UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

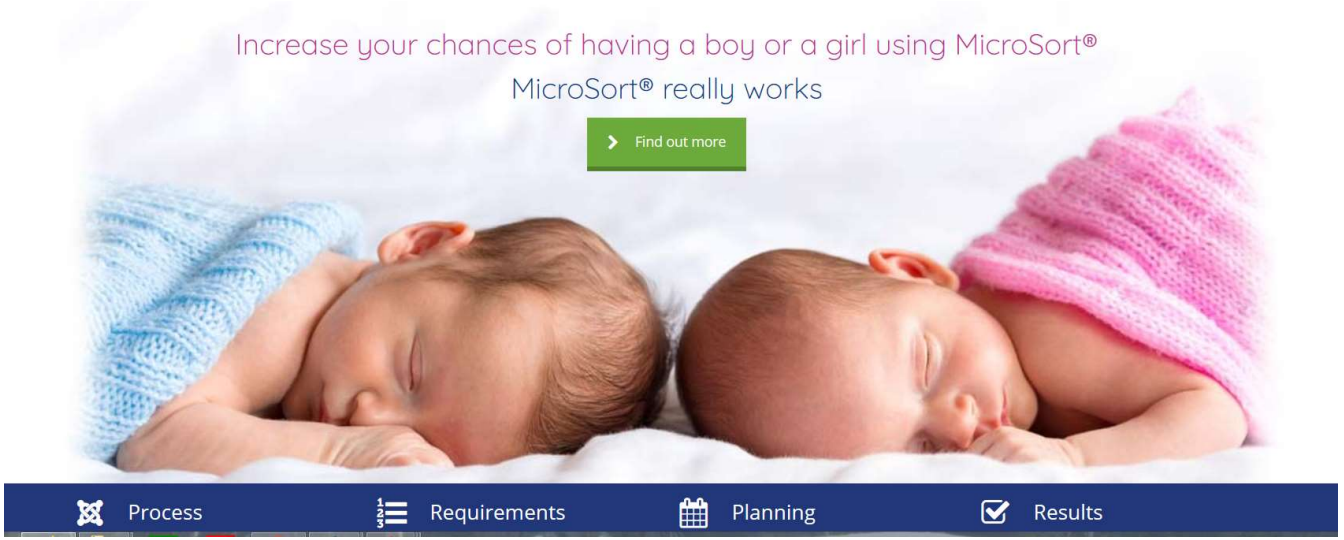
Hypothesis testing

Paola Rebora

Review

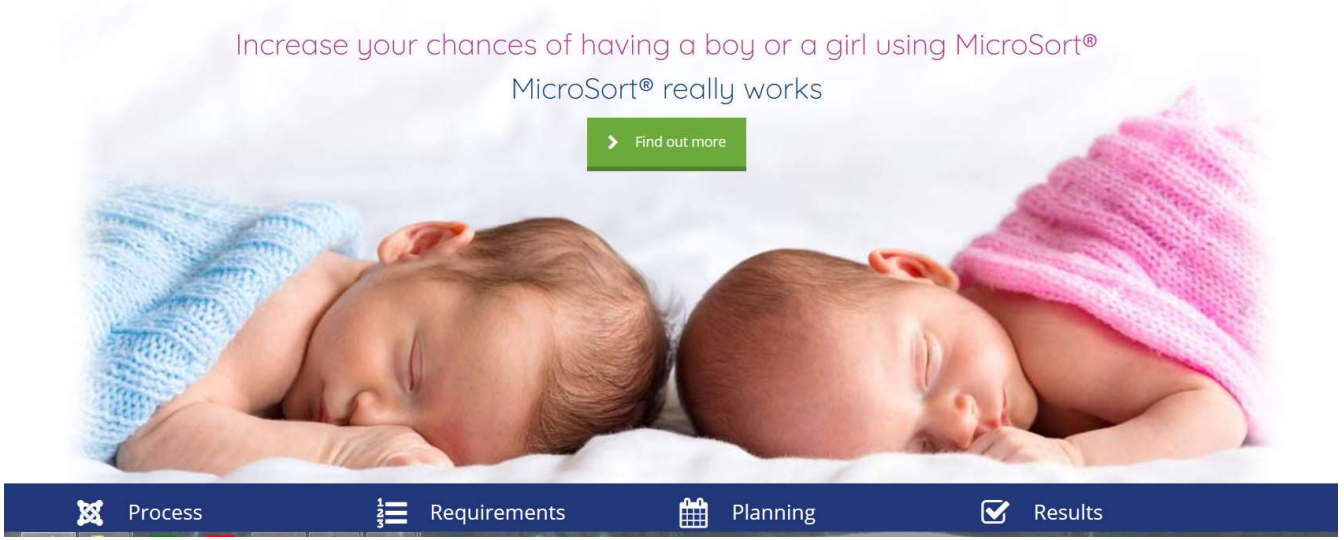
In previous lectures we used "descriptive statistics" when we summarized data using tools such as graphs, and statistics such as the mean and standard deviation. Methods of inferential statistics use sample data to make an inference or conclusion about a population. The two main activities of inferential statistics are using sample data to (1) estimate a population parameter (such as estimating a population parameter with a confidence interval), and (2) test a hypothesis or claim about a population parameter. In last lecture we presented methods for estimating a population parameter with a confidence interval, and in this chapter we present the method of hypothesis testing.

Example: Does the MicroSort Method of Gender Selection Increase the Likelihood That a Baby Will Be a Girl?



The Genetics & IVF Institute claims that its XSORT method allows couples to increase the probability of having a baby girl.

Example: Does the MicroSort Method of Gender Selection Increase the Likelihood That a Baby Will Be a Girl?



Preliminary results:

- 14 babies born to couples using the XSORT method of gender selection
- 13 of the babies were girls.

Under normal circumstances with no special treatment, girls occur in about 50% of births. (Actually, the current birth rate of girls is 48.8%, but we will use 50% to keep things simple.)

Can we actually support the claim that the XSORT technique is effective in increasing the probability of a girl?

Hypothesis test

In statistics, a hypothesis is a claim or statement about a property of a population. A hypothesis test (or test of significance) is a procedure for testing a claim about a property of a population.

- A result of 8 girls (or 57.1%) could easily occur by chance under normal circumstances with no treatment, so 8 is not significantly high.
- The actual result of 13 girls (or 92.9%) appears to be significantly high.

The method of hypothesis testing gives us a standard and widely accepted procedure for deciding whether such results are significant.

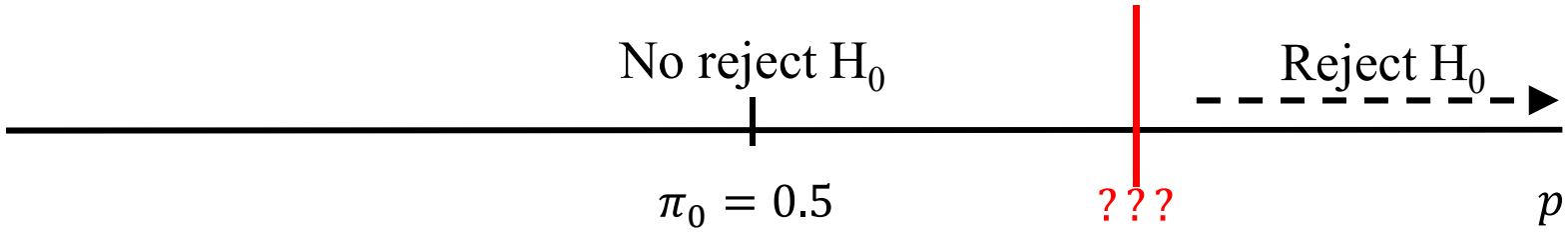
The test to verify a hypothesis is a rule that, based on experimental data, leads to the DECISION OF REJECT or NO the hypothesis under study.

Hypothesis test

The test to verify a hypothesis is a rule that, based on experimental data, leads to the DECISION OF REJECT or NO the hypothesis under study.

1. Identify the claim to be tested and the null and alternative hypothesis to test:
   H_0: $\pi = \pi_0 = 0.5$ (null hypothesis: e.g. *the probabilty to get a girl is 50%*)
   H_1: $\pi > 0.5$ (alternative hypothesis e.g. *the probabilty to get a girl is higher then 50%*)

2. Build a rule that allows to reject the null if sample data are not consistent with the null



Null Hypothesis: H_0

- **The null hypothesis (denoted by H_0) is a statement that the value of a population parameter (such as proportion, mean, or standard deviation) is equal to some claimed value.**
- **We test the null hypothesis directly.**
- **Either reject H_0 or fail to reject H_0.**

Alternative Hypothesis: H_1

- **The alternative hypothesis (denoted by H_1 or H_a or H_A) is the statement that the parameter has a value that somehow differs from the null hypothesis.**
- **The symbolic form of the alternative hypothesis must use one of these symbols: $\neq$, <, >.**

The statistical issue

Even if the true probability of getting a girl is 50%, it is possible that by chance we observe a sample probability which is higher than 50%.

Even if the true probability of getting a girl is higher than 50%, it is however possible that a sample probability is observed that is very close to 50% (or even lower).

↓

In defining the reject region we need to control randomness or the probability of making mistakes and this can be done!

