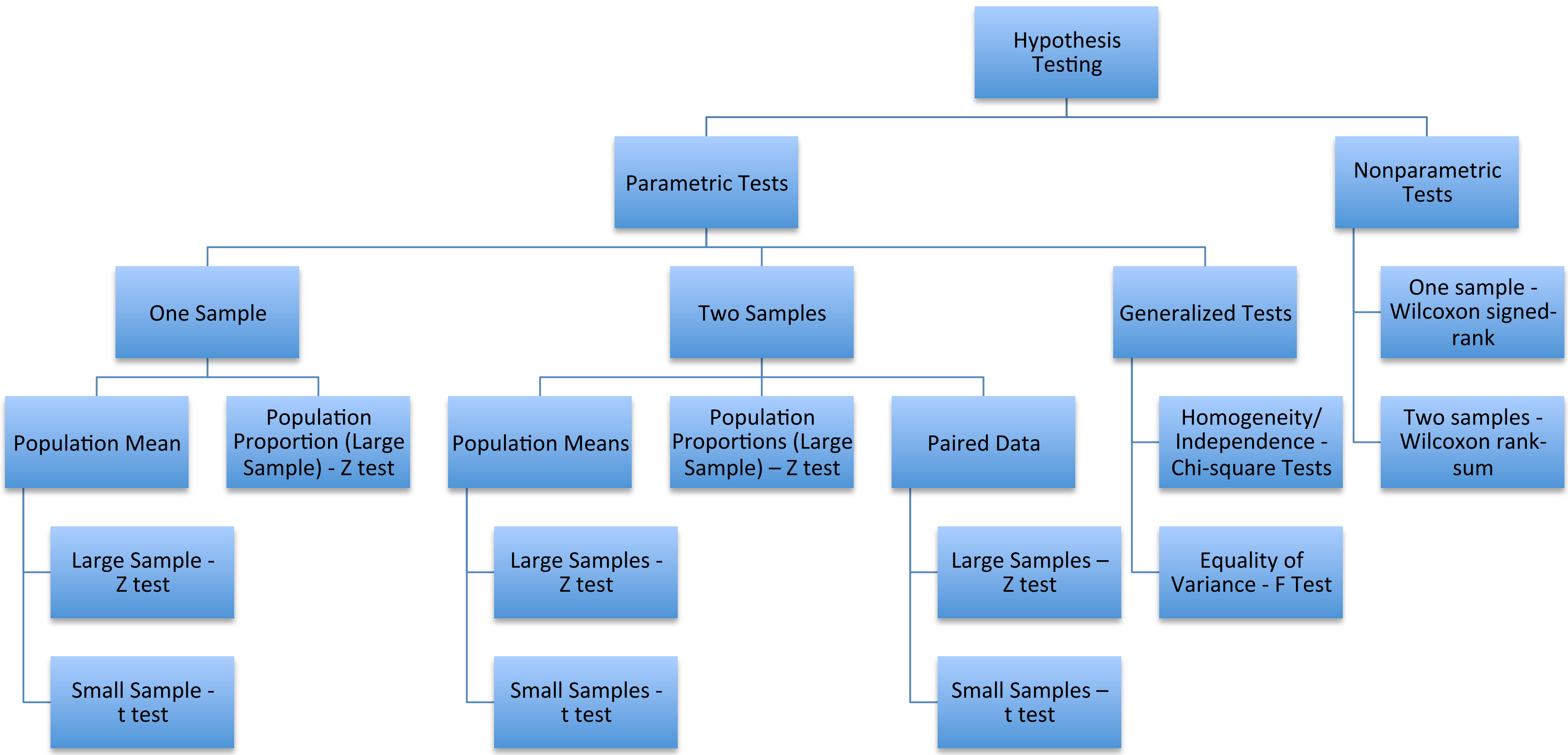
Properties of Hypothesis Tests:
Critical Points, Type I/II Error, Power, and Multiple Testing

Keegan Korthauer
Department of Statistics
UW Madison

HT Toolbox


Hypothesis Testing
Parametric Tests
Nonparametric Tests
One Sample
Two Samples
Generalized Tests
One sample - Wilcoxon signed-rank
Population Mean
Population Proportion (Large Sample) - Z test
Population Means
Population Proportions (Large Sample) – Z test
Paired Data
Homogeneity/ Independence - Chi-square Tests
Two samples - Wilcoxon rank-sum
Large Sample - Z test
Large Samples - Z test
Large Samples – Z test
Equality of Variance - F Test
Small Sample - t test
Small Samples - t test
Small Samples – t test

Fixed-Level Testing

- The Smaller the p-value, the more evidence against the null we have
- There is no “correct” P-value cutoff: recall our discussion of **practical vs statistical** significance
- **Fixed-level test**: a test with a fixed cutoff point α (called the ‘significance level’) for the P-value

Critical Point and Rejection Region

- In a fixed-level test, a **critical point** is a value of the test statistic that produces a P-value exactly equal to α
  - How to find: Set the p-value equal to α and solve for the test statistic

- If the test statistic is *more extreme* than the critical value, the P-value will be less than α, and H_0 will be rejected

- The region that captures the values of the test statistic that will lead to a rejection of the null hypothesis is called the **rejection region**

Example 6.27

- A scale is to be calibrated by weighing a 1000g test weight 60 times (sample standard deviation 2g)
- The null and alternative hypotheses are:

$$H_0: \mu = 1000 \text{ versus } H_1: \mu \neq 1000$$

- Find the critical points and rejection region for this test at the 5% level

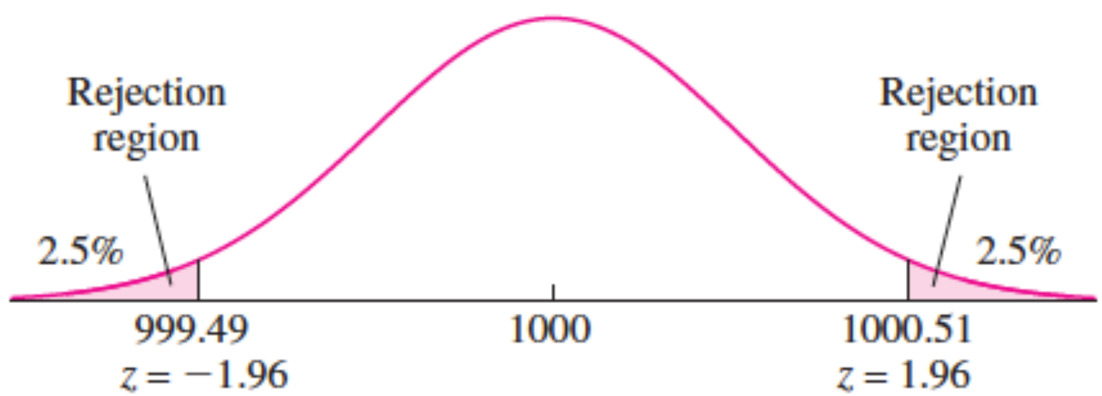


FIGURE 6.24 The rejection region for this two-tailed test consists of both the lower and the upper 2.5% of the null distribution. There are two critical points, 999.49 and 1000.51.

Example 6.26

- Let μ be the mean compressive strength (MPa) of a new concrete mix that has a population sd $\sigma = 70$ Mpa
- We'll sample 100 concrete blocks and test the hypotheses:

$$H_0: \mu \leq 1350 \text{ versus } H_1: \mu > 1350$$

- Find the critical point and rejection region for this test

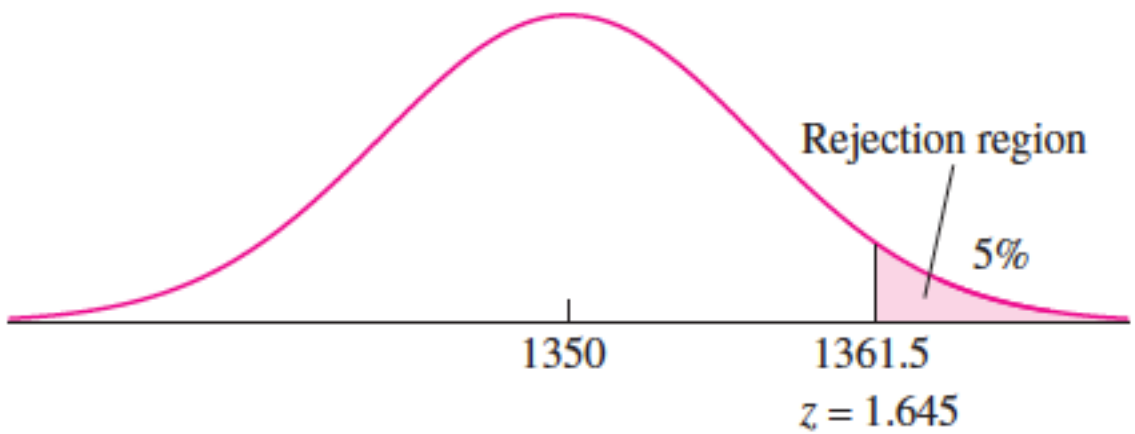


FIGURE 6.23 The rejection region for this one-tailed test consists of the upper 5% of the null distribution. The critical point is 1361.5, on the boundary of the rejection region.

One-sided vs Two-sided Tests

- It is easier to reject a one-sided test than a two-sided test
- Deciding between the two depends on the scientific question at hand, and is best done before the data is collected
  - Unless we are truly not interested in the other extreme direction, a two-sided test should be used
  - Do not use the data to choose (leads to bias and increased error rates)
- Example from Homework 8, Problem 8 – Difference in coding time between two languages

TYPE I AND TYPE II ERROR

Two Types of Errors in HTs

		The Truth	
		H_0 is True	H_0 is False
Result of the Hypothesis Test	Reject H_0		
	Fail to reject H_0		

Two Types of Errors in HTs

		The Truth	
		H_0 is True	**H_0 is False**
Result of the Hypothesis Test	**Reject H_0**		Correct
	Fail to reject H_0	Correct	

Two Types of Errors in HTs

		The Truth	
		H_0 is True	**H_0 is False**
Result of the Hypothesis Test	**Reject H_0**	Type I Error	Correct
	Fail to reject H_0	Correct	Type II Error

Two Types of Errors in HTs

Expressed mathematically,

$$P(\text{Type I Error}) = P(\text{Reject } H_0 \mid H_0 \text{ True})$$

$$P(\text{Type II Error}) = P(\text{Fail to Reject } H_0 \mid H_0 \text{ False})$$

Example

A test is made of the hypotheses H_0: $\mu \leq 10$ versus H_1: $\mu > 10$

For each of the following situations, state whether the correct decision was made, a type I error occurred, or a type II error occurred:

1. $\mu = 12$ and H_0 is rejected

   Correct decision!

2. $\mu = 11$ and H_0 is not rejected

   Type II error

3. $\mu = 9$ and H_0 is rejected

   Type I error

Type I Error and α

The probability of a type I error is never greater than α

Why is this true?

- A type I error occurs if H_0 is true but we reject it
- In order to reject, the test statistic must have been in most the extreme 100α% of the null distribution (rejection region)
- But if H_0 is true then we know that the null distribution is the **true** distribution so we should only land in that rejection region 100α% of the time
- Nice discussion in the book that explains this in more detail

Type I/II Error Tradeoff: Choice of α

- We would like to have very small error rates for both type I and type II errors

- Since P(type I error) $\leq \alpha$, can't we just make α really small?
  - Unfortunately, decreasing α generally increases the rate of type II errors

- The choice of α depends on balancing the "costs" of the two types of errors:
  - If Type I error "costs" more, then we choose a smaller α
  - This can depend on the context of the problem
  - Default: use the (reasonably small) value 0.05

Example

In a court of law, we assume innocence until guilt is proven:

H_0: the defendant is innocent
H_1: the defendant is guilty

Type I error: send the innocent person to jail
Type II error: let the guilty person go free

A type I error here is (generally) