could be as large as 0.05 that we've rejected H_0 wrongly.

The 2 methods are really the same thing. Figure A shows the distribution function for $\text{Bin}(9000, \frac{1}{3})$ with the observed quantile 3164 and associated with it is $P(X > 3164)$ The dashed lines show the upper $\alpha = 0.05$ probability and the quantile $C_{1-\alpha}$. The event that $X > C_{1-\alpha}$ has a probability $p < \alpha$.

The rejection region can be defined either by the probabilities or the quantiles.

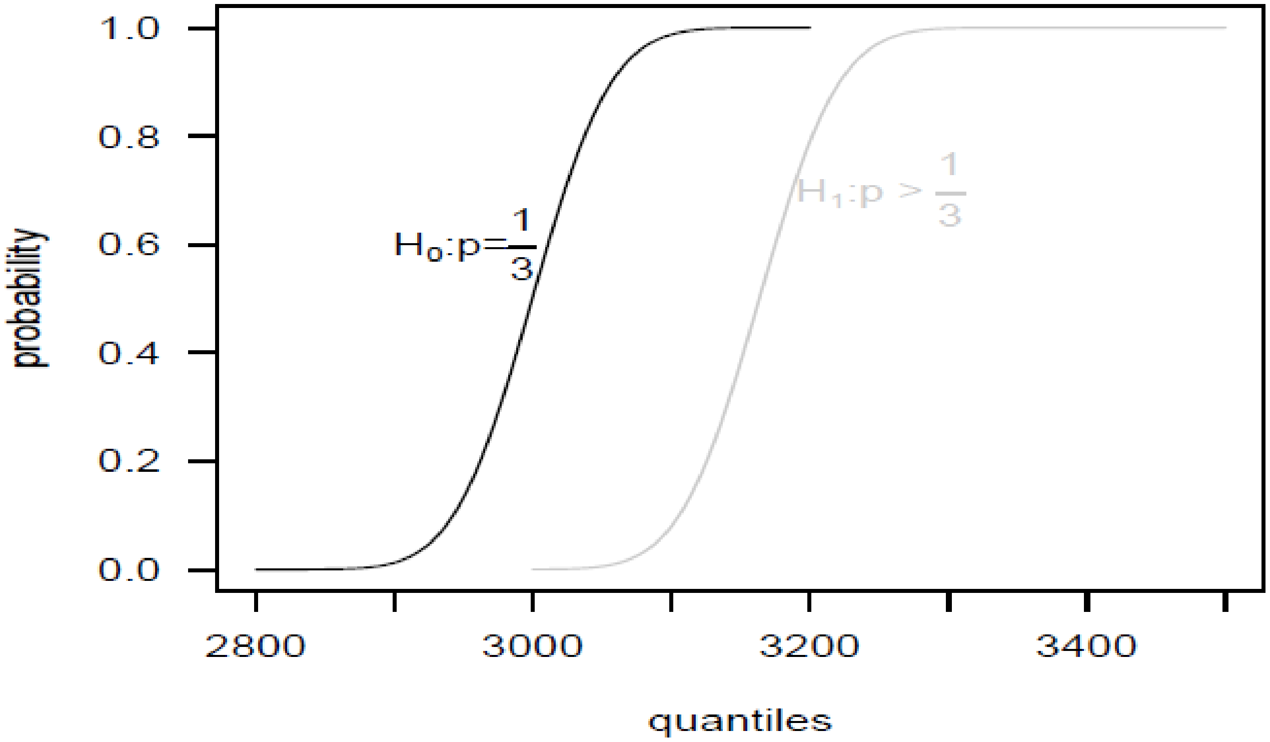
Figure A : using either quantiles or probability to test the null hypothesis



In doing this sort of problem it helps to draw a diagram, or at least try to visualize the partitioning of the sample space as suggested in Figure

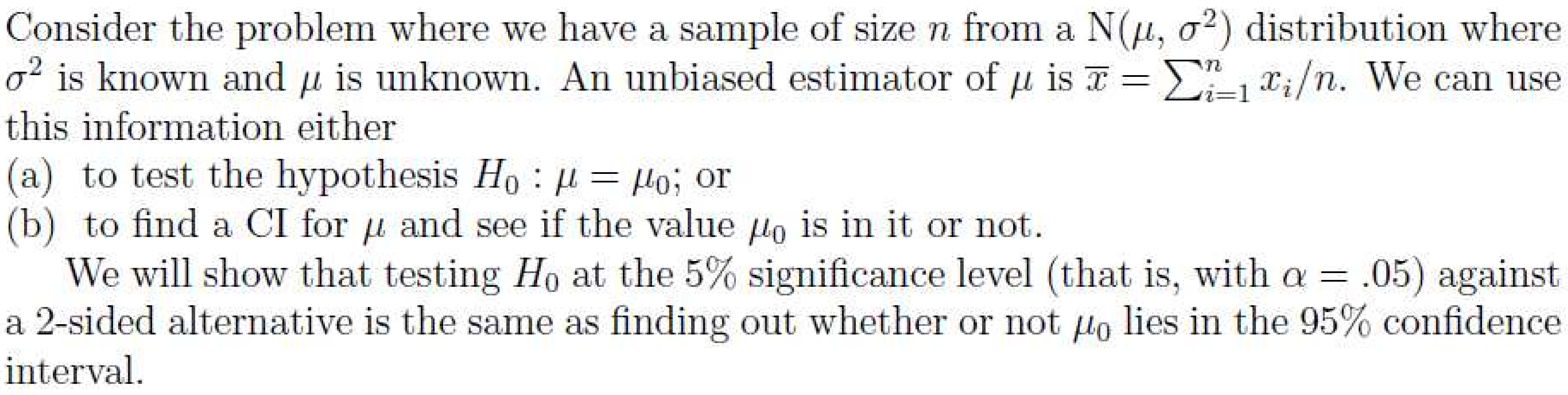
If $x \in R$ it seems much more likely that the actual distribution of X is given by a curve similar to the one on the right hand side, with mean somewhat greater than 3000.

Figure : One Sided Alternative – Binomial.



3 Connection between Hypothesis testing and CI’s

3.1 Two faces of the same coin



Consider the problem where we have a sample of size n from a N(μ, σ^2) distribution where σ^2 is known and μ is unknown. An unbiased estimator of μ is $\overline{x} = \sum_{i=1}^{n} x_i/n$. We can use this information either

(a) to test the hypothesis $H_0 : \mu = \mu_0$; or

(b) to find a CI for μ and see if the value μ_0 is in it or not.

We will show that testing H_0 at the 5% significance level (that is, with $\alpha = .05$) against a 2-sided alternative is the same as finding out whether or not μ_0 lies in the 95% confidence interval.

(a) For $H_1 : \mu \neq \mu_0$ we reject H_0 at the 5% significance level if

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > 1.96 \quad \text{or} \quad \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} < -1.96.$$

That is, if

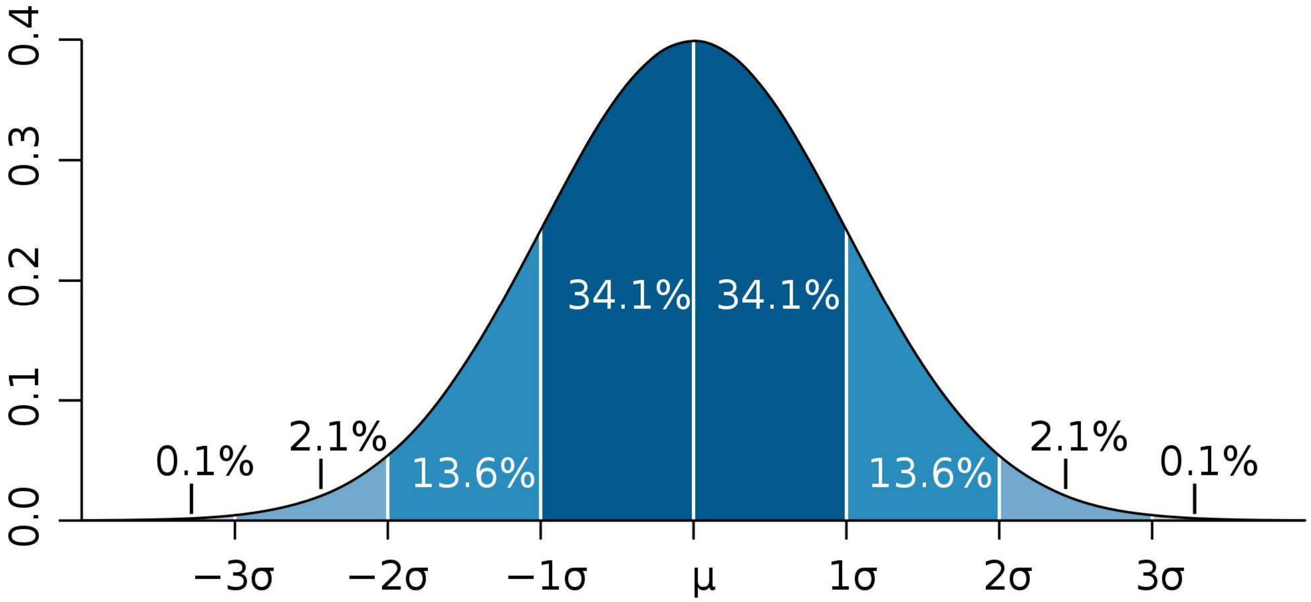
$$\frac{|\overline{x} - \mu_0|}{\sigma/\sqrt{n}} > 1.96.$$

Or, using the "P-value", if $\overline{x} > \mu_0$ we calculate the probability of a value as extreme or more extreme than this, in either direction. That is, calculate

$$P = 2 \times P(\overline{X} > \overline{x}) = 2 \times P_N\left(Z > \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right).$$

If $P < .05$ the result is significant at the 5% level.

http://www.statsoft.com/textbook/distribution-tables/



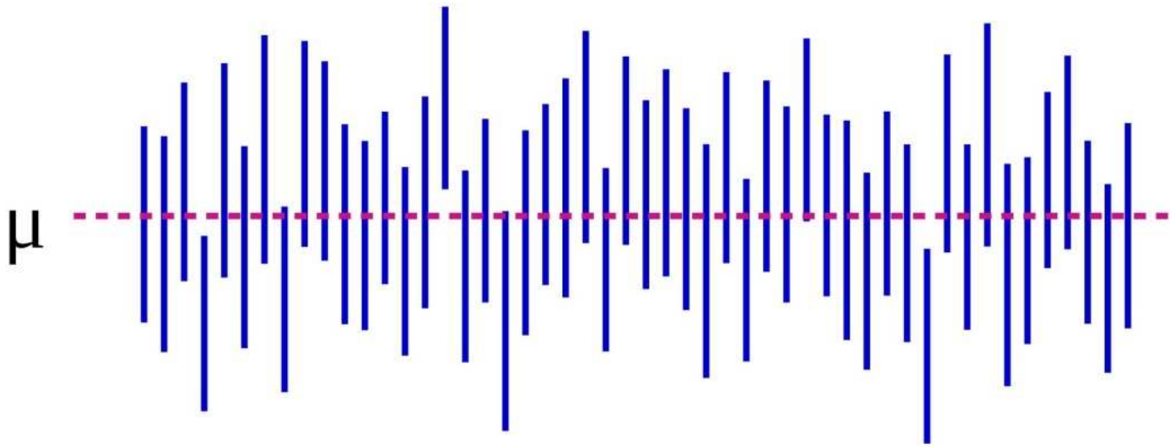
Dark blue is less than one standard deviation from the mean. For the normal distribution, this accounts for about 68% of the set, while two standard deviations from the mean (medium and dark blue) account for about 95%, and three standard deviations (light, medium, and dark blue) account for about 99.7%.

(b) A symmetric 95% confidence interval for μ is $\overline{x}\pm 1.96\sigma/\sqrt{n}$ which arose from considering the inequality

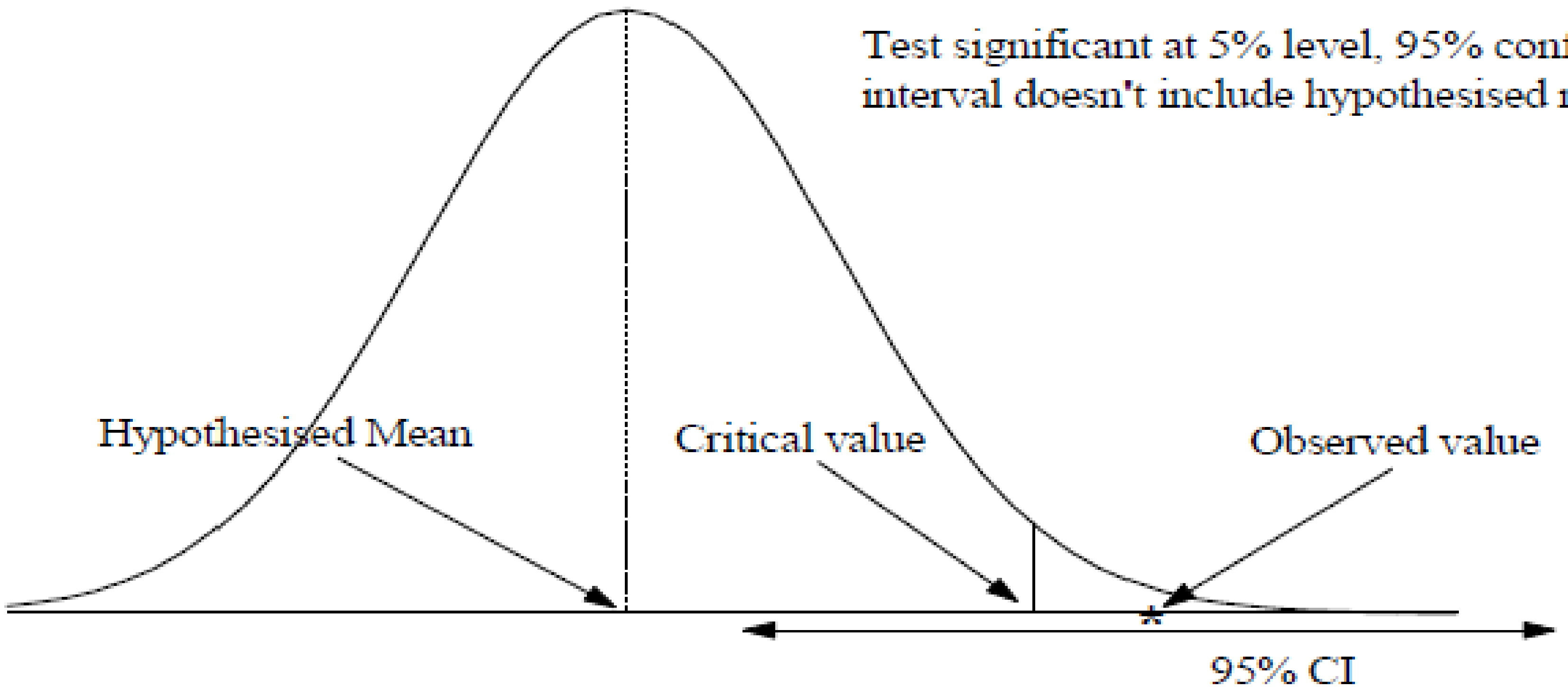
$$-1.96 < \frac{\overline{x}-\mu}{\sigma/\sqrt{n}} < 1.96$$

So, to reject H_0 at the 5% significance level is equivalent to saying that "the hypothesized value is <u>not</u> in the 95% CI". Likewise, to reject H_0 at the 1% significance level is equivalent to saying that "the hypothesized value is <u>not</u> in the 99% CI", which is equivalent to saying that "the P-value is less than 1%".

If $1\% < P < 5\%$ the hypothesized value of μ will not be within the 95% CI but it will lie in the 99% CI.



The vertical line segments represent 50 realizations of a confidence interval for μ.


Test significant at 5% level, 95% confidence interval doesn't include hypothesised mean.
Hypothesised Mean
Critical value
Observed value
*
95% CI

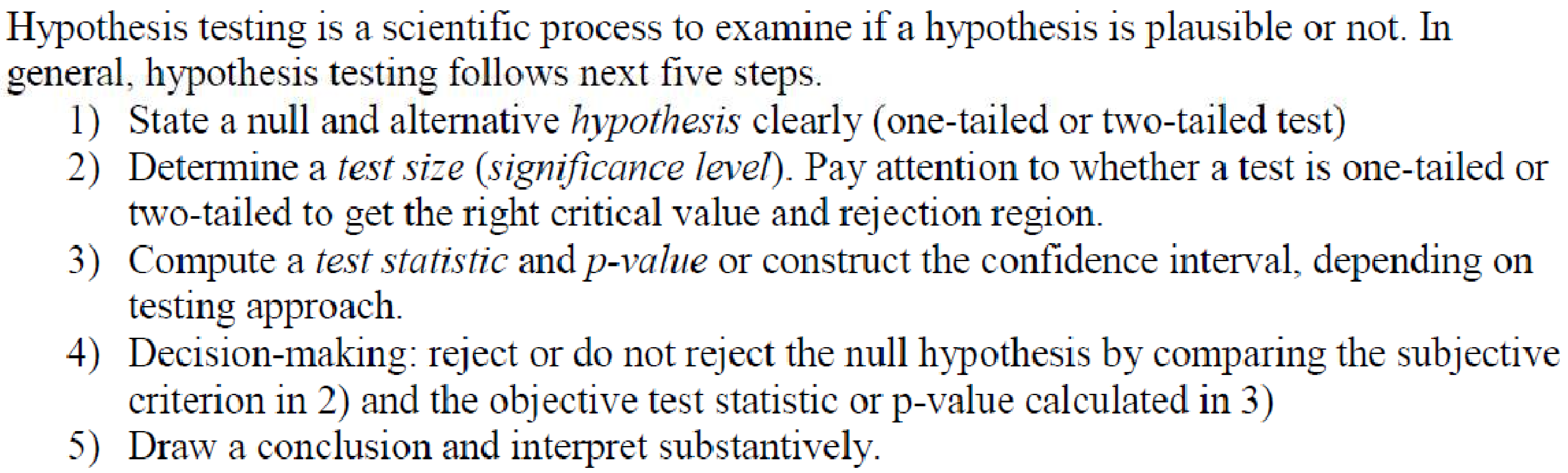
If $1\% < P < 5\%$ the hypothesized value of μ will not be within the 95% CI but it will lie in the 99% CI.

3.2 The concept of a p-value

- The statistical significance of a result is the probability that the observed relationship (e.g., between variables) or a difference (e.g., between means) in a sample occurred by pure chance ("luck of the draw"), and that in the population from which the sample was drawn, no such relationship or differences exist.
- Using less technical terms, we could say that the statistical significance of a result tells us something about the degree to which the result is "true" (in the sense of being "representative of the population").
- More technically, the value of the p-value represents a decreasing index of the reliability of a result (see Brownlee, 1960).
    - The higher the p-value, the less we can believe that the observed relation between variables in the sample is a reliable indicator of the relation between the respective variables in the population.

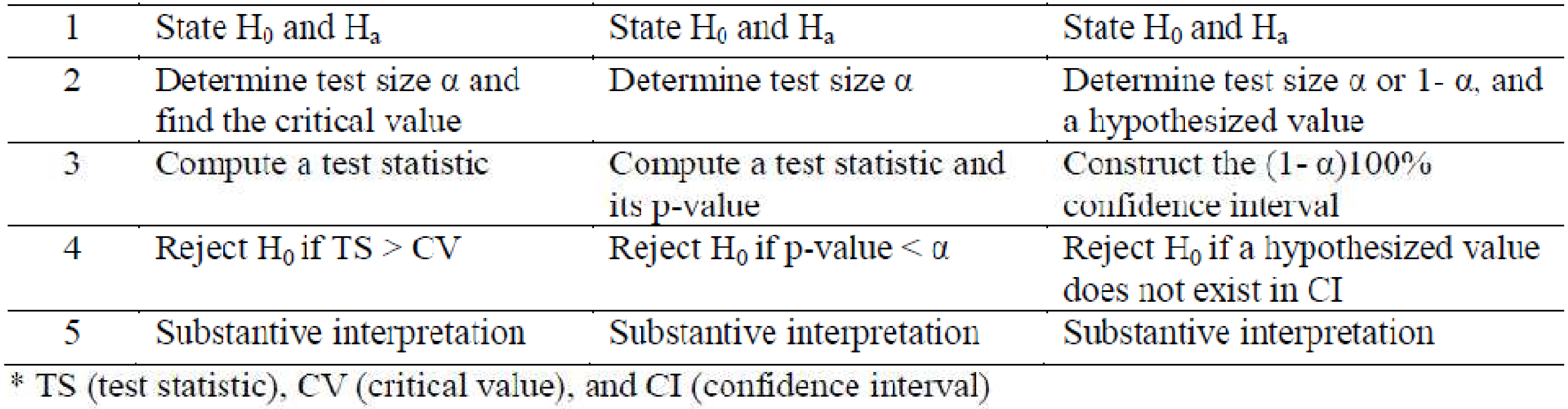
- Specifically, the p-value represents the probability of error that is involved in accepting our observed result as valid, that is, as "representative of the population."
  - If the P value is 0.03, that means that there is a 3% chance of observing a difference as large as you observed even if the two population means are identical.
  - It is tempting to conclude, therefore, that there is a 97% chance that the difference you observed reflects a real difference between populations and a 3% chance that the difference is due to chance. **Wrong.**
  - What you can say is that random sampling from identical populations would lead to a difference smaller than you observed in 97% of experiments and larger than you observed in 3% of experiments.
- When there IS a relationship between the variables in the population, the probability of replicating the study and finding that relationship is related to the statistical power of the design.

3.3 Three approaches for hypothesis testing



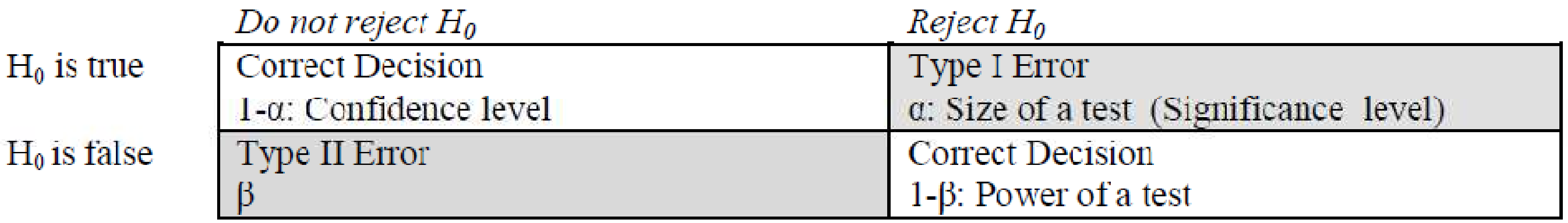
Hypothesis testing is a scientific process to examine if a hypothesis is plausible or not. In general, hypothesis testing follows next five steps.

1) State a null and alternative *hypothesis* clearly (one-tailed or two-tailed test)
2) Determine a *test size* (*significance level*). Pay attention to whether a test is one-tailed or two-tailed to get the right critical value and rejection region.
3) Compute a *test statistic* and *p-value* or construct the confidence interval, depending on testing approach.
4) Decision-making: reject or do not reject the null hypothesis by comparing the subjective criterion in 2) and the objective test statistic or p-value calculated in 3)
5) Draw a conclusion and interpret substantively.

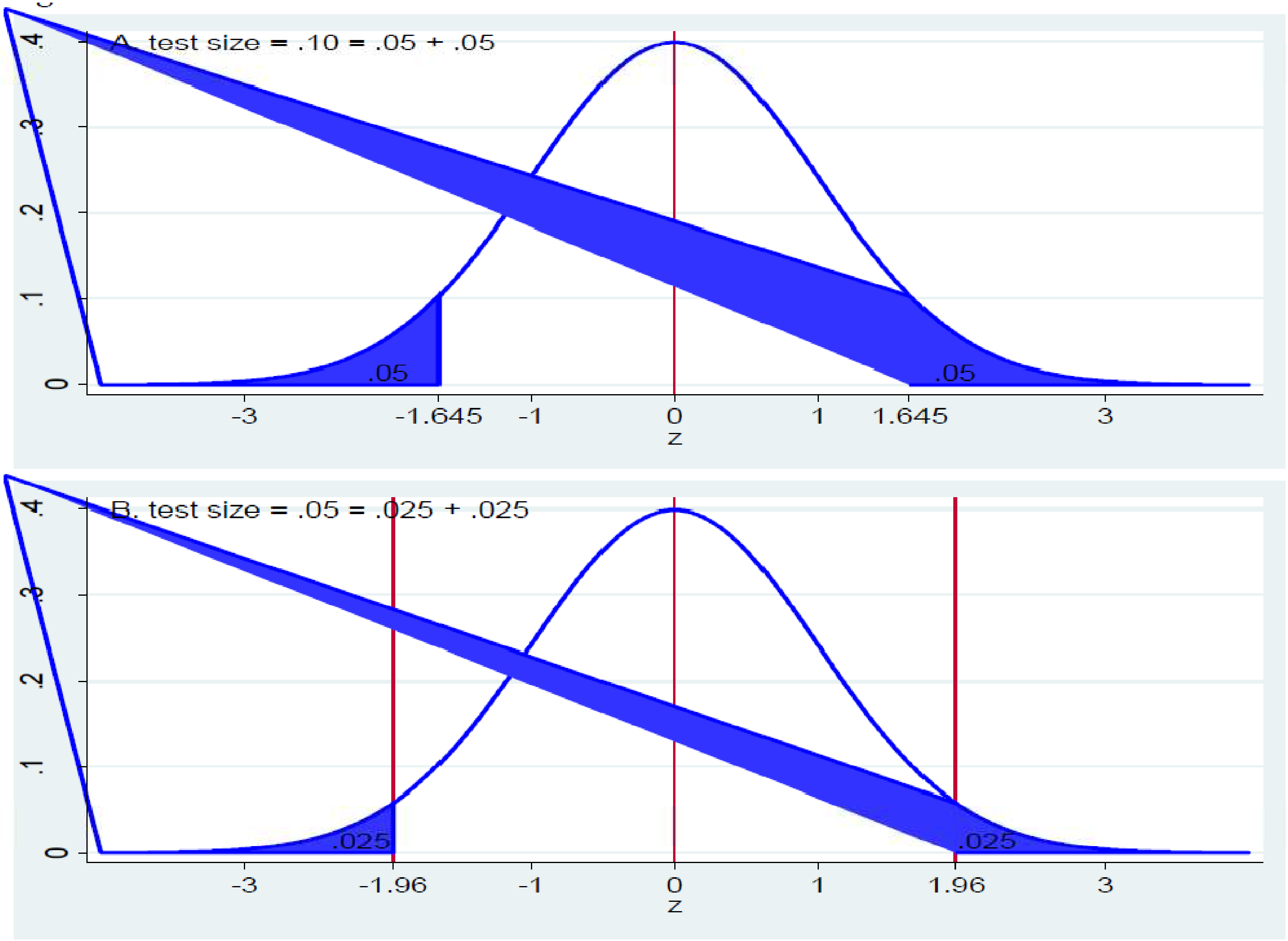


	Quantile (TS)	Probability	Confidence Interval (CI)
1	State H_0 and H_a	State H_0 and H_a	State H_0 and H_a
2	Determine test size α and find the critical value	Determine test size α	Determine test size α or 1- α, and a hypothesized value
3	Compute a test statistic	Compute a test statistic and its p-value	Construct the (1- α)100% confidence interval
4	Reject H_0 if TS > CV	Reject H_0 if p-value < α	Reject H_0 if a hypothesized value does not exist in CI
5	Substantive interpretation	Substantive interpretation	Substantive interpretation

* TS (test statistic), CV (critical value), and CI (confidence interval)

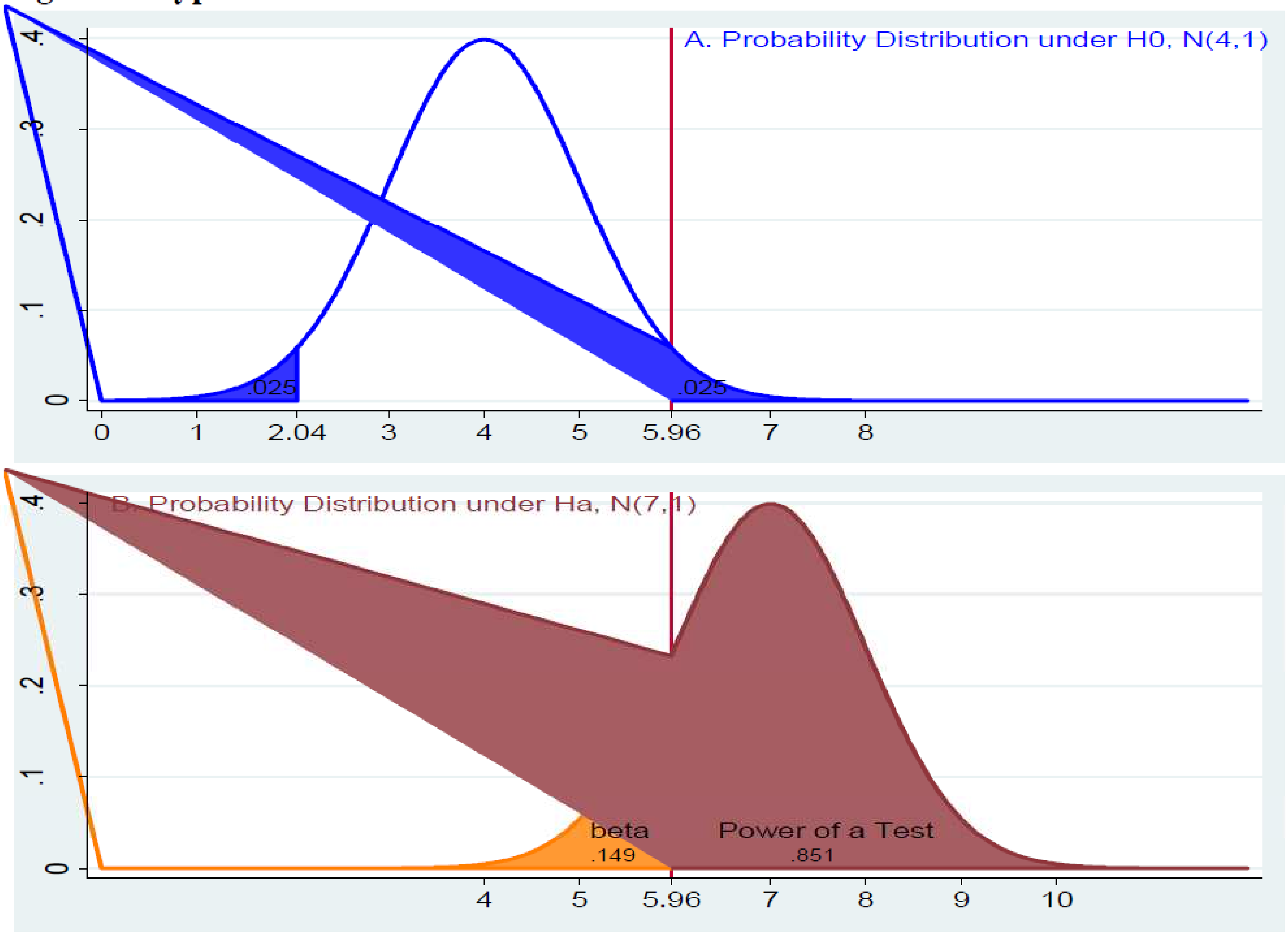


	Do not reject H_0	*Reject H_0*
H_0 is true	Correct Decision 1-α: Confidence level	Type I Error α: Size of a test (Significance level)
H_0 is false	Type II Error β	Correct Decision 1-β: Power of a test

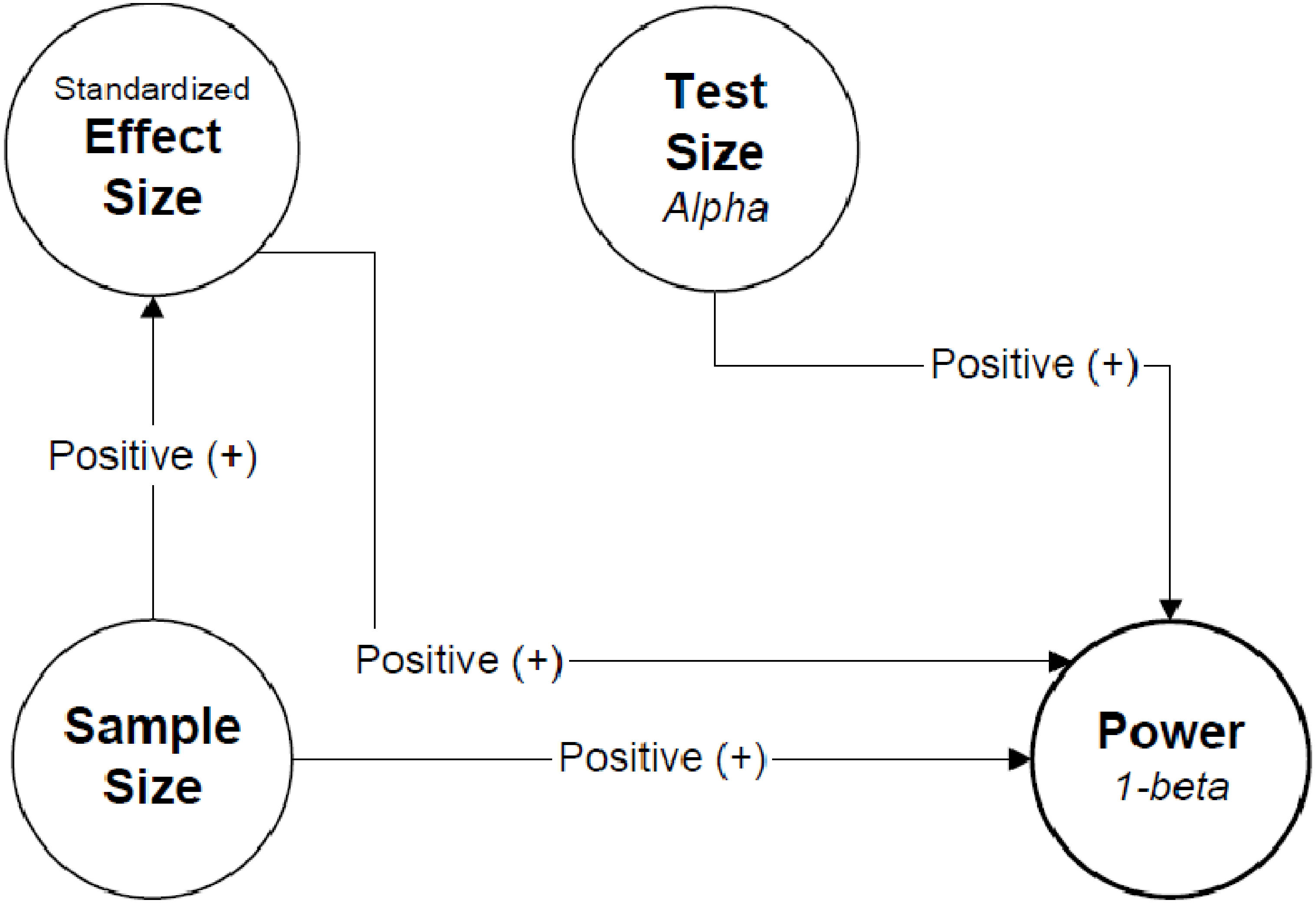


Test Size and Critical Value in the Standard Normal Distribution

Type II Error and Statistical Power of a Test



Components of a statistical power analysis



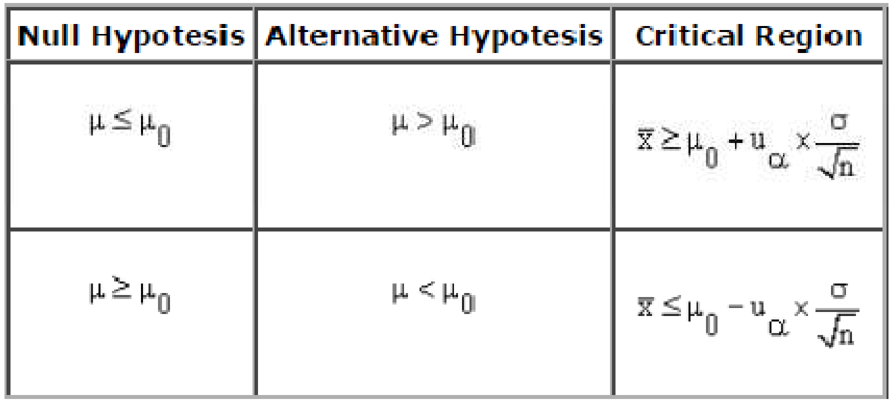
A note about effect sizes

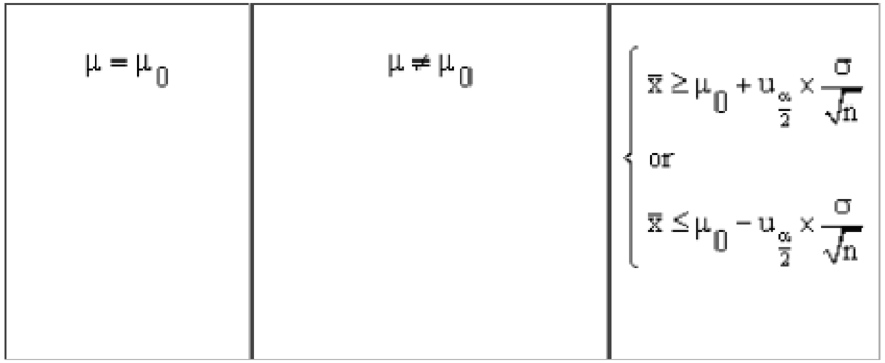
- The effect size encodes the selected research findings on a numeric scale
- There are many different types of effect size measures (OR, difference in means, correlations, …) , each suited to different research situations
- Each effect size type may also have multiple methods of computation
- An example of a standardized effect size ES is:

$$ES = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{s_{pooled}} \qquad s_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

4 One-sample problems

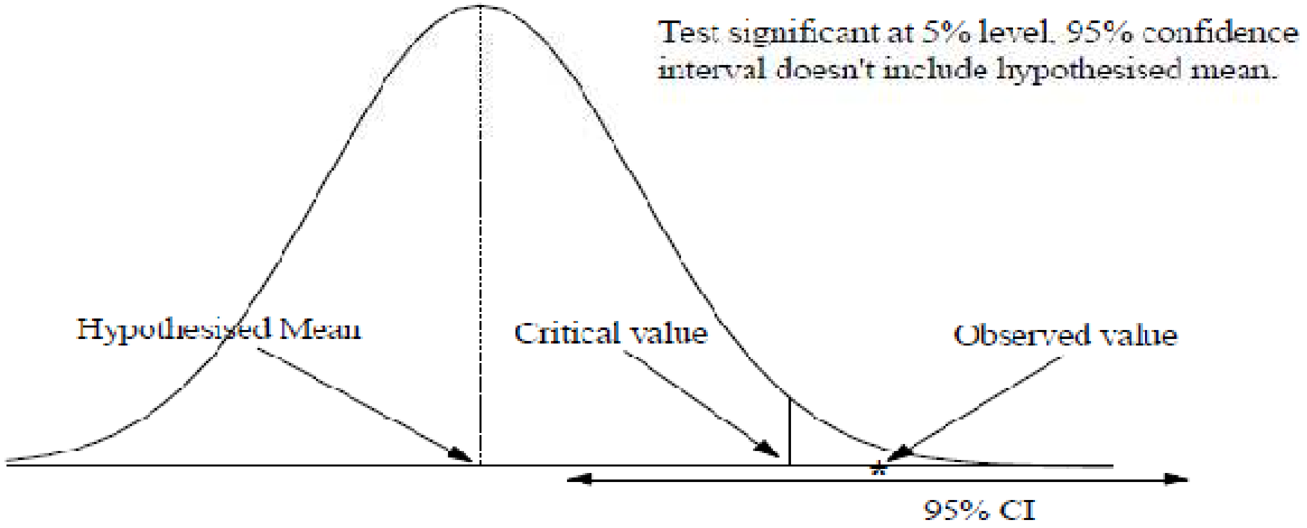
4.1 Testing hypotheses about μ





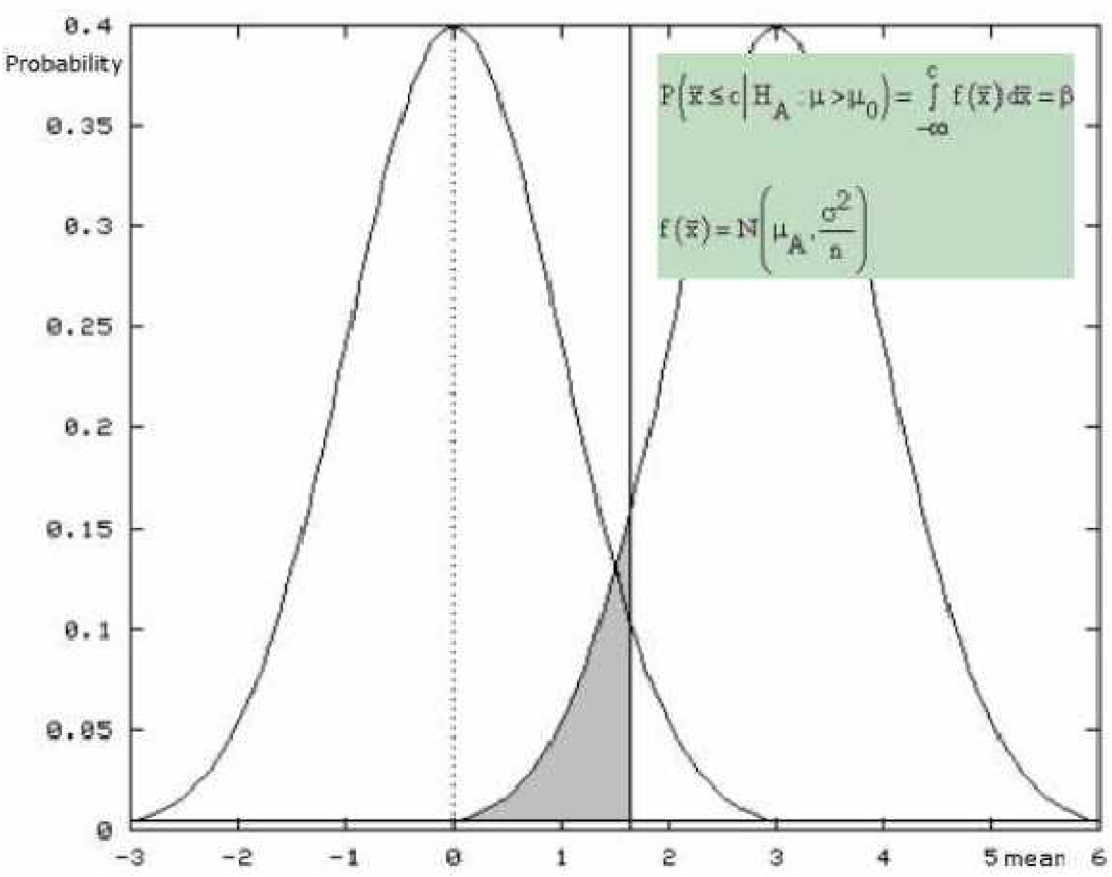
Null Hypotesis	Alternative Hypotesis	Critical Region
$\mu \leq \mu_0$	$\mu > \mu_0$	$\bar{x} \geq \mu_0 + u_\alpha \times \frac{\sigma}{\sqrt{n}}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\bar{x} \leq \mu_0 - u_\alpha \times \frac{\sigma}{\sqrt{n}}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\bar{x} \geq \mu_0 + u_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$ or $\bar{x} \leq \mu_0 - u_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$

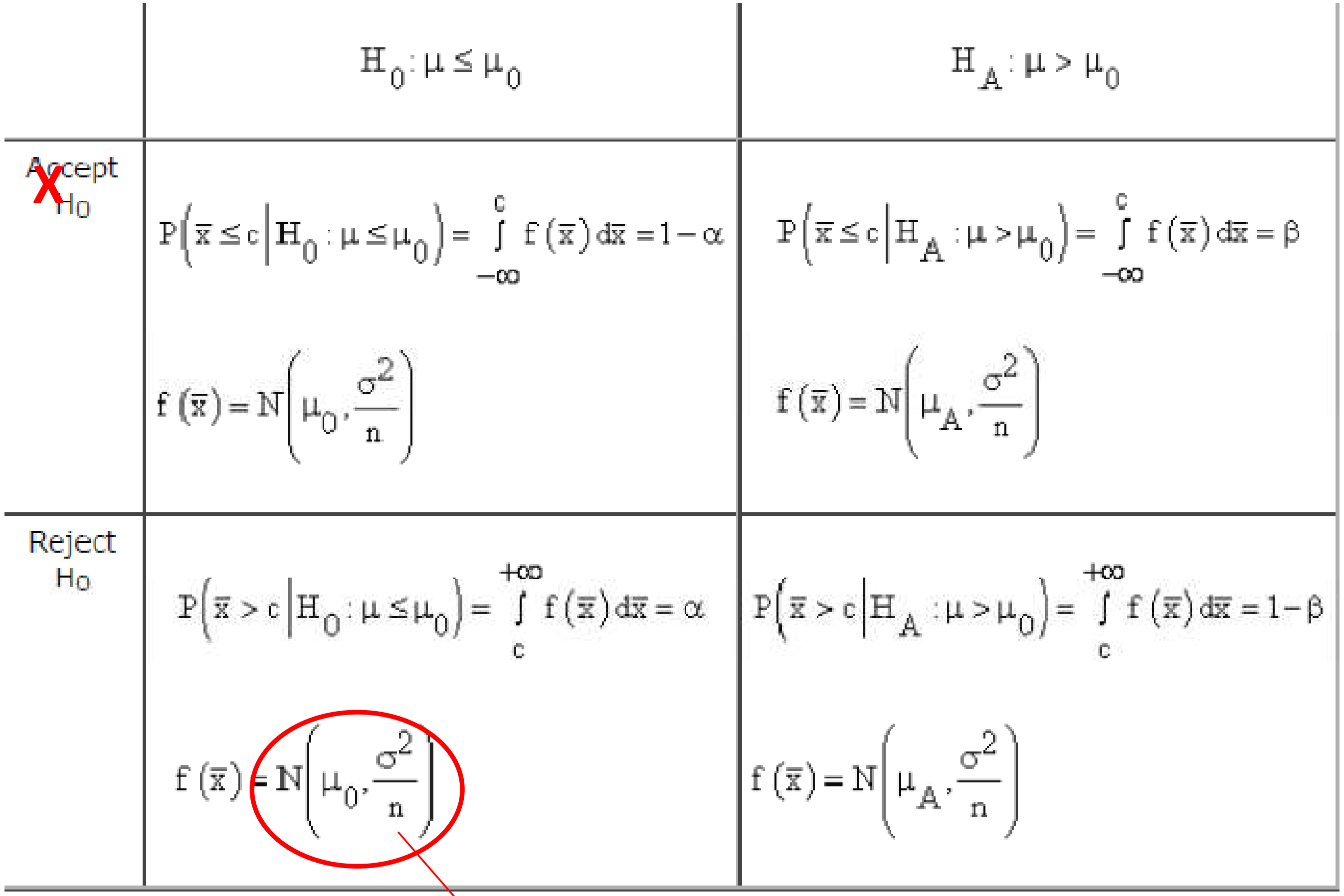
	$H_0: \mu \leq \mu_0$	$H_A: \mu > \mu_0$
~~Accept~~ H_0	Correct Probability $p = 1 - \alpha$	Type II Error Probability $p = \beta$
Reject H_0	Type I Error Probability $p = \alpha$	Correct Probability $p = 1 - \beta$



Test significant at 5% level. 95% confidence interval doesn't include hypothesised mean.

	$H_0 : \mu \le \mu_0$	$H_A : \mu > \mu_0$
Accept H0	$P\left(\bar{x} \le c \middle\| H_0 : \mu \le \mu_0\right) = \int_{-\infty}^{c} f(\bar{x})\, d\bar{x} = 1-\alpha$ $f(\bar{x}) = N\left(\mu_0, \frac{\sigma^2}{n}\right)$	$P\left(\bar{x} \le c \middle\| H_A : \mu > \mu_0\right) = \int_{-\infty}^{c} f(\bar{x})\, d\bar{x} = \beta$ $f(\bar{x}) = N\left(\mu_A, \frac{\sigma^2}{n}\right)$
Reject H0	$P\left(\bar{x} > c \middle\| H_0 : \mu \le \mu_0\right) = \int_{c}^{+\infty} f(\bar{x})\, d\bar{x} = \alpha$ $f(\bar{x}) = N\left(\mu_0, \frac{\sigma^2}{n}\right)$	$P\left(\bar{x} > c \middle\| H_A : \mu > \mu_0\right) = \int_{c}^{+\infty} f(\bar{x})\, d\bar{x} = 1-\beta$ $f(\bar{x}) = N\left(\mu_A, \frac{\sigma^2}{n}\right)$


0.4
Probability
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-3
-2
-1
0
1
2
3
4
5 mean
6
$P\left(\bar{x} \le c \middle| H_A : \mu > \mu_0\right) = \int_{-\infty}^{c} f(\bar{x})\, d\bar{x} = \beta$
$f(\bar{x}) = N\left(\mu_A, \frac{\sigma^2}{n}\right)$



	$H_0 : \mu \le \mu_0$	$H_A : \mu > \mu_0$
X Accept H_0	$P\left(\bar{x} \le c \middle\| H_0 : \mu \le \mu_0\right) = \int_{-\infty}^{c} f(\bar{x})\, d\bar{x} = 1-\alpha$ $f(\bar{x}) = N\left(\mu_0, \frac{\sigma^2}{n}\right)$	$P\left(\bar{x} \le c \middle\| H_A : \mu > \mu_0\right) = \int_{-\infty}^{c} f(\bar{x})\, d\bar{x} = \beta$ $f(\bar{x}) = N\left(\mu_A, \frac{\sigma^2}{n}\right)$
Reject H_0	$P\left(\bar{x} > c \middle\| H_0 : \mu \le \mu_0\right) = \int_{c}^{+\infty} f(\bar{x})\, d\bar{x} = \alpha$ $f(\bar{x}) = N\left(\mu_0, \frac{\sigma^2}{n}\right)$	$P\left(\bar{x} > c \middle\| H_A : \mu > \mu_0\right) = \int_{c}^{+\infty} f(\bar{x})\, d\bar{x} = 1-\beta$ $f(\bar{x}) = N\left(\mu_A, \frac{\sigma^2}{n}\right)$

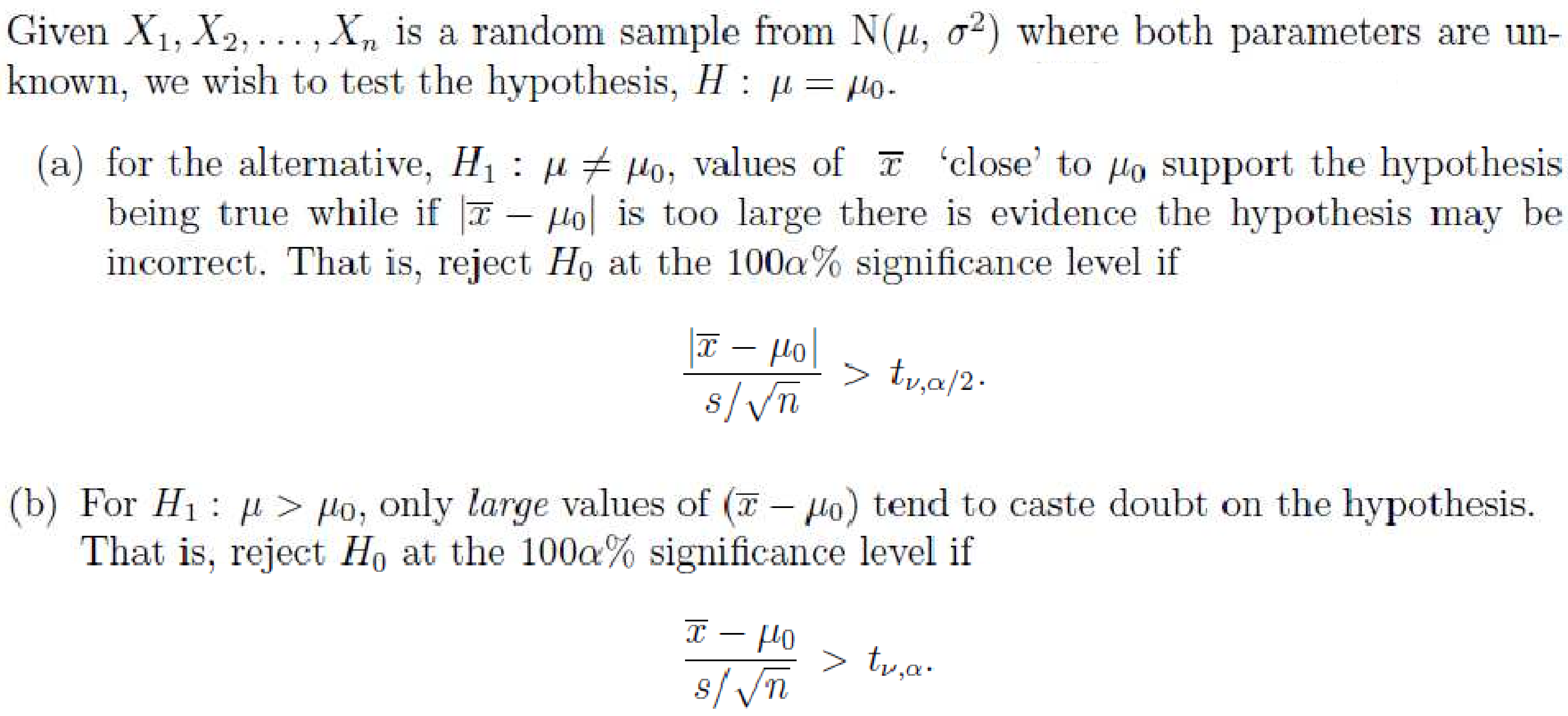
Think about the 3 ways to perform statistical hypothesis testing …

Recall

	Population σ^2	Estimation of μ	Test statistic and distribution
Case (i)	σ^2 **Known**	$\bar{X} = 1/n \sum_1^n x_i$	$Z = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$
Case (ii)	σ^2 **Unknown**	$\bar{X} = 1/n \sum_1^n x_i$	$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$ $\frac{S}{\sqrt{n}} = \widehat{sd}(\bar{X})$, S^2 unbiased for σ^2

Unknown mean

According to the table, the crucial distribution is now a t-distribution



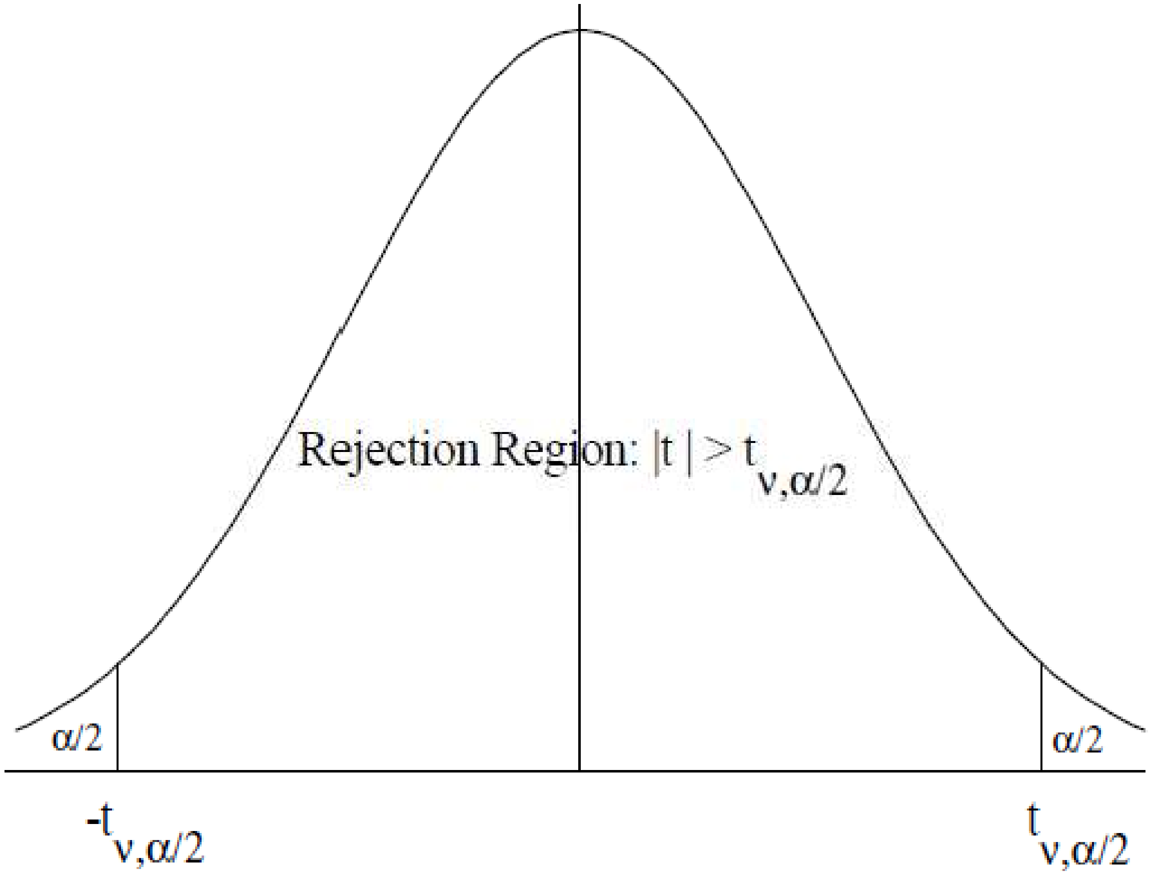
Given $X_1, X_2, \ldots, X_n$ is a random sample from $\mathrm{N}(\mu,\ \sigma^2)$ where both parameters are unknown, we wish to test the hypothesis, $H : \mu = \mu_0$.

(a) for the alternative, $H_1 : \mu \neq \mu_0$, values of $\overline{x}$ 'close' to μ_0 support the hypothesis being true while if $|\overline{x} - \mu_0|$ is too large there is evidence the hypothesis may be incorrect. That is, reject H_0 at the $100\alpha\%$ significance level if

$$\frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} > t_{\nu,\alpha/2}.$$

(b) For $H_1 : \mu > \mu_0$, only *large* values of $(\overline{x} - \mu_0)$ tend to caste doubt on the hypothesis. That is, reject H_0 at the $100\alpha\%$ significance level if

$$\frac{\overline{x} - \mu_0}{s/\sqrt{n}} > t_{\nu,\alpha}.$$

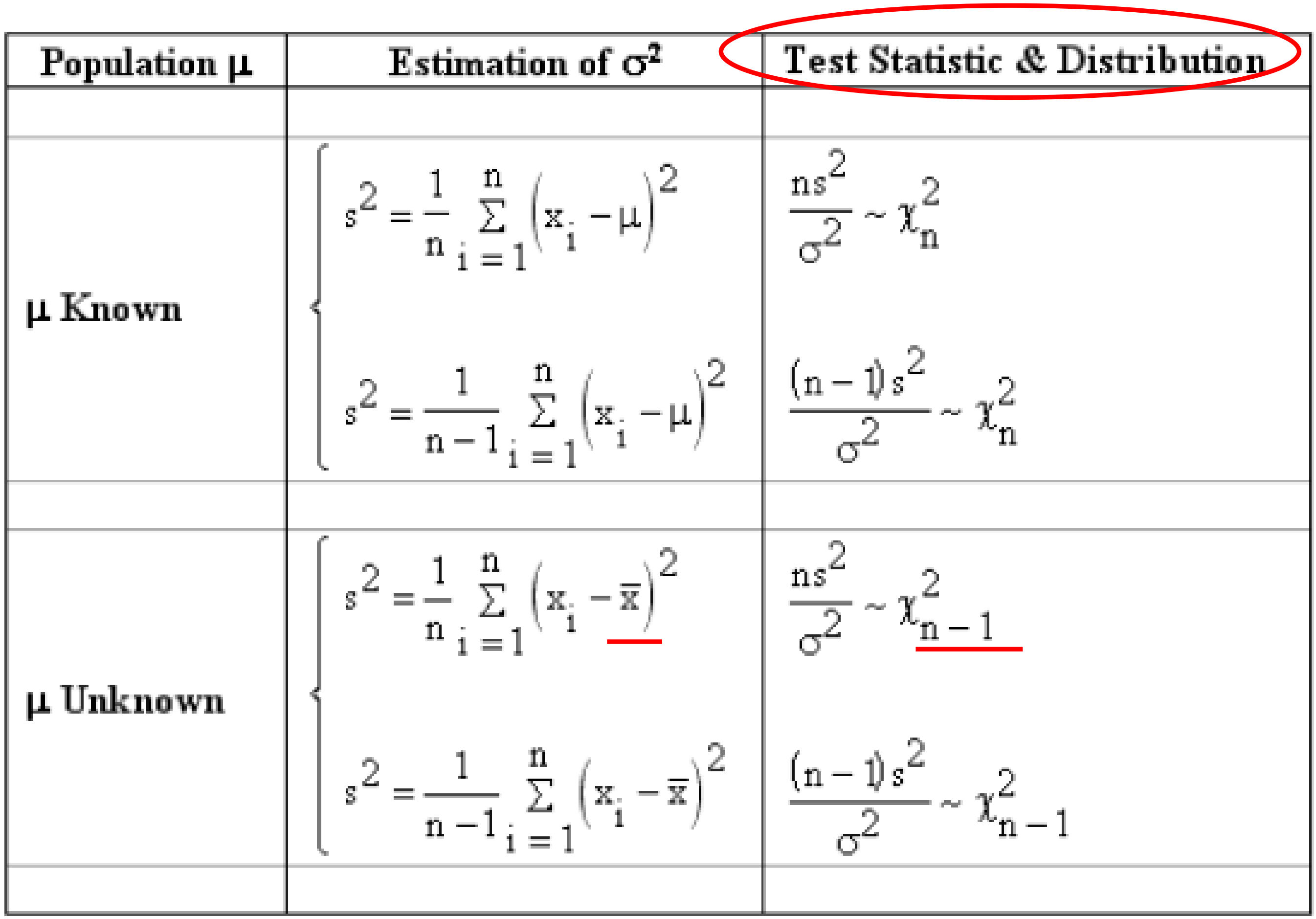


An alternative $H_1 : \mu < \mu_0$, would be treated similarly to (b) but with lower critical value $-t_{\nu,\alpha}$.

4.2 Testing hypotheses about σ^2

Case (i)

Case (ii)



Population μ	Estimation of σ^2	Test Statistic & Distribution
μ Known	$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$	$\frac{ns^2}{\sigma^2} \sim \chi^2_n$ $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_n$
μ Unknown	$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \underline{\bar{x}})^2$ $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$\frac{ns^2}{\sigma^2} \sim \chi^2_{\underline{n-1}}$ $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

Again the cases (i) μ unknown; and (ii)μ known are considered separately.

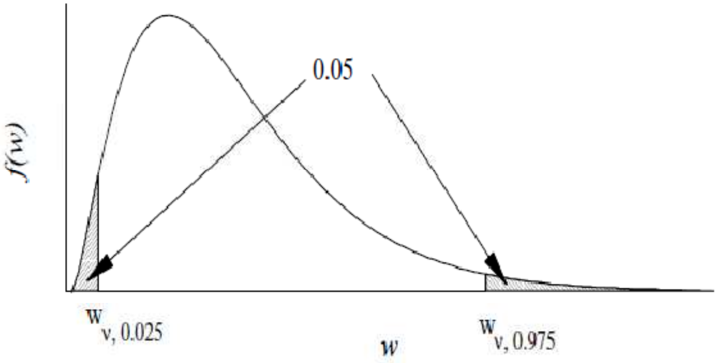
Case (ii)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a N(μ, σ^2) distribution where μ **is unknown**, and suppose we wish to test the hypothesis

$$H : \sigma^2 = \sigma_0^2 \ \text{ against } \ A : \sigma^2 \neq \sigma_0^2.$$

Under H, $\nu S^2/\sigma_0^2 \sim \chi_\nu^2$ and values of $\nu s^2/\sigma_0^2$ too large or too small would support A. For $\alpha = .05$, say, and equal-tail probabilities we have as critical region

$$R = \left\{ s^2 : \ \frac{\nu s^2}{\sigma_0^2} > w_{\nu,.025} \ \text{ or } \ \frac{\nu s^2}{\sigma_0^2} < w_{\nu,.975} \right\}.$$



Consider now a one-sided alternative. Suppose we wish to test

$$H : \sigma^2 = \sigma_0^2 \ \text{ against } \ A : \sigma^2 > \sigma_0^2.$$

Large values of s^2 would support this alternative. That is, for $\alpha = .05$, use as critical region

$$\{s^2 : \ \nu s^2/\sigma_0^2 > w_{\nu,.05}\}.$$

Similarly, for the alternative A: $\sigma^2 < \sigma_0^2$, a critical region is

$$\{s^2 : \ \nu s^2/\sigma_0^2 < w_{\nu,.95}\}.$$

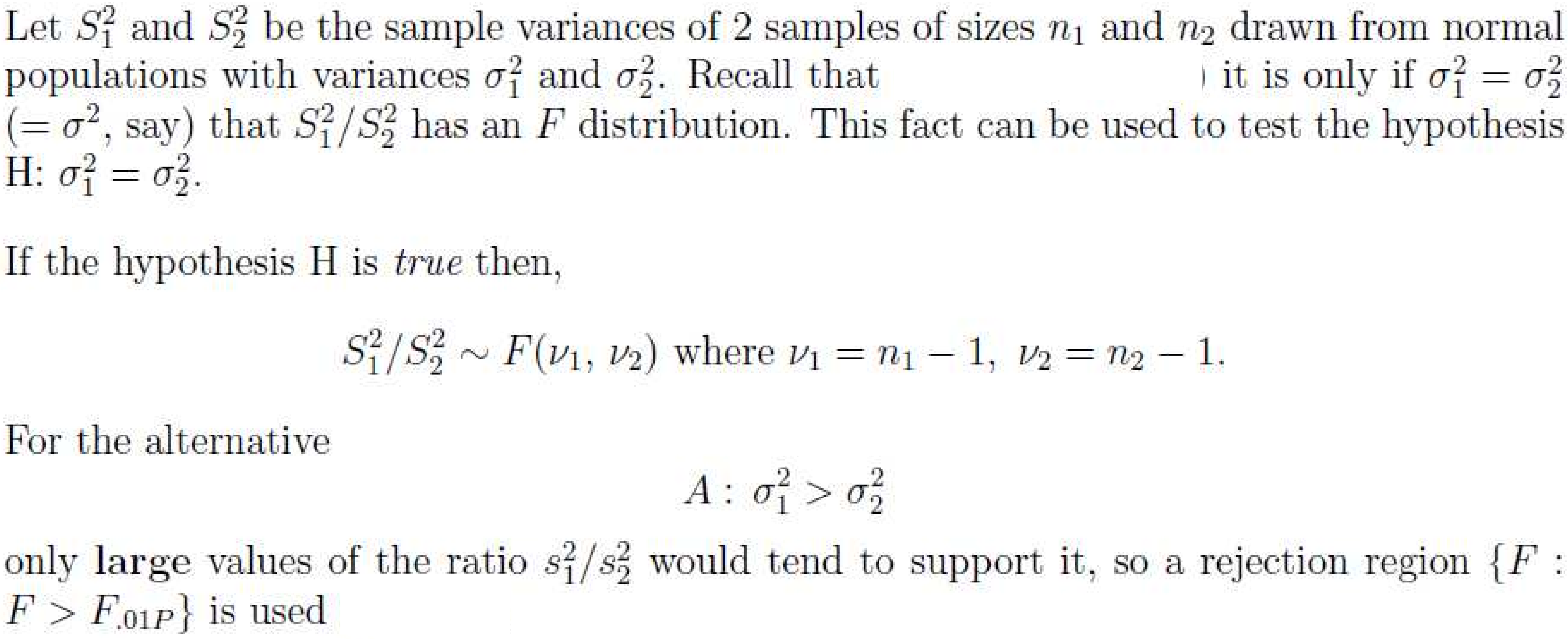
Case (i)

Let $X_1, X_2, \ldots, X_n$ be a random sample from N(μ, σ^2) where μ is **known**, and suppose we wish to test H: $\sigma^2 = \sigma_0^2$. Again we use the fact that **if H is true**, $nS^{*2}/\sigma_0^2 \sim \chi_n^2$ where $S^{*2} = \sum_{i=1}^n (X_i - \mu)^2/n$, and the rejection region for a size-α 2-tailed test, for example, would be

$$\left\{ s^{*2} : \ \frac{ns^{*2}}{\sigma_0^2} > w_{n,\alpha/2} \ \text{ or } \ \frac{ns^{*2}}{\sigma_0^2} < w_{n,1-(\alpha/2)} \right\}$$

5 Two-Sample Problems

5.1 Testing equality of sample variances



Let S_1^2 and S_2^2 be the sample variances of 2 samples of sizes n_1 and n_2 drawn from normal populations with variances σ_1^2 and σ_2^2. Recall that) it is only if $\sigma_1^2 = \sigma_2^2$ $(= \sigma^2$, say) that S_1^2/S_2^2 has an F distribution. This fact can be used to test the hypothesis H: $\sigma_1^2 = \sigma_2^2$.

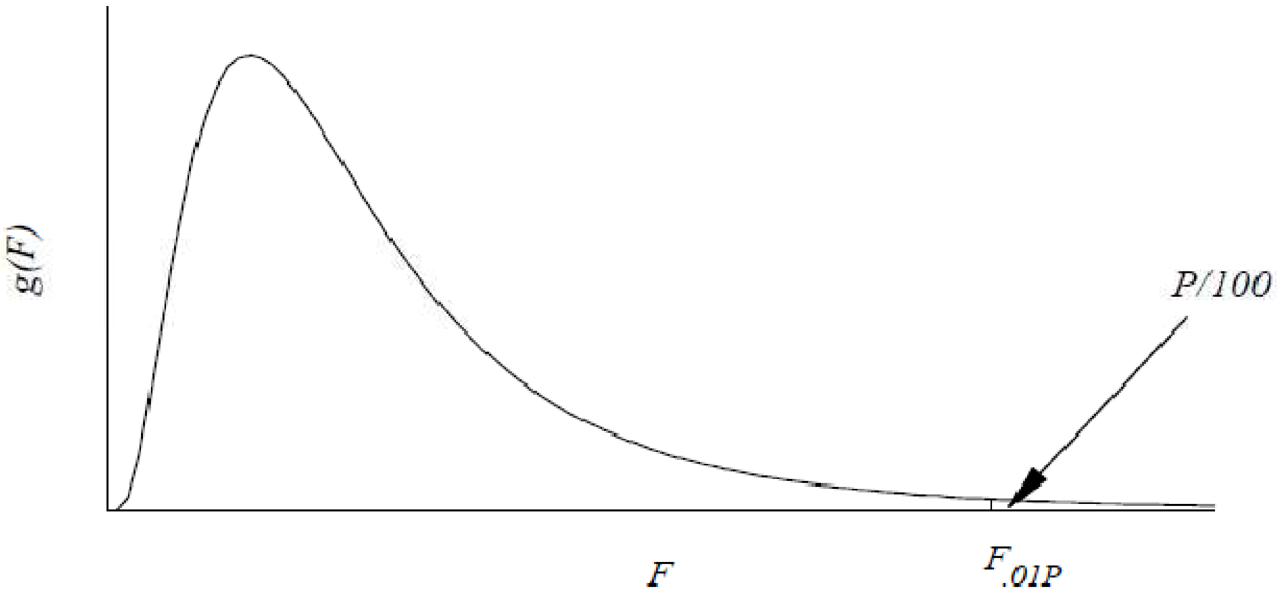
If the hypothesis H is *true* then,

$$S_1^2/S_2^2 \sim F(\nu_1,\, \nu_2) \text{ where } \nu_1 = n_1 - 1,\ \nu_2 = n_2 - 1.$$

For the alternative

$$A:\ \sigma_1^2 > \sigma_2^2$$

only **large** values of the ratio s_1^2/s_2^2 would tend to support it, so a rejection region $\{F : F > F_{.01P}\}$ is used



Since only the right hand tail areas of the distribution are tabulated it is convenient to always use $s_i^2/s_j^2 > 1$. That is, always put the larger sample variance in the numerator.

Reciprocal of an F-variate

Let the random variable $F \sim F(\nu_1, \nu_2)$ and let $Y = 1/F$. Then Y has p.d.f.

$$\begin{aligned}
f(y) &= g(F)\left|\frac{dF}{dy}\right| \\
&= \frac{\nu_1^{(\nu_1/2)} y^{1-(\nu_1/2)} \nu_2^{\nu_2/2} y^{(\nu_1+\nu_2)/2}}{B(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2)(\nu_2 y + \nu_1)^{(\nu_1+\nu_2)/2}} \frac{1}{y^2} \\
&= \frac{\nu_2^{\nu_2/2} \nu_1^{\nu_1/2} y^{(\nu_2/2)-1}}{B(\frac{1}{2}\nu_2, \frac{1}{2}\nu_1)(\nu_1 + \nu_2 y)^{(\nu_1+\nu_2)/2}}, \ y \in [0, \infty).
\end{aligned}$$

Thus if $F \sim F(\nu_1, \nu_2)$ and $Y = 1/F$ then $Y \sim F(\nu_2, \nu_1)$.

One-sided test

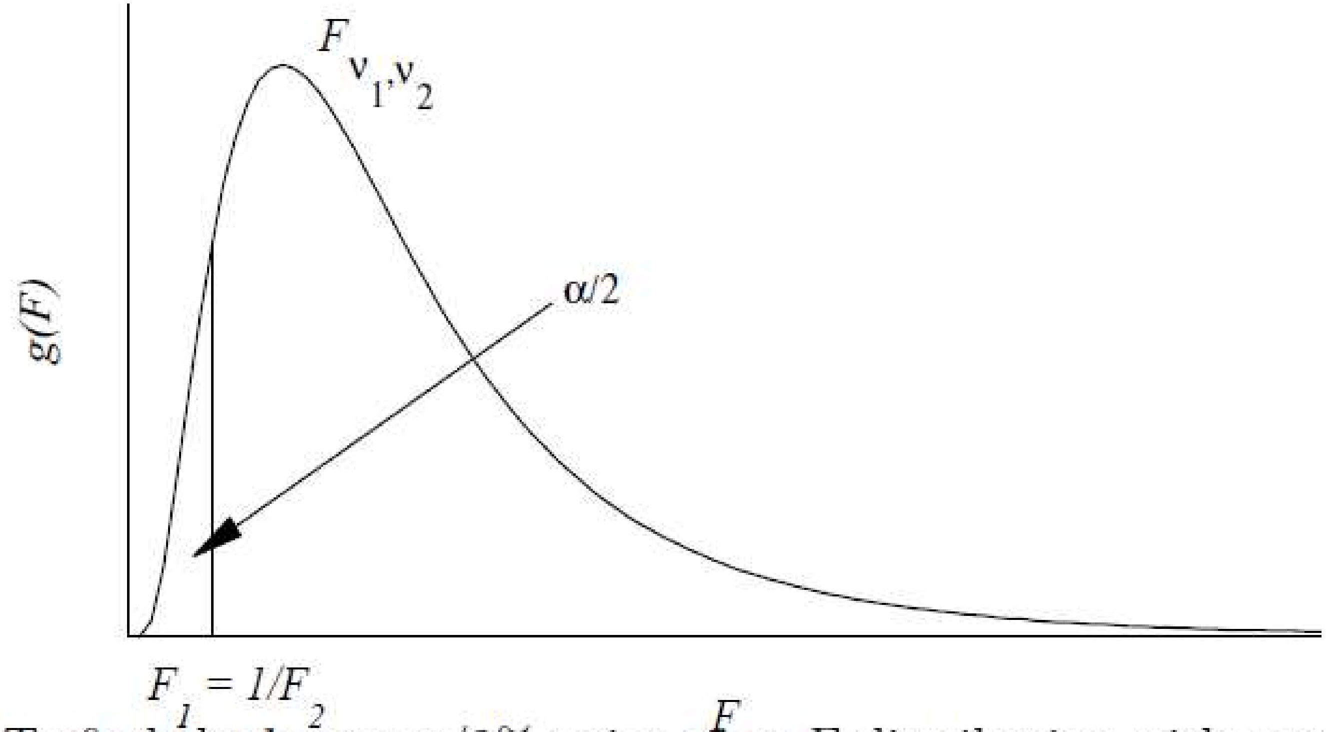
If the alternative is $\sigma_1^2 \neq \sigma_2^2$, then both tails of the distribution could be used for rejection regions, so it may be necessary to find the lower critical value. Let $F \sim F(\nu_1, \nu_2)$. That is we want find a value F_1 so that

$$\int_0^{F_1} g(F)\, dF = \alpha/2.$$

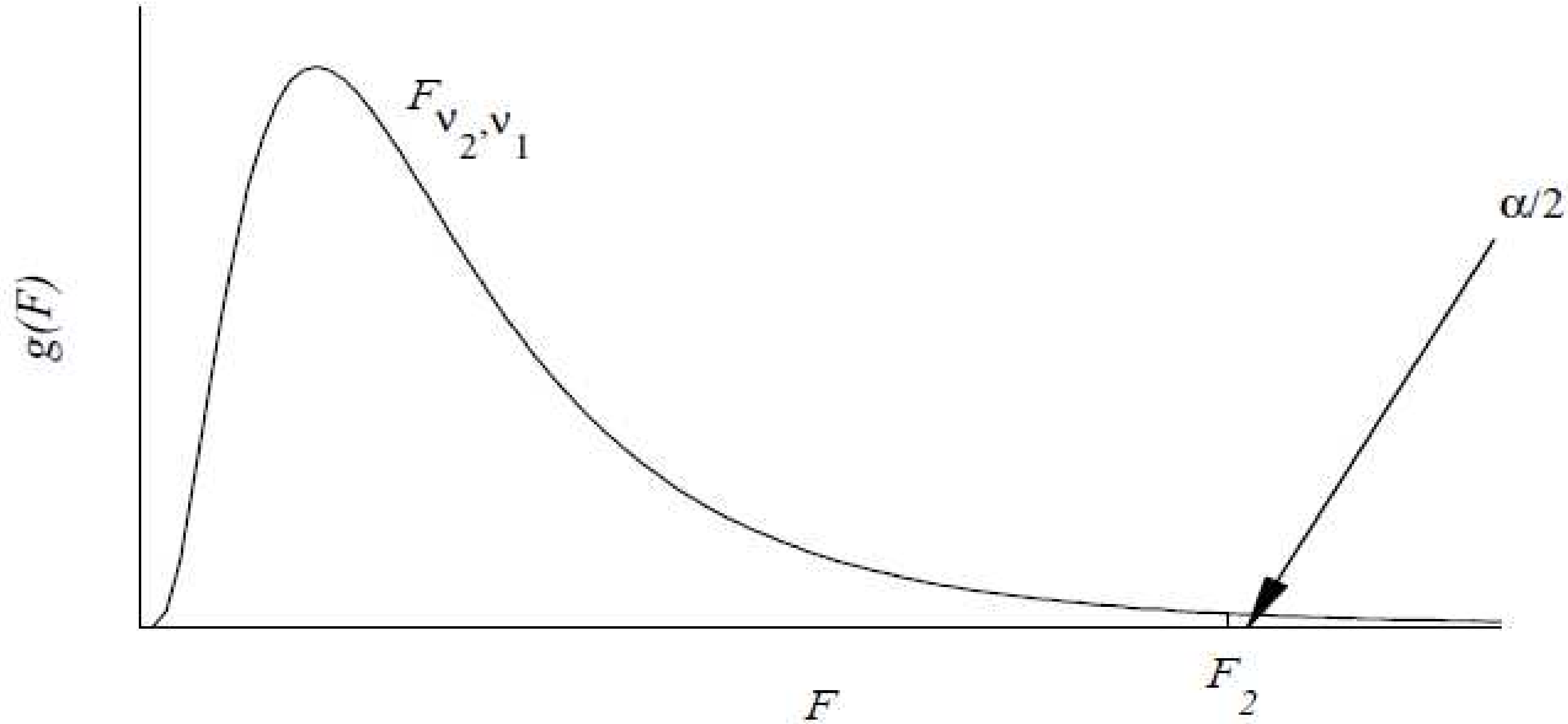
Put $Y = 1/F$ so that $Y \sim F(\nu_2, \nu_1)$. Then

$$\int_0^{F_1} g(F) dF = P(F \leq F_1) = P(Y > 1/F_1) = P(Y > F_2), \text{ say.}$$

Thus to find the lower $\frac{\alpha}{2}\%$ critical value, F_1, first find the upper $\frac{\alpha}{2}\%$ critical value, F_2 from tables of $F(\nu_2, \nu_1)$, and then calculate F_1 as $F_1 = 1/F_2$.



To find the **lower** $\alpha/2\%$ point of an F distribution with parameters ν_1, ν_2, take the reciprocal of the **upper** $\alpha/2\%$ point of an F distribution with parameters ν_2, ν_1.


F_{ν_2,ν_1}
$\alpha/2$
$g(F)$
F
F_2

5.2 Testing equality of normal means

Given $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$ are independent random samples from $\mathrm{N}(\mu_1, \sigma_1^2)$ and $\mathrm{N}(\mu_2, \sigma_2^2)$ respectively, we may wish to test $H: \mu_1 - \mu_2 = \delta_0$, say.) we can see that, under H_0,

$$\frac{\overline{X} - \overline{Y} - \delta_0}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

So H_0 can be tested against one- or two-sided alternatives.

Note however, that we have assumed that both populations have the same variance σ^2, and this in general is not known. More generally, let $X_1, X_2, \ldots, X_{n_1}$ be a random sample from $\mathrm{N}(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_{n_2}$ be an independent random sample from $\mathrm{N}(\mu_2, \sigma_2^2)$ where μ_1, μ_2, σ_1^2, σ_2^2 are unknown, and suppose we wish to test $H: \mu_1 - \mu_2 = \delta_0$. From the samples of sizes n_1, n_2 we can determine $\overline{x}$, $\overline{y}$, s_1^2, s_2^2. We first test the preliminary hypothesis that $\sigma_1^2 = \sigma_2^2$ and if evidence supports this, then we regard the populations as having a common variance σ^2. So the procedure is:

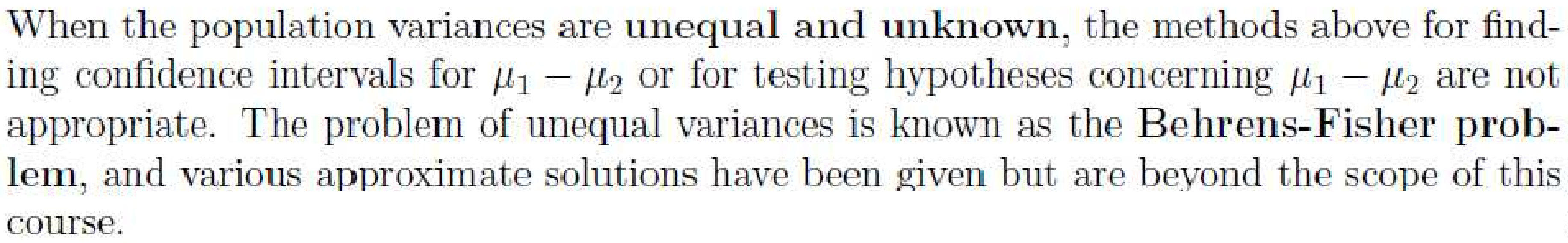
(i) Test $H_0 : \sigma_1^2 = \sigma_2^2 (= \sigma^2)$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$, using the fact that under H_0, $S_1^2/S_2^2 \sim F_{\nu_1,\nu_2}$. [This is often referred to as testing sample variances for compatibility.] A two-sided alternative and a two-tailed test is always appropriate here. We don't have any prior information about the variances. If this test is "survived" (that is, if H_0 is not rejected), proceed to (ii).

(ii) Pool s_1^2 and s_2^2 using $s^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2}$ which is now an estimate of σ^2 based on $\nu_1 + \nu_2$ degrees of freedom.

(iii) Test $H_0 : \mu_1 - \mu_2 = \delta_0$ against the appropriate alternative using the fact that, under H_0,

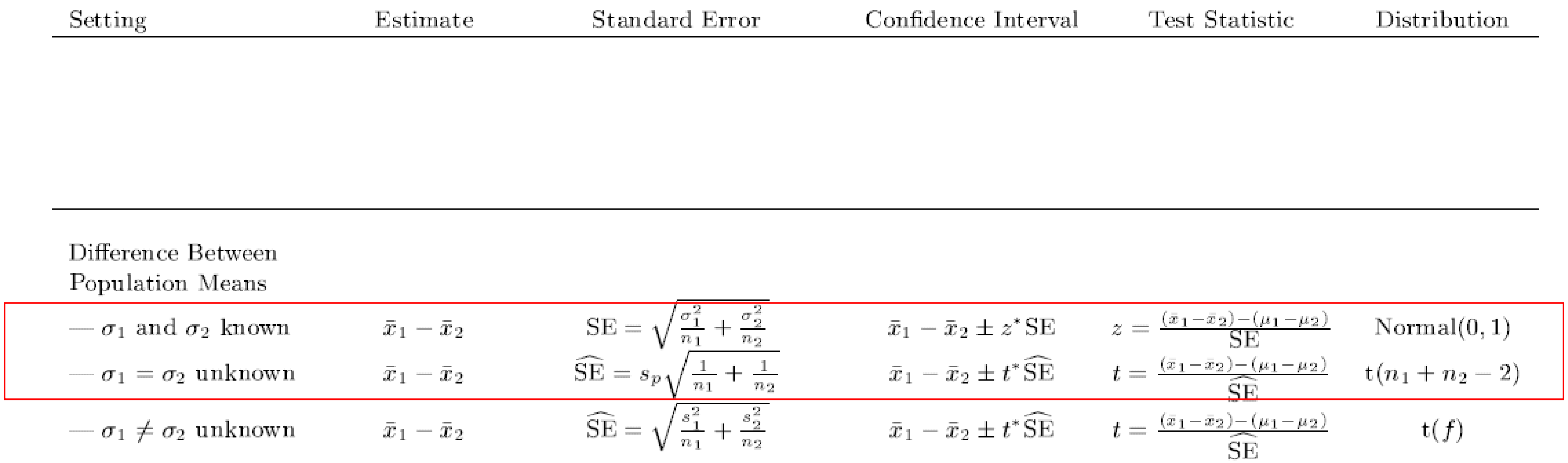
$$\frac{\overline{X} - \overline{Y} - \delta_0}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{\nu_1+\nu_2}.$$

Comment



When the population variances are **unequal and unknown,** the methods above for finding confidence intervals for $\mu_1 - \mu_2$ or for testing hypotheses concerning $\mu_1 - \mu_2$ are not appropriate. The problem of unequal variances is known as the **Behrens-Fisher problem**, and various approximate solutions have been given but are beyond the scope of this course.

Recall



Setting	Estimate	Standard Error	Confidence Interval	Test Statistic	Distribution
Difference Between Population Means					
— σ_1 and σ_2 known	$\bar{x}_1 - \bar{x}_2$	$\mathrm{SE} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\bar{x}_1 - \bar{x}_2 \pm z^* \mathrm{SE}$	$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\mathrm{SE}}$	Normal(0, 1)
— $\sigma_1 = \sigma_2$ unknown	$\bar{x}_1 - \bar{x}_2$	$\widehat{\mathrm{SE}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$\bar{x}_1 - \bar{x}_2 \pm t^* \widehat{\mathrm{SE}}$	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\widehat{\mathrm{SE}}}$	$\mathrm{t}(n_1 + n_2 - 2)$
— $\sigma_1 \neq \sigma_2$ unknown	$\bar{x}_1 - \bar{x}_2$	$\widehat{\mathrm{SE}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$\bar{x}_1 - \bar{x}_2 \pm t^* \widehat{\mathrm{SE}}$	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\widehat{\mathrm{SE}}}$	$\mathrm{t}(f)$

6 Course concluding remarks

Fallacies of Statistical Testing

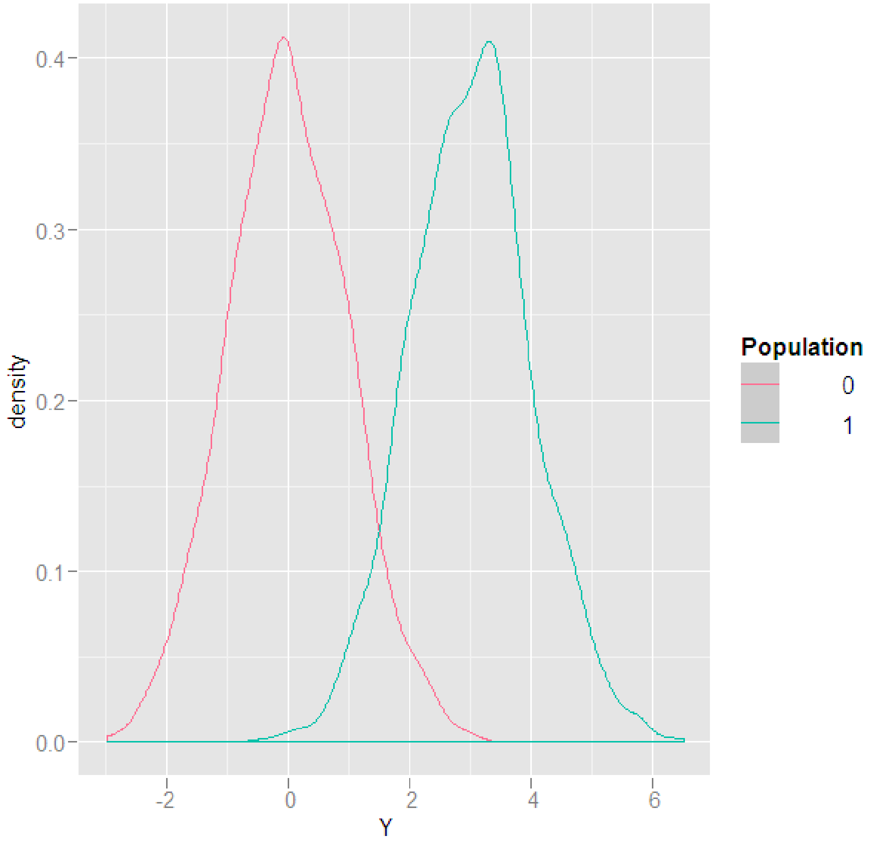
1. Failure to reject the null hypothesis leads to its acceptance. (**WRONG!** Failure ro reject the null hypothesis implies insufficient evidence for its rejection.)
2. The *p* value is the probability that the null hypothesis is incorrect. (**WRONG!** The *p* value is the probability of the current data or data that is more extreme assuming *H*0 is true.)
3. α = .05 is a standard with an objective basis. (**WRONG!** α= .05 is merely a *convention* that has taken on unwise mechanical use.)

Fallacies of Statistical Testing (continued)

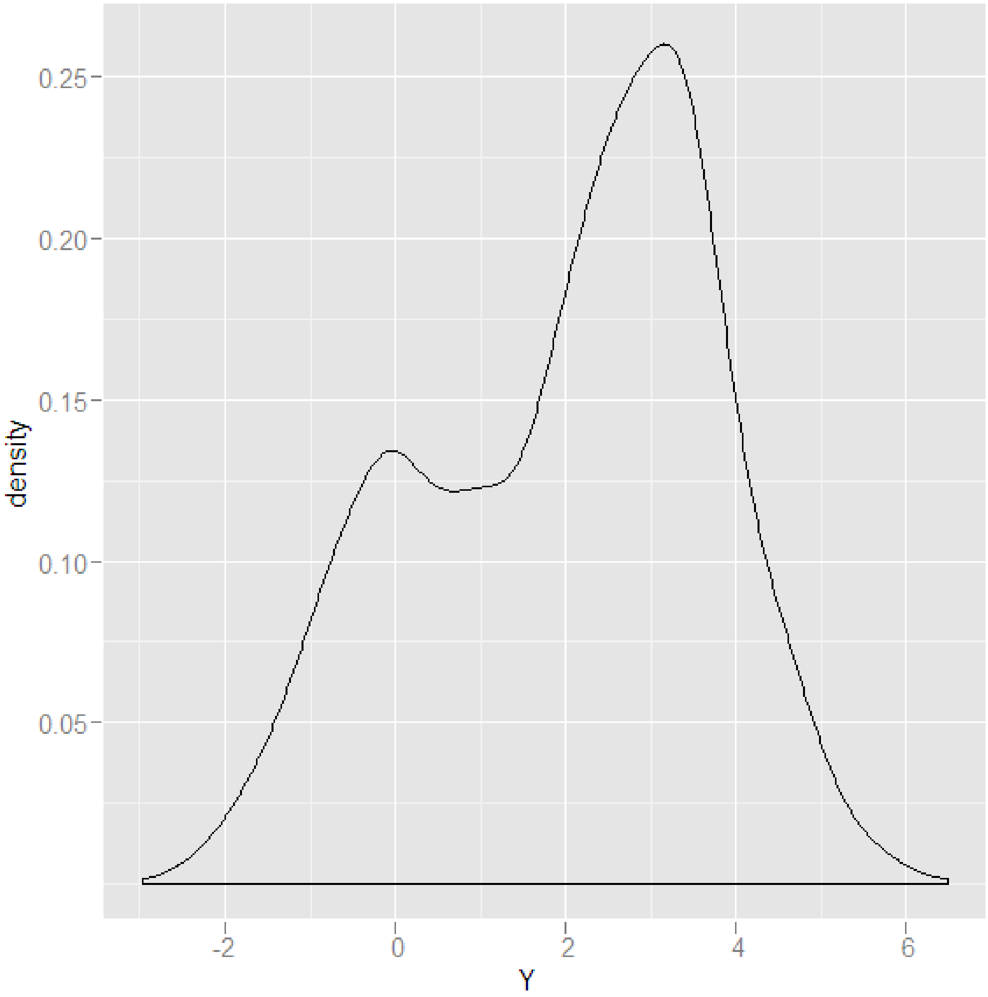
4. Small *p* values indicate large effects. (**WRONG!** *p* values tell you next to nothing about the size of a difference.)
5. Data show a theory to be true or false. (**WRONG!** Data can at best serve to bolster or refute a theory or claim.)
6. Statistical significance implies importance. (**WRONG! WRONG! WRONG!** Statistical significance says very little about the importance of a relation.)

The importance of the normal distribution

~ Central Limit Theorem

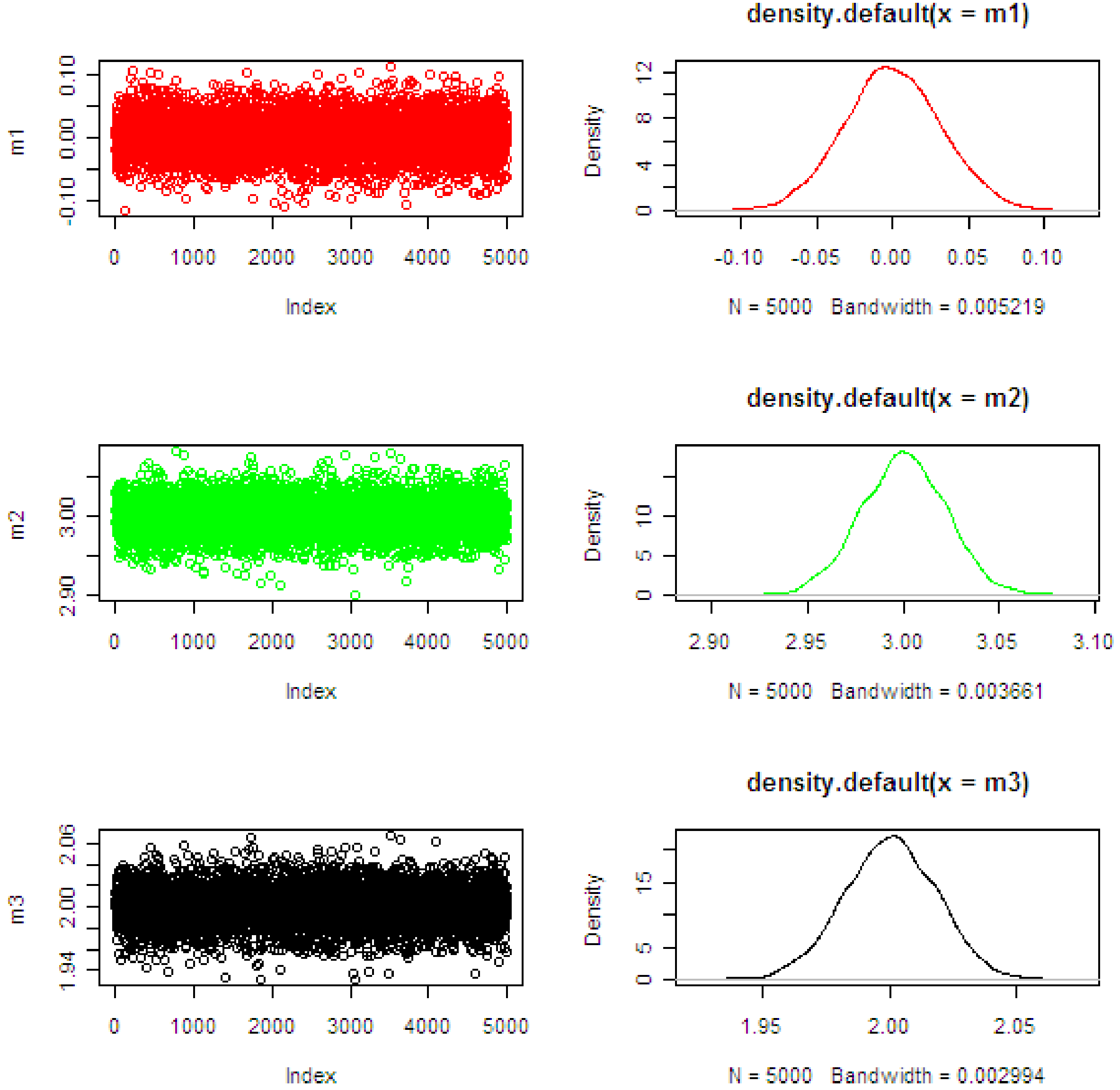


```
set.seed(7)
norm1 <- rnorm(1000)
norm2 <- rnorm(2000,mean=3)
mixnorm <- c(norm1,norm2)

mixnorm <-
data.frame(cbind(mixnorm,c(rep(0,1000),rep
(1,2000))))
names(mixnorm) <- c("Y","Population")

library(ggplot2)
qplot(Y, colour=Population, data=mixnorm,
geom="density")
```



Mixing 2 normal distributions

Sample distributions for the mean of the 3 considered populations



Main reference:

STAT261 Statistical inference notes – School of mathematics, statistics and computer science. University of New England, Oct 4, 2007