We know the theoretical distribution of the sample probability:

if $n\pi \geq 5$ and $n(1-\pi) \geq 5$, the distribution of p approximates a normal distribution centered on the true value π

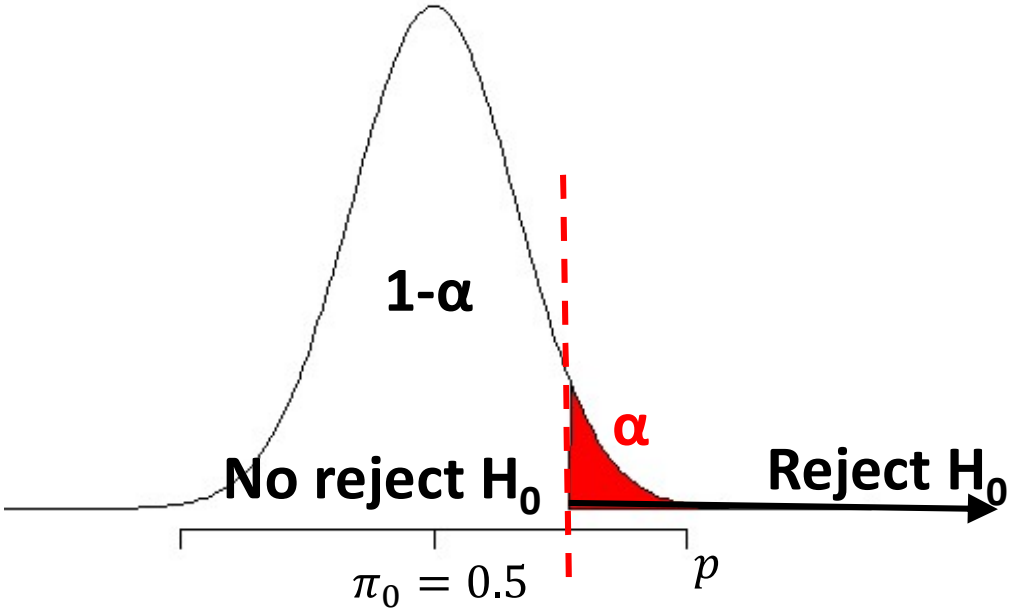
$$p \sim N(\pi, \sqrt{\pi * (1-\pi)/\mathrm{n}})$$

Under the null (H_0): XSORT does not work

Under the null (true probability of getting a girl is 50%,): $\boldsymbol{H_0}: \boldsymbol{\pi = \pi_0 = 0.5}$

the theoretical distribution of the sample probability: $\boldsymbol{p \sim N(0.5, \sqrt{0.5 * 0.5/14})}$

We can then define the **critical rejection region** in order to establish a priori the probability of making mistakes when we reject H_0.
This probability is called significance level α.



α is the level of significance on the basis of which the critical waste region is defined

Under the null (H_0): XSORT does not work

Better standardise: $$z = \frac{p - \pi_0}{se(p|H_0)} \qquad Z \sim N(0,1)$$

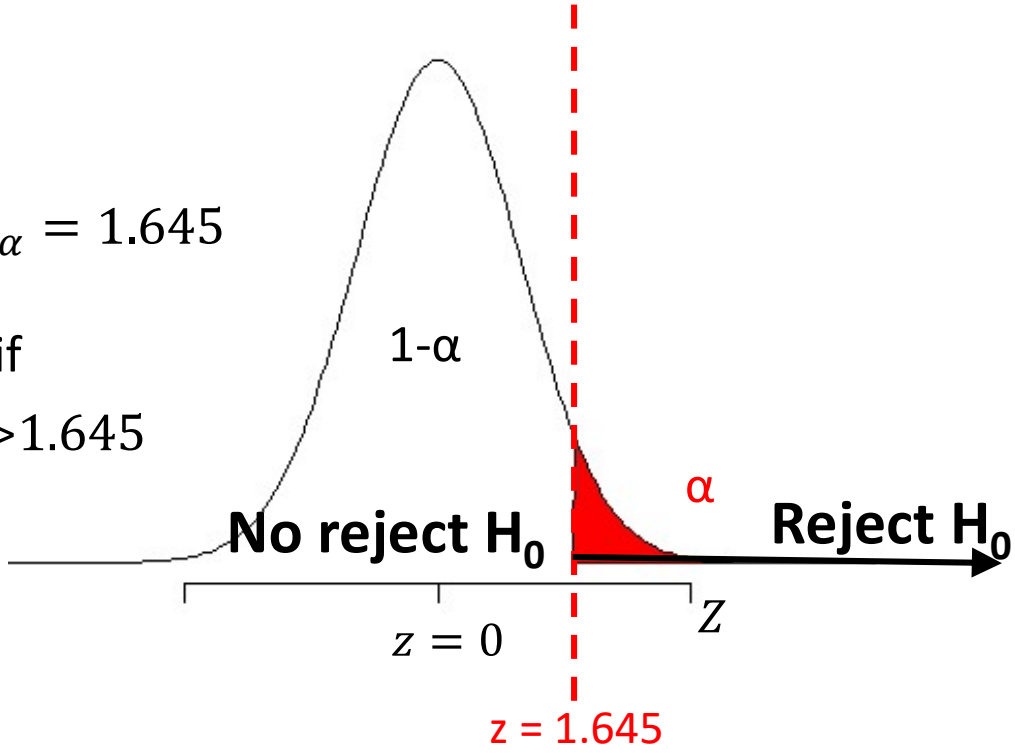
$$z = \frac{p - 0.5}{\sqrt{0.5 * 0.5/14}}$$

When the level of significance α is set, the threshold is the (1-α)th percentile $z_{1-\alpha}$

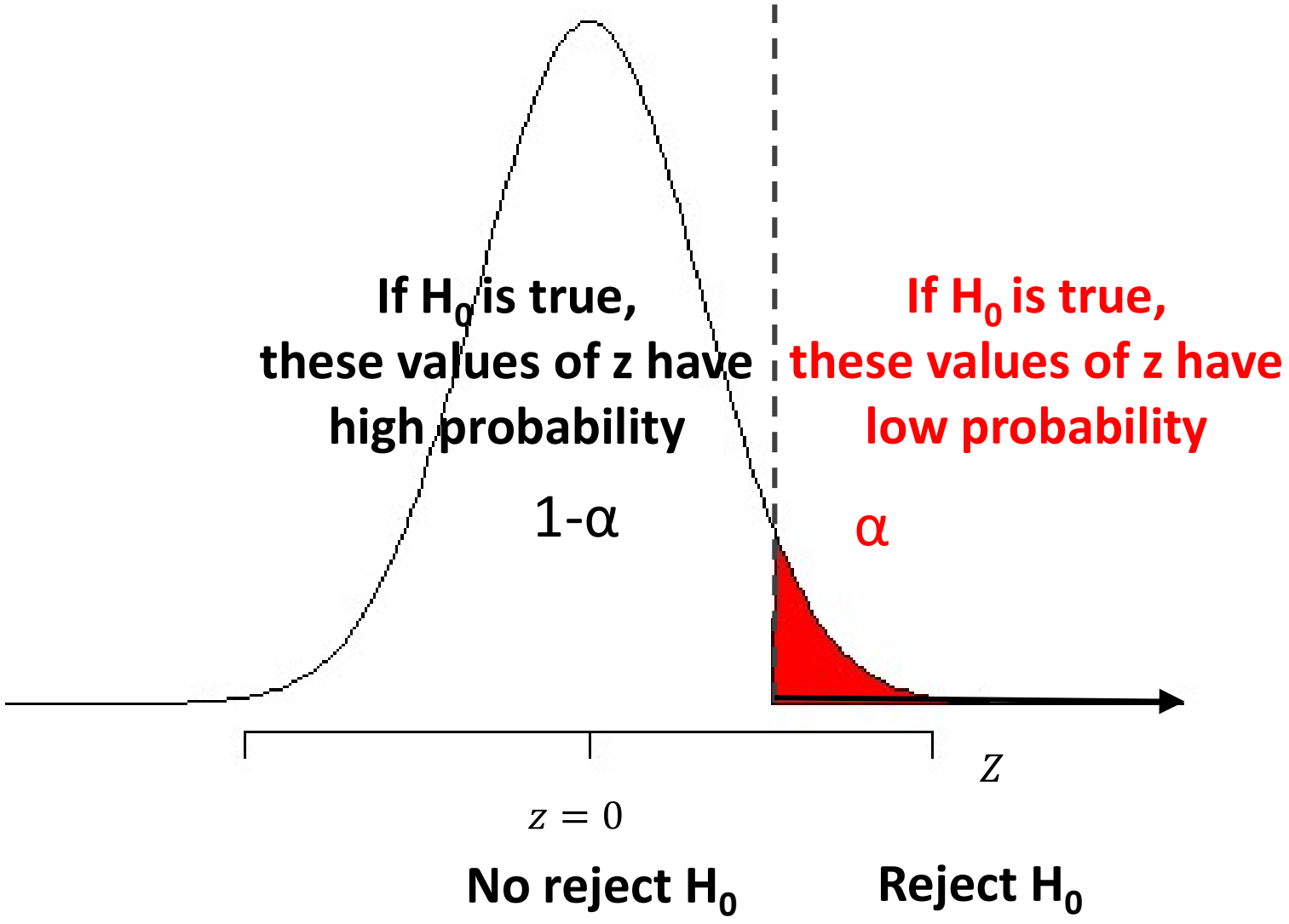
Ex. $\alpha = 0.05$ -> $z_{1-\alpha} = 1.645$

Thus we will reject if

$$z = \frac{p-0.5}{\sqrt{(0.5*0.5)/14}} > 1.645$$



Hypothesis test



But values of z over the threshold are not impossible under the null!

Critical Region

The critical region (or rejection region) is the set of all values of the test statistic that cause us to reject the null hypothesis. For example, see the red-shaded region in the previous figure.

Significance Level

The significance level (denoted by α) is the probability that the test statistic will fall in the critical region when the null hypothesis is actually true. This is the same α introduced with confidence intervals. Common choices for α are 0.05, 0.01, and 0.10.

Example

$n=14$

$p=13/14=0.929$

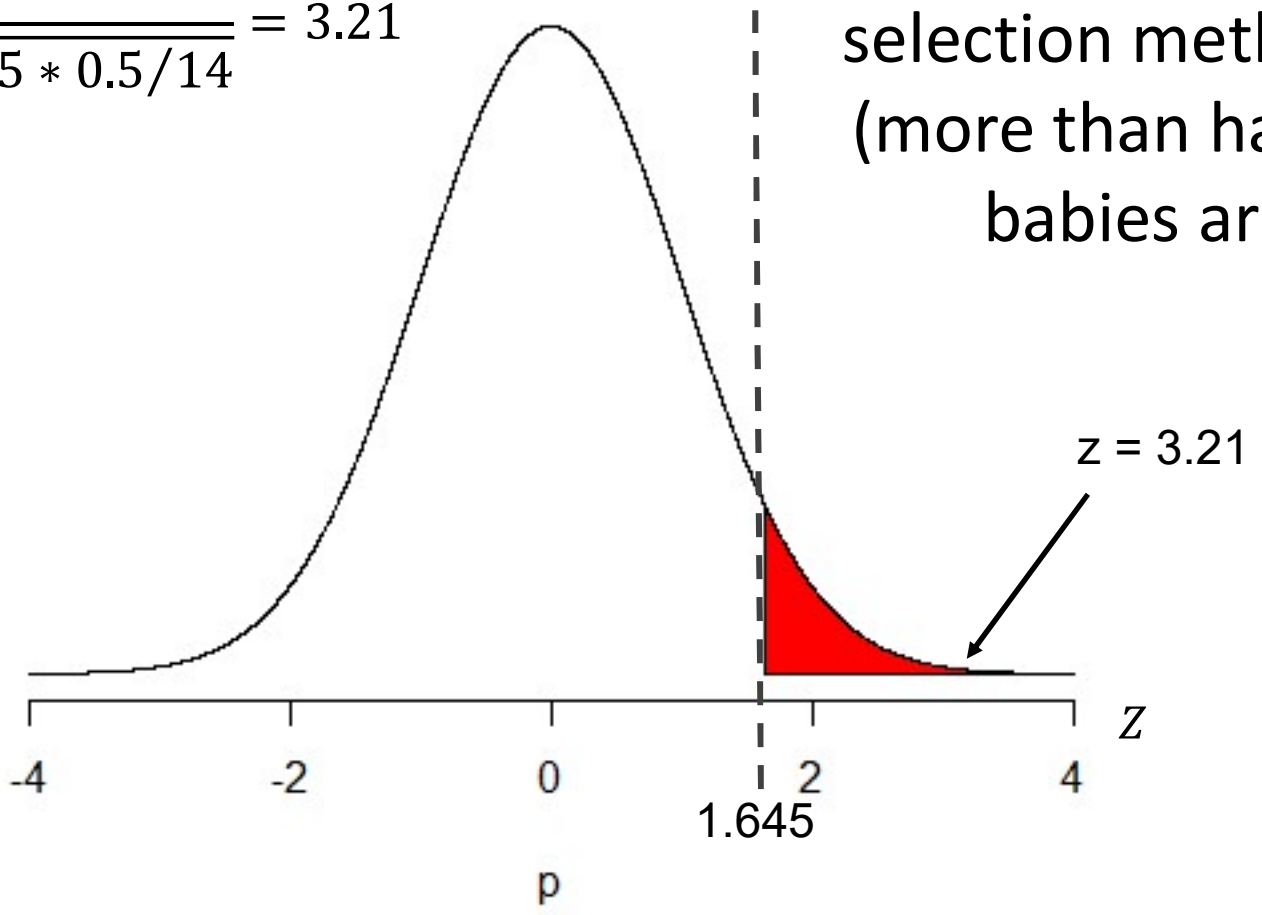
$\alpha = 0.05 \;\; z_{0.95} = 1.645$

$$z = \frac{p-\pi_0}{se(p)}$$

$$z = \frac{0.929 - 0.5}{\sqrt{0.5 * 0.5/14}} = 3.21$$

Reject the null hypothesis!

There is sufficient sample evidence to support the claim that for couples using the «XSORT gender selection method», most (more than half) of their babies are girls.

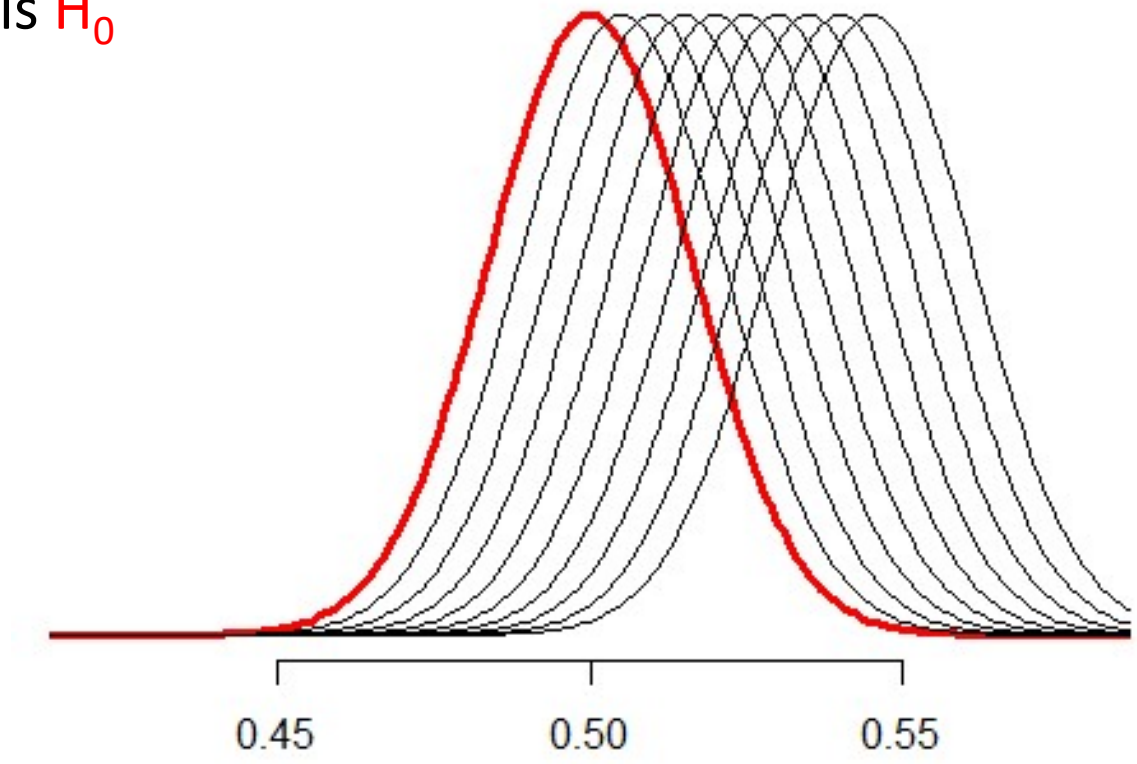


Hypothesis test: the method

Mathematical logic	Statistical logic	Operative
Hypothesis	Null hypothesis H_0	Formulate H_0
↓	↓	↓
logical-deductive argument	Statistical test and sampling distribution under H_0	Find the reject region
↓	↓	↓
Contradiction (reductio ad absurdum)	"Distance" between the sample result and expected theoretically	Compute statistical test and evaluate plausibility of H_0 given sample data
↓	↓	↓
Conclusion (thesis) The hypothesis was false!	Conclusion : H_0 is true and the distance is random H_0 is false and the distance is due to treatment	Conclusion : Not reject H_0 Reject H_0

The logic of the hypothesis test:

The statistical hypthesis test is based on the disproof of a specific hypothesis H_0



H_0: π=0.5
Under a specific single hypothesis is possible to find the sampling distribution

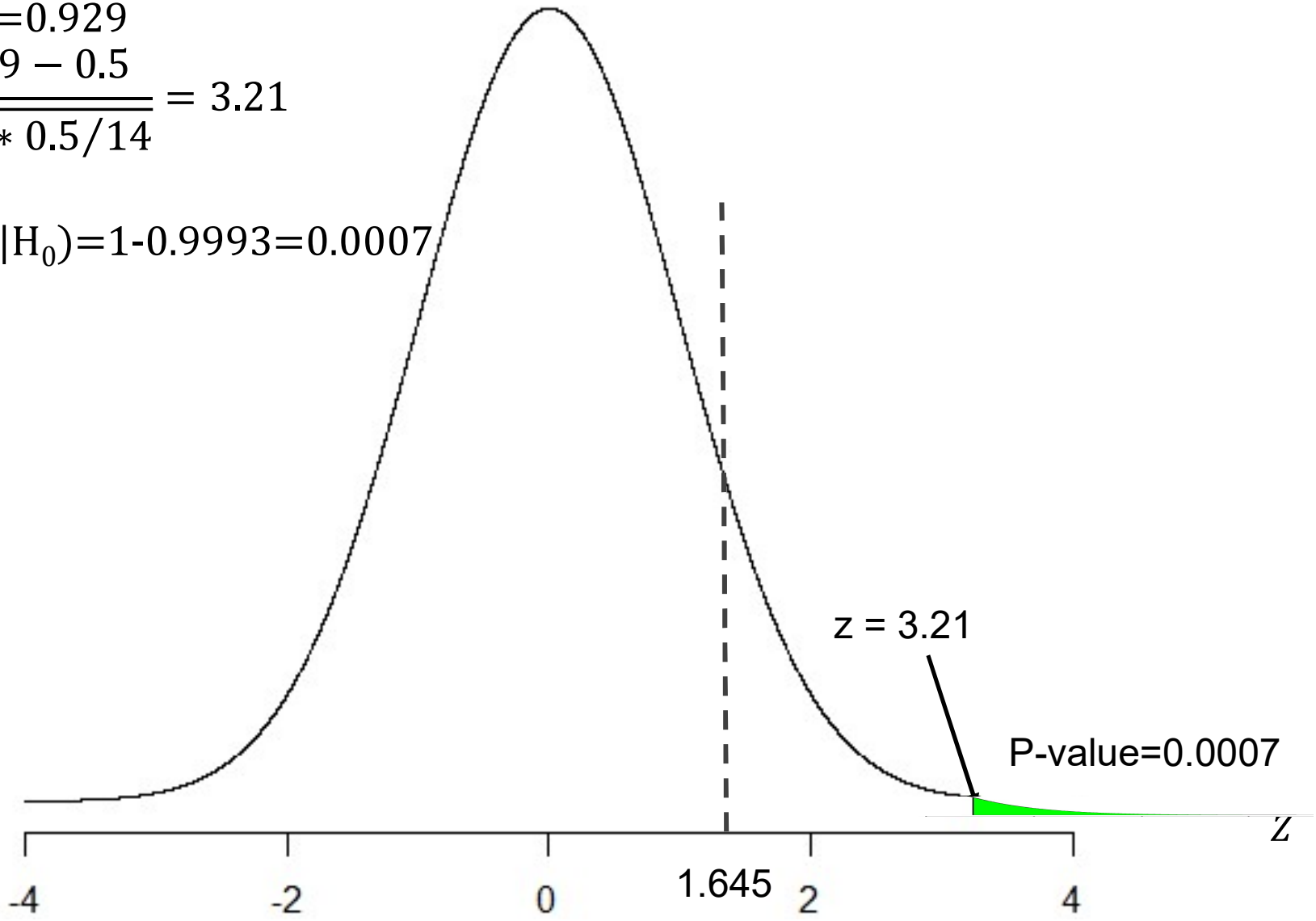
H_1: π>0.5
The alternative hypothesis includes an infinity of values and their related sampling distributions

Example

$p=13/14=0.929$

$$z = \frac{0.929 - 0.5}{\sqrt{0.5 * 0.5/14}} = 3.21$$

$P(Z>3.21|H_0)=1\text{-}0.9993=0.0007$



P-Value

The *P*-value (or *p*-value or probability value) is the probability of getting a value of the test statistic that is at least as extreme as the one representing the sample data, assuming that the null hypothesis is true.

Critical region in the left tail:	***P*-value = area to the left of the test statistic**
Critical region in the right tail:	***P*-value = area to the right of the test statistic**
Critical region in two tails:	***P*-value = twice the area in the tail beyond the test statistic**

P-Value

The null hypothesis is rejected if the *P*-value is very small, such as 0.05 (1 out of 20) or less.

Here is a memory tool useful for interpreting the P-value:

If the P is low, the null must go.
If the P is high, the null will fly.

The P-value expresses the force of evidence against the null hypothesis

Caution

Don't confuse a *P*-value with a proportion *p*. Know this distinction:

***P*-value = probability of getting a test statistic at least as extreme as the one representing sample data**

π= population proportion

***p* = sample proportion**

The P-value

P-value = probability of getting a test statistic at least as extreme as the one representing the sample data, assuming that the null hypothesis H_0 is true

NOT the probability that H_0 is true!

The P-value expresses the force of evidence against the null hypothesis

P-Value Method
With the P-value method of testing hypotheses, we make a decision by comparing the P-value to the significance level α.

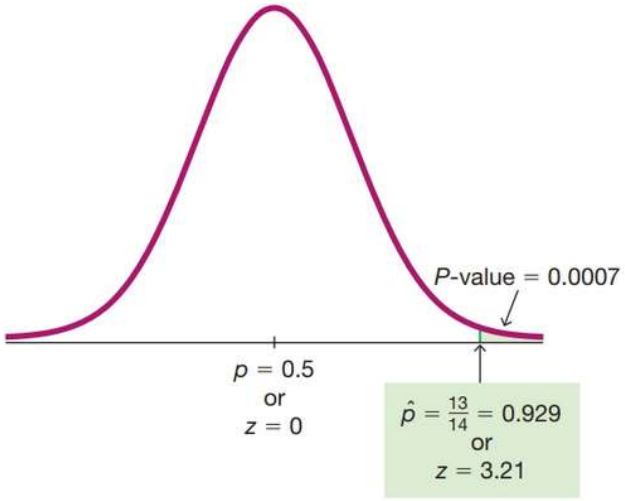
- Reject H_0 if P-value $\leq \alpha$.
- Fail to reject H_0 if P-value $> \alpha$.

Example XSORT

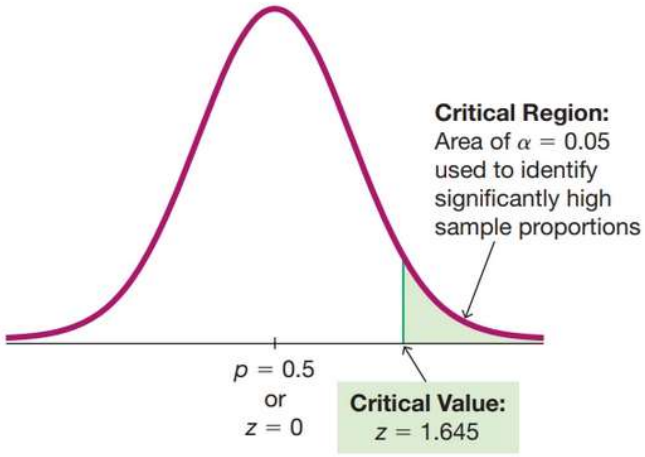
$$p=13/14=0.929$$

$$z = \frac{0.929 - 0.5}{\sqrt{0.5 * 0.5/14}} = 3.21$$

P-value method:



Critical value method:



If the XSORT method would not be effective, a sample result equal to or more extreme (in the tail of the distribution) than that observed in the sample (13 girls out of 14 births) would occur 7 times out of 10000.

It is **possible** that an ineffective method will provide such result, however ... is **unlikely** (p≈0.0007)

It is **more plausible** that the new method rise the probability of getting a girl!

The experiment suggests that the method is effective

Types of Hypothesis Tests: Two-tailed, Left-tailed, Right-tailed

The tails in a distribution are the extreme regions bounded by critical values.

Determinations of *P*-values and critical values are affected by whether a critical region is in two tails, the left tail, or the right tail. It therefore becomes important to correctly characterize a hypothesis test as two-tailed, left-tailed, or right-tailed.

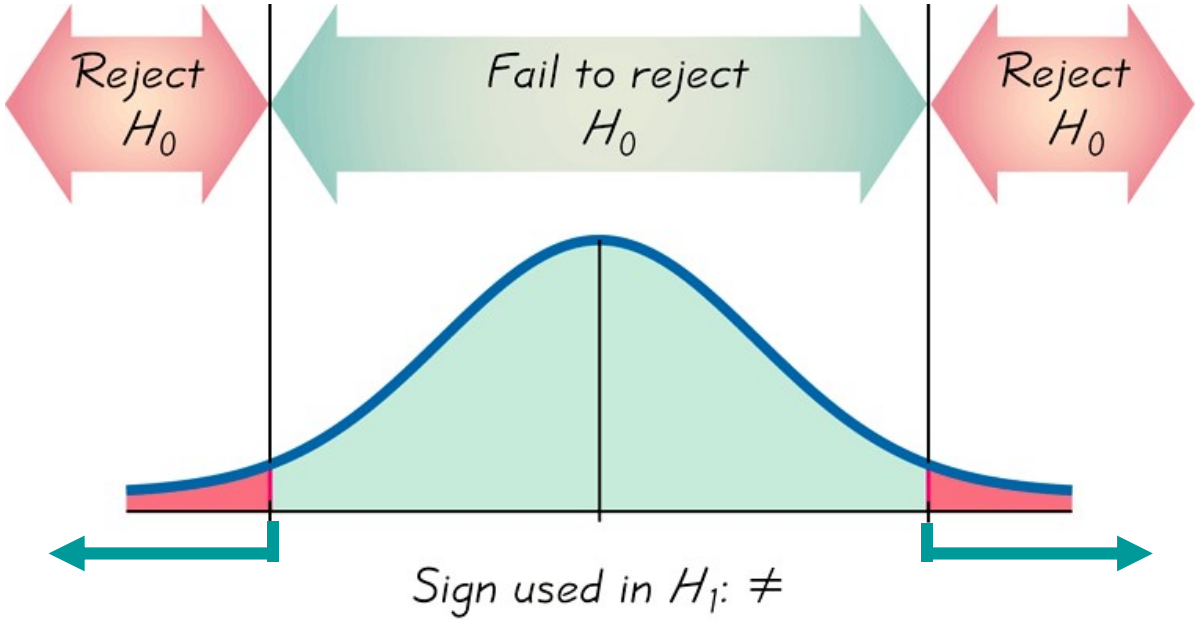
Two-tailed Test

H_0: =

H_1: ≠

α **is divided equally between the two tails of the critical region**

Means less than or greater than



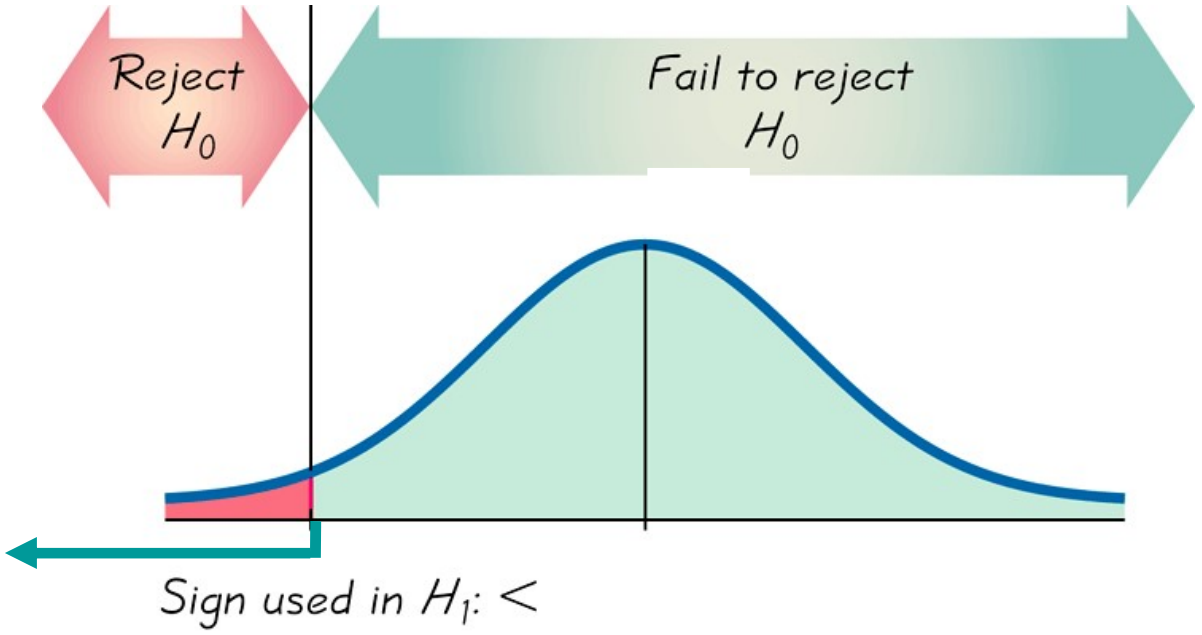
Left-tailed Test

H_0: =

H_1: <

α the left tail

Points Left

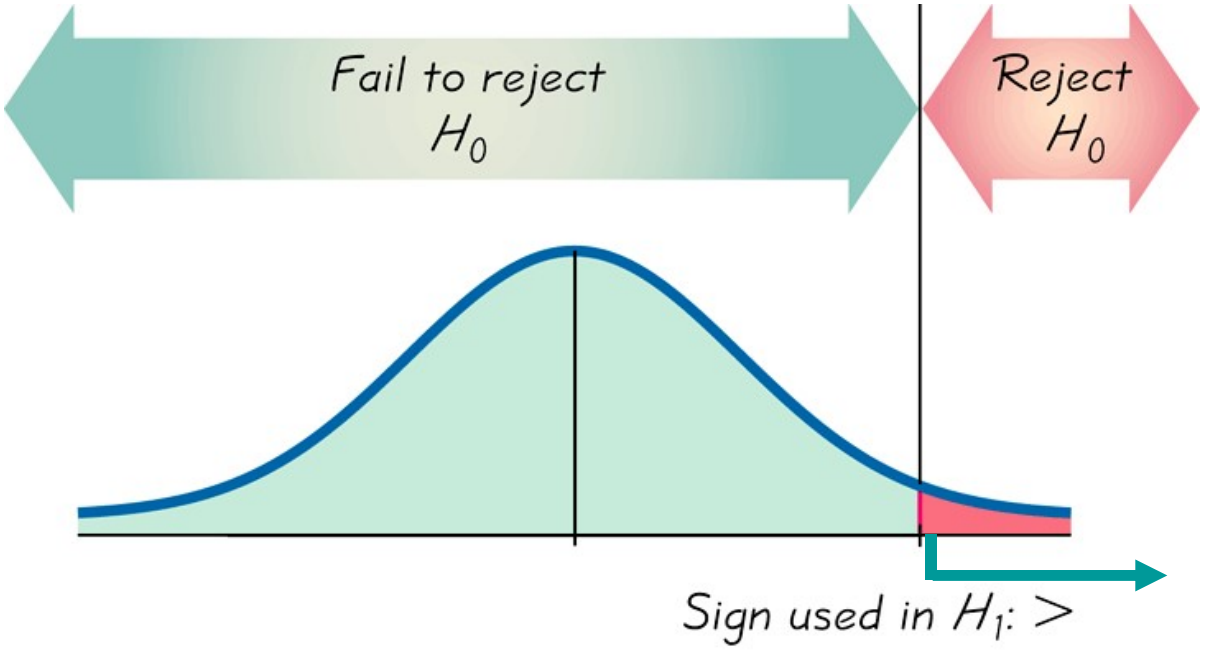


Right-tailed Test

H_0: =

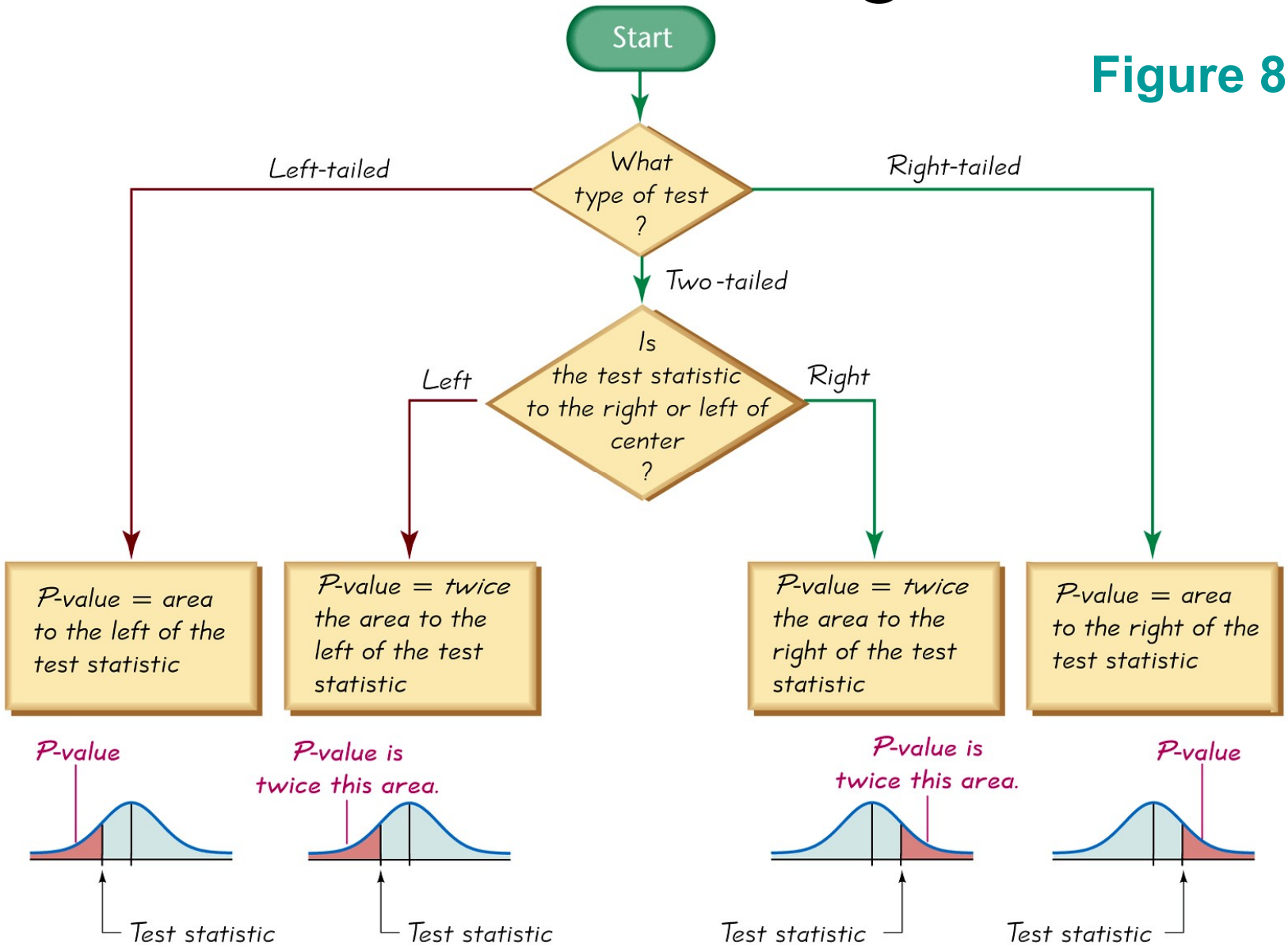
H_1: >

Points Right



Procedure for Finding *P*-Values

Figure 8-5



Conclusions in Hypothesis Testing

We always test the null hypothesis. The initial conclusion will always be one of the following:

1. Reject the null hypothesis.

2. Fail to reject the null hypothesis.

Decision Criterion

P-value method:

Using the significance level α:

If *P*-value $\leq \alpha$, reject H_0.

If *P*-value > α , fail to reject H_0.

Decision Criterion

Traditional method:

If the test statistic falls within the critical region, reject H_0.

If the test statistic does not fall within the critical region, fail to reject H_0.

Decision Criterion

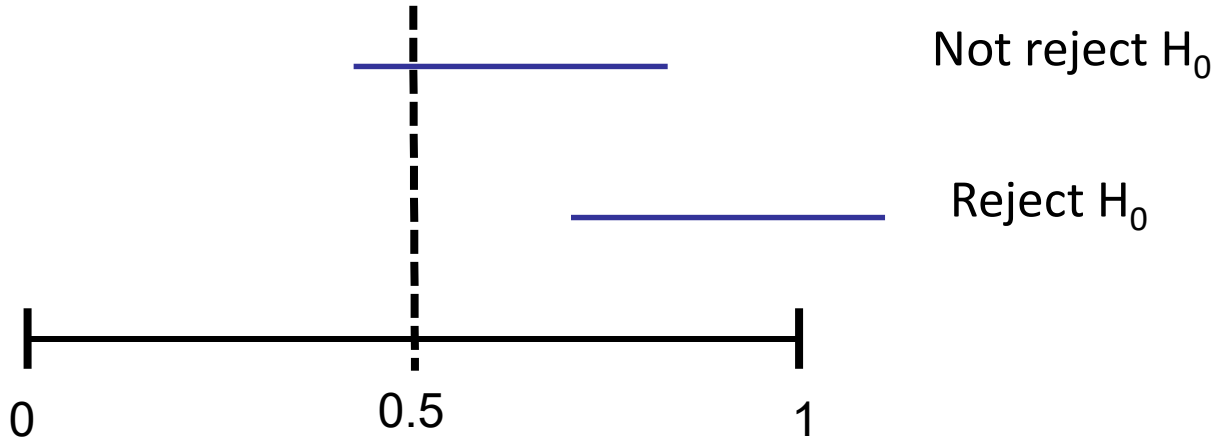
Confidence Intervals method:

A confidence interval estimate of a population parameter contains the likely values of that parameter.

If a confidence interval does not include a claimed value of a population parameter, reject that claim.

The confidence interval method

Because a confidence interval estimate of a population parameter contains the likely values of that parameter, reject a claim that the population parameter has a value that is not included in the confidence interval.

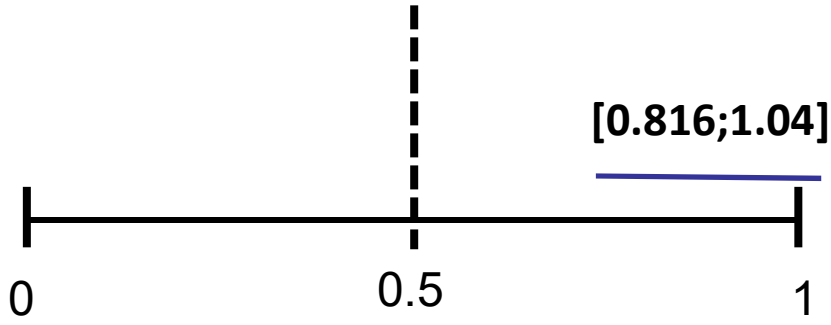


Example

$$p=13/14=0.929$$

$$p \pm z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}$$

$$0.929 \pm 1.645\sqrt{\frac{0.929(1-0.929)}{14}} \text{ -> } [0.816;1.04]$$



The entire range of values in this confidence interval is greater than 0.5.

Because we are 90% confident that the limits of 0.816 and 1.04 contain the true value of π, the sample data appear to support the claim that most (more than 0.5) XSORT babies are girls.

In this case, the conclusion is the same as with the P-value method and the critical value method, but that is not always the case. *It is possible that a conclusion based on the confidence interval can be slightly different from the conclusion based on the P-value method or critical value method.*

Example: Does Touch Therapy Work?

Touch Therapy: Structured and standardized healing practice performed by practitioners trained to be sensitive to the receiver's energy field that surrounds the body...no touching is required.

Emily' science fair project:

Each touch therapist would put both hands through the two holes, and Emily would place her hand just above one of the therapist's hands; then the therapist was asked to identify the hand that Emily had selected. Emily used a coin toss to randomly select the hand to be used. This test was repeated 280 times.

Among the 280 trials, the touch therapists identified the correct hand 123 times.

A Close Look at Therapeutic Touch" Journal of the American Medical Association, Vol. 279, No. 13

Exercise: Does Touch Therapy Work?

Emily' science fair project:

If the touch therapists really did have the ability to sense a human energy field, they should have identified the correct hand significantly more than 50% of the time. If they did not have the ability to detect the energy field and they just guessed, they should have been correct about 50% of the time.

Does the Touch Therapy Work?

A Close Look at Therapeutic Touch" Journal of the American Medical Association, Vol. 279, No. 13

Exercise: Does Touch Therapy Work?

$H_0: \pi = \pi_0 = 0.5$
$H_1: \pi > 0.5$
Let's use a significance level of 0.01, so the critical value will be $z_{0.99}$=2.33

Among the 280 trials, the touch therapists identified the correct hand 123 times, so p=123/280=0.4393

1. Critical value method : $z = \frac{\mathbf{0.4393-0.5}}{\sqrt{\mathbf{0.5*0.5/280}}} = \mathbf{-2.03}$

Since -2.03 is lower than the critical value we cannot reject H_0

2. P- value

P-value=0.97882 If the touch therapists just guessed (as they do not have the ability to detect the energy field), a sample result equal to or more than that observed in the sample would occur 98 times out of 100.

3. Confidence interval

$0.4393 \pm 2.33\sqrt{\frac{0.4393(1-0.4393)}{280}}$ -> $[0.3702;0.5084]$

A Close Look at Therapeutic Touch" Journal of the American Medical Association, Vol. 279, No. 13

Exercise

In a study 57 out of 104 pregnant women correctly guessed the sex of their babies. Use these sample data to test the claim that the success rate of such guesses is no different from the 50% success rate expected with random chance guesses. Use a 0.05 significance level.

Decision criteria: type I and II errors

When testing a null hypothesis, we arrive at a conclusion of rejecting it or failing to reject it. Our conclusions are sometimes correct and sometimes wrong (even if we apply all procedures correctly).

	If H_0 is true	If H_1 is true
and based on the sample…I do not reject H_0	**Correct decision** protection: $(1-\alpha)$	**Wrong decision** Fail to reject a false H_0 P(Type II error)= β
and based on the sample…I do reject H_0	**Wrong decision** Reject a true H_0 P(Type I error)=α	**Correct decision** protection: $(1-\beta)$

Type I Error

- **A Type I error is the mistake of rejecting the null hypothesis when it is actually true.**
- **The symbol α (alpha) is used to represent the probability of a type I error.**

Decision criteria: type II error

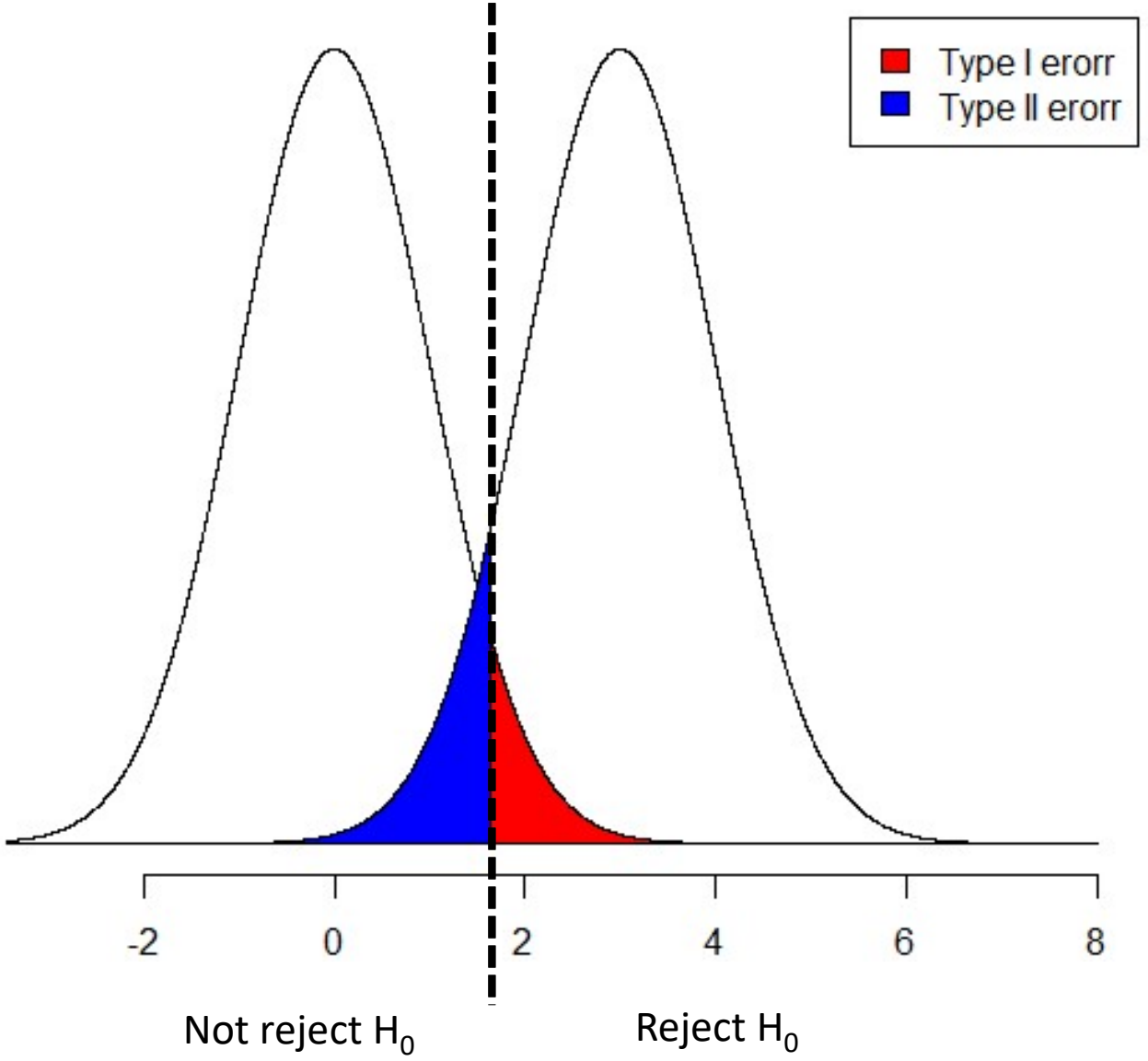
Based on the sample	If H_0 is true	If H_1 is true
Not reject H_0	**Correct decision** protection: $(1-\alpha)$	**Wrong decision** Fail to reject a false H_0 P(Type II error)= β
Reject H_0	**Wrong decision** Reject a true H_0 P(Type I error)=α	**Correct decision** protection: $(1-\beta)$

Type II Error

- **A Type II error is the mistake of failing to reject the null hypothesis when it is actually false.**
- **The symbol β (beta) is used to represent the probability of a type II error.**

Type I and II errors

$H_0: \pi = \pi_0 = 0.5$ $H_1: \pi > \pi_0 = 0.5$



Example:

Assume that we are conducting a hypothesis test of the claim that a method of gender selection increases the likelihood of a baby girl, so that the probability of a baby girls is $\pi > 0.5$. Here are the null and alternative hypotheses: H_0: $\pi = 0.5$, and H_1: $\pi > 0.5$.

a) **Identify a type I error.**
b) **Identify a type II error.**

Example:

a) A type I error is the mistake of rejecting a true null hypothesis, so this is a type I error: Conclude that there is sufficient evidence to support $\pi > 0.5$, when in reality $\pi = 0.5$.

b) A type II error is the mistake of failing to reject the null hypothesis when it is false, so this is a type II error: Fail to reject $\pi = 0.5$ (and therefore fail to support $\pi > 0.5$) when in reality $\pi > 0.5$.

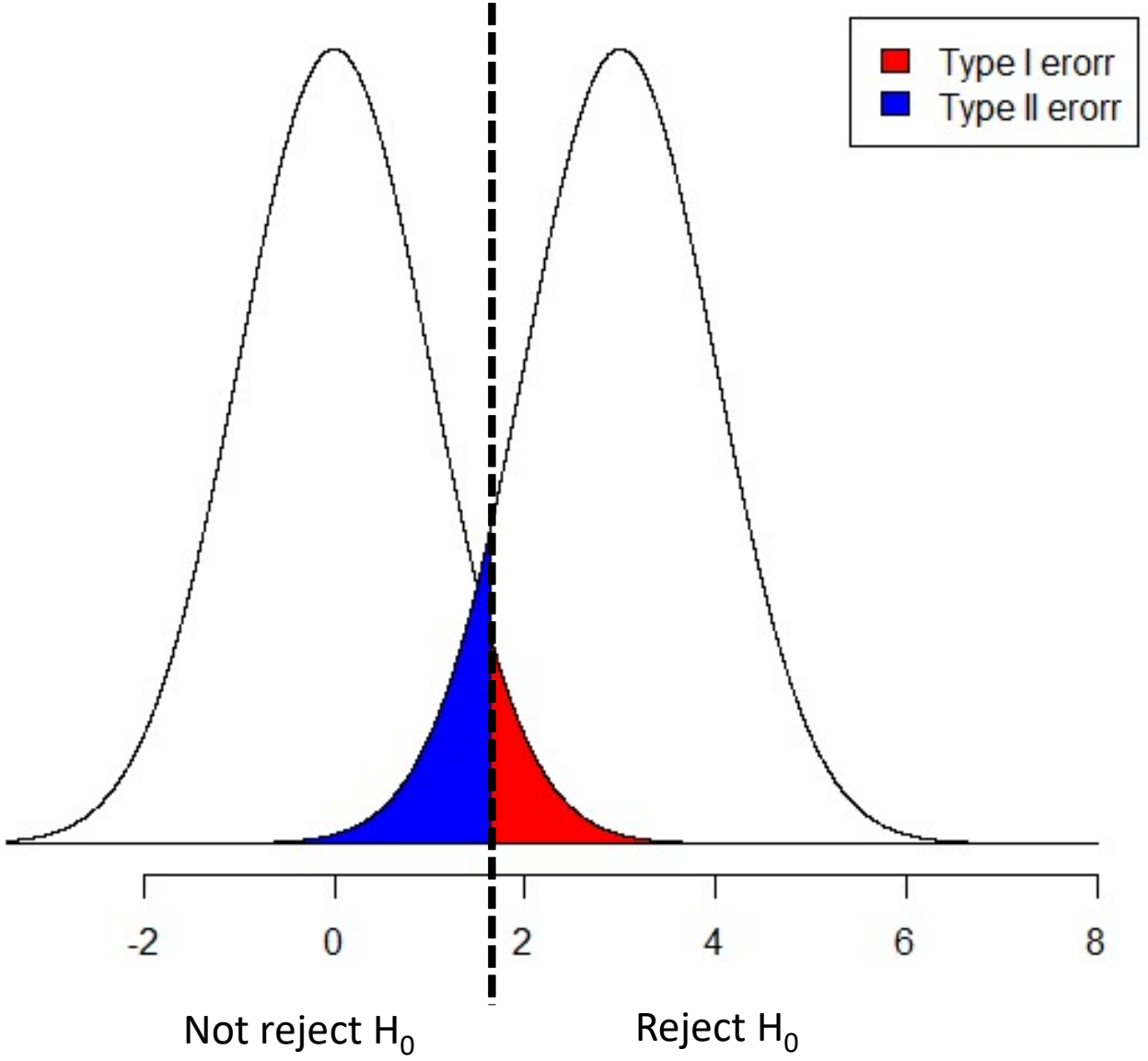
Controlling Type I and Type II Errors

- **For any fixed sample size *n*, a decrease in α will cause an increase in β. Conversely, an increase in α will cause a decrease in β.**

- **To decrease both α and β, increase the sample size.**

Type I and II errors

$H_0: \pi = \pi_0 = 0.5$ $H_1: \pi > \pi_0 = 0.5$



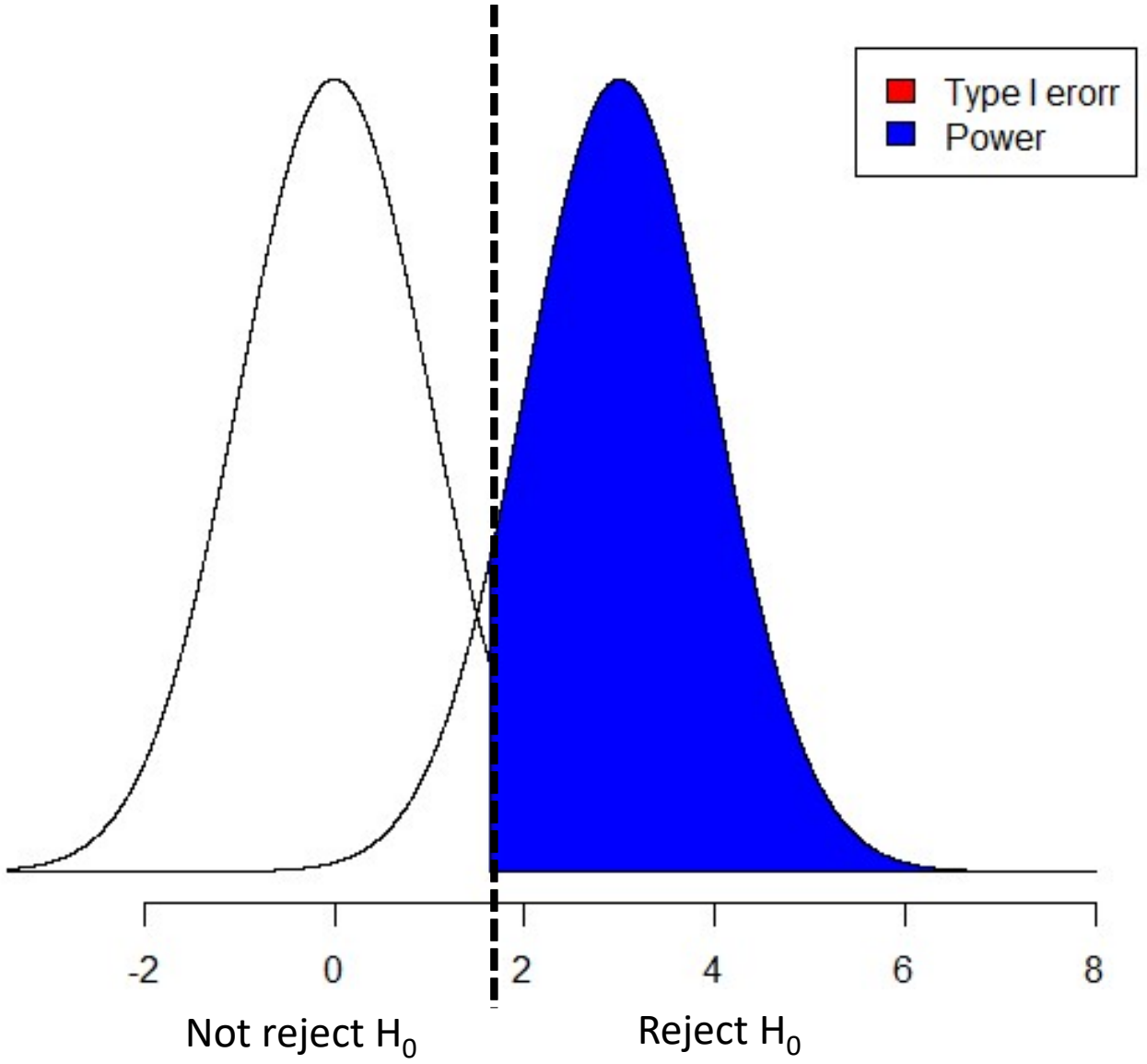
Power

The power of a hypothesis test is the probability 1 - β of rejecting a false null hypothesis. The value of the power is computed by using a particular significance level α and a particular value of the population parameter that is an alternative to the value assumed true in the null hypothesis.

Because determination of power requires a particular value that is an alternative to the value assumed in the null hypothesis, a hypothesis test can have many different values of power, depending on the particular values of the population parameter chosen as alternatives to the null hypothesis.

Power

H_0: $\pi = \pi_0 = 0.5$ H_1: $\pi > \pi_0 = 0.5$



Power and the Design of Experiments

Just as 0.05 is a common choice for a significance level, a power of at least 0.80 is a common requirement for determining that a hypothesis test is effective.

When designing an experiment, we might consider how much of a difference between the claimed value of a parameter and its true value is an important amount of difference. When designing an experiment, a goal of having a power value of at least 0.80 can often be used to determine the minimum required sample size.

Type I error and Power

Type I error (α):

Probability of rejecting H_0 when it is true H_0

e.g. it is concluded that B is better (or worse) than A when in reality it is not (treatments do not differ).

Usually it is fixed ≤ 5%

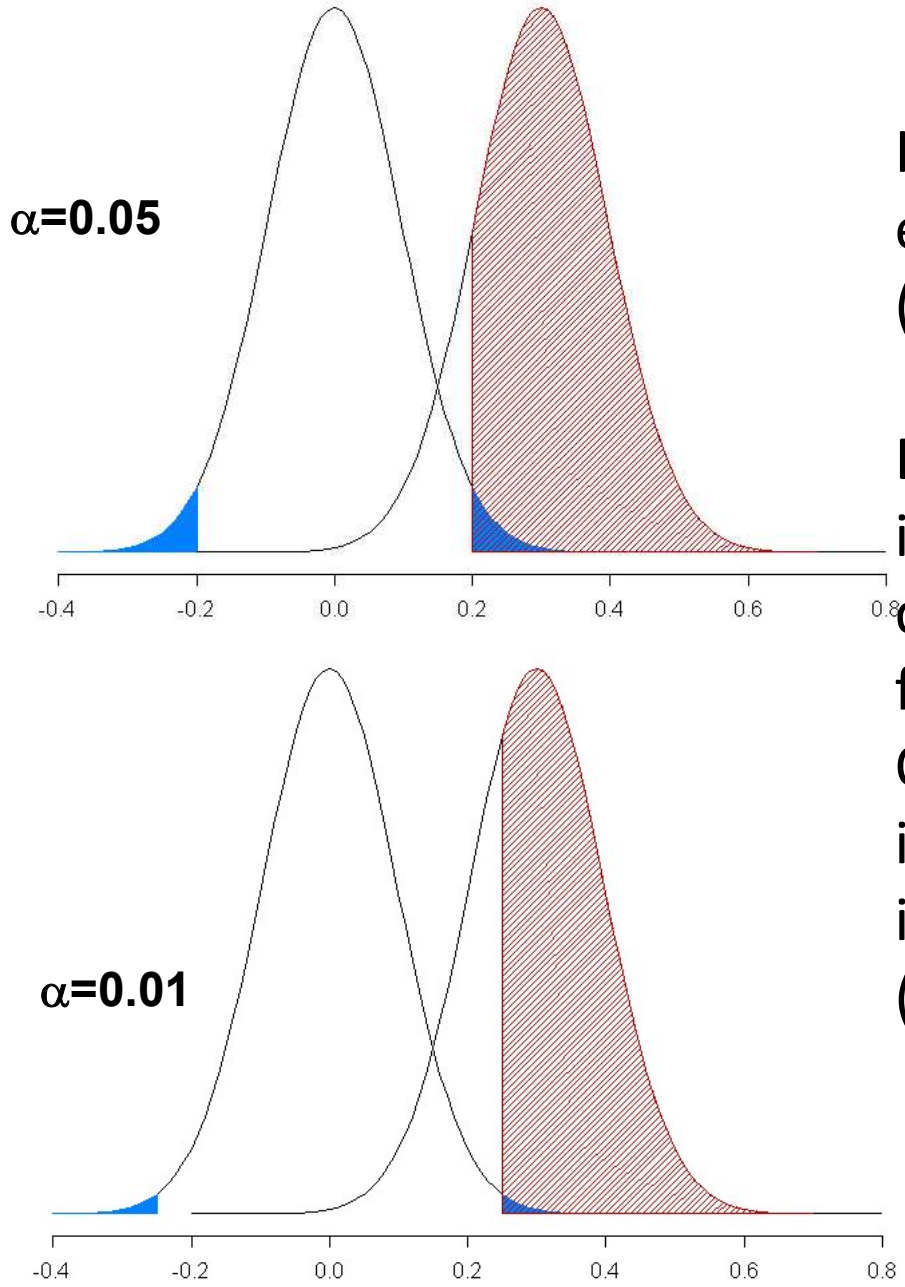
Power (1-β):

Probability of rejecting H_0 when it is true a specification H_1

e.g. it is concluded that B differs from A when actually B is better or worse than A.

It is usually fixed ≥ 80%

Type I error and Power



It would be optimal to have the error α close to 0 and the power (1-β) close to 1!

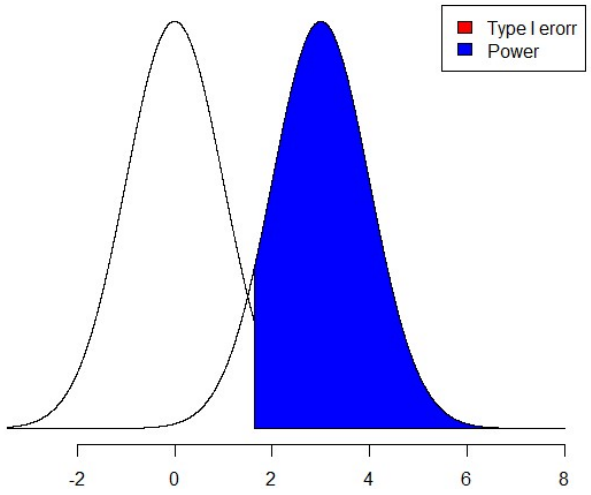
But it happens that:
if α decreases, then the power decreases as well as shown in the figure (from top to bottom)
Or:
if the power increases, then α increases
(from bottom to top)

Example: calculate the power

Consider the preliminary results from the XSORT method of gender selection:
n=14
H_0: π = 0.5 H_1:π > 0.5
Let's use a significance level of α = 0.05.

In addition to all given test components, finding power requires that we select a particular value of π that is an alternative to the value assumed in the null hypothesis H_0: π = 0.5.

Find the values of power corresponding to these alternative values of π : 0.6, 0.7, 0.8, and 0.9.



Example: calculate the power (by STATA)

H_0: $\pi = 0.5$ H_1:$\pi > 0.5$

$\alpha = 0.05$. Find the values of power corresponding to these alternative values of p: 0.6, 0.7, 0.8, and 0.9.

Specific Alternative Value of p	β	Power of Test $= 1 - \beta$
0.6	0.820	0.180
0.7	0.564	0.436
0.8	0.227	0.773
0.9	0.012	0.988

INTERPRETATION

On the basis of the power values listed above, we see that this hypothesis test has a power of 0.180 (or 18.0%) of rejecting H_0: $p = 0.5$ when the population proportion p is actually 0.6. That is, if the true population proportion is actually equal to 0.6, there is an 18.0% chance of making the correct conclusion of rejecting the false null hypothesis that $p = 0.5$. That low power of 18.0% is not so good.

There is a 0.436 probability of rejecting $p = 0.5$ when the true value of p is actually 0.7. It makes sense that this test is more effective in rejecting the claim of $p = 0.5$ when the population proportion is actually 0.7 than when the population proportion is actually 0.6. (When identifying animals assumed to be horses, there's a better chance of rejecting an elephant as a horse—because of the greater

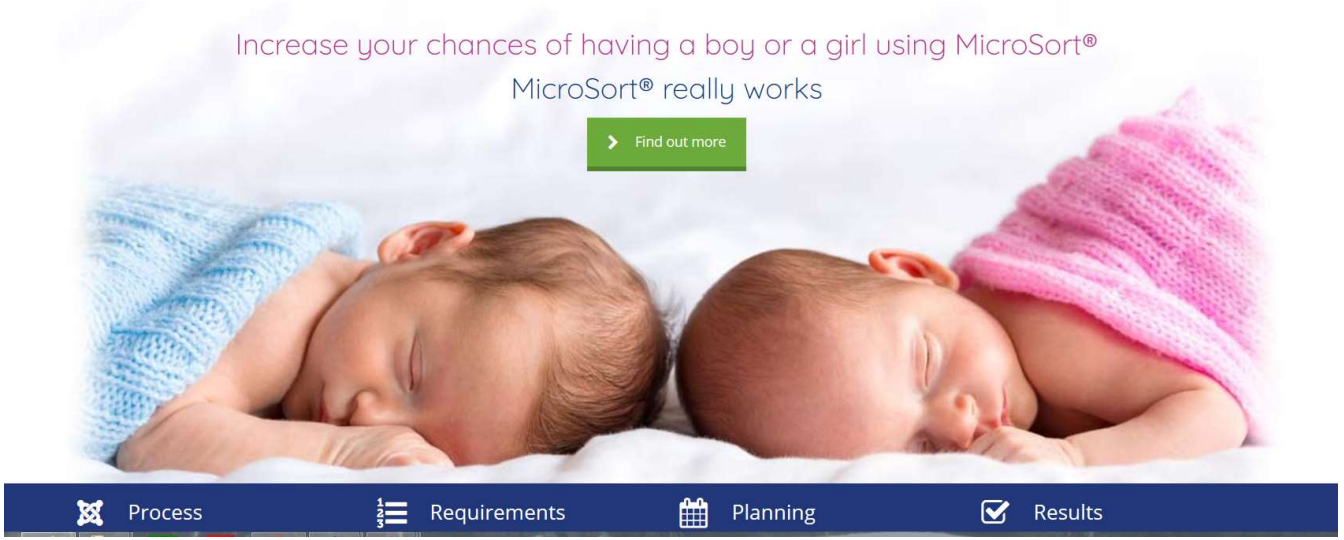
STATA

```
. power oneproportion 0.5 (0.6 0.7 0.8 0.9), test(wald) n(14)
onesided parallel

Estimated power for a one-sample proportion test
Wald z test
H0: p = p0  versus  Ha: p > p0
```

alpha	power	N	delta	p0	pa
.05	.1891	14	.1	.5	.6
.05	.4953	14	.2	.5	.7
.05	.8773	14	.3	.5	.8
.05	.9996	14	.4	.5	.9

Exercise: Does the MicroSort Method of Gender Selection Increase the Likelihood That a Baby Will Be a Girl?



In clinical trials, among 945 babies born to parents who used the XSORT method in trying to have a baby girl, 879 couples did have baby girls, for a success rate of 93%. Under normal circumstances with no special treatment, girls occur in about 50% of births. (Actually, the current birth rate of girls is 48.8%, but we will use 50% to keep things simple.) **Can we actually support the claim that the XSORT technique is effective in increasing the probability of a girl?**

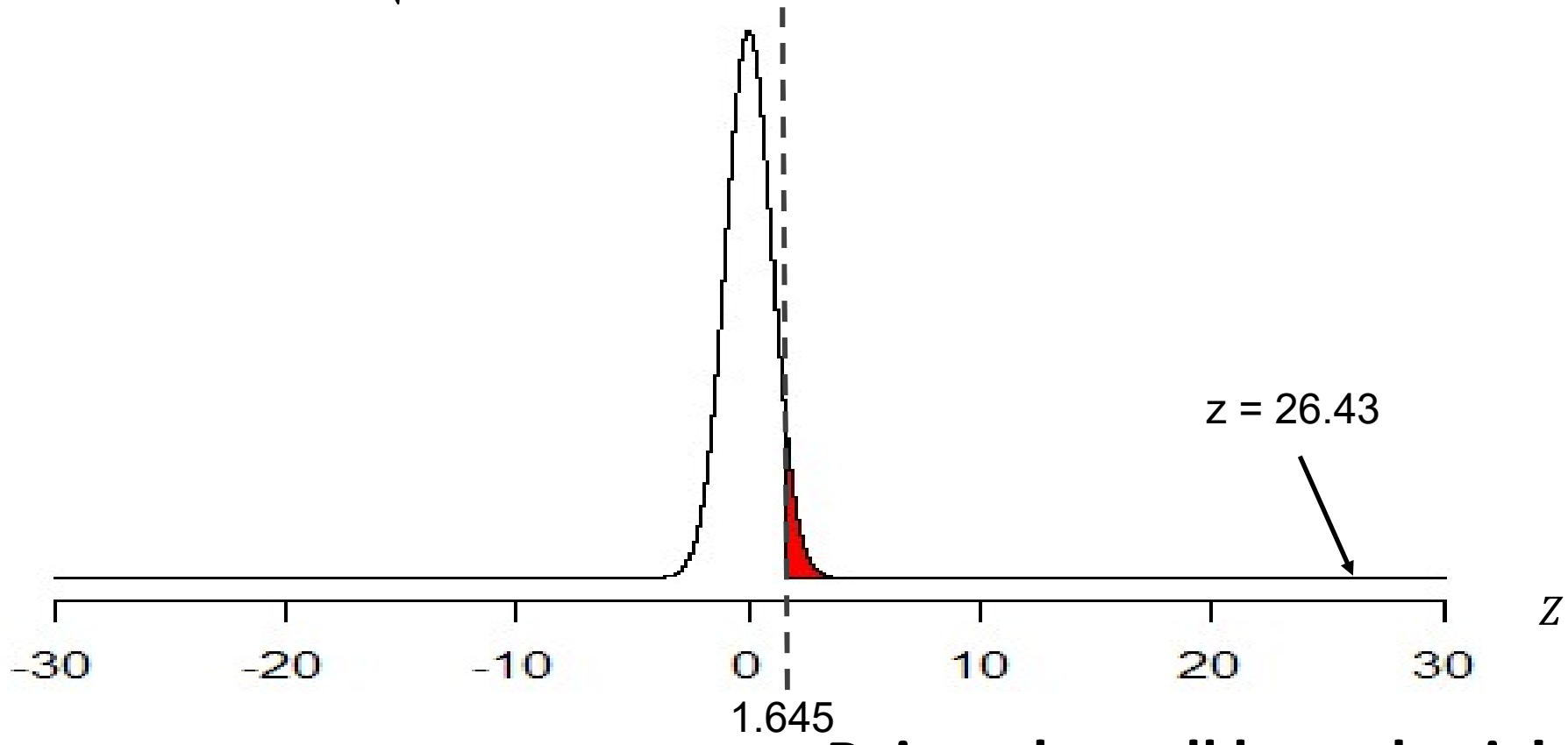
Exercise

N=945

p=879/945=0.93

$\alpha = 0.05 \;\; Z_{0.95} = 1.645$

$$\mathbf{Z = \frac{p - \pi_0}{se(p)}} \qquad \mathbf{z = \frac{0.93 - 0.5}{\sqrt{0.5 * 0.5 / 945}} = 26.43}$$



Reject the null hypothesis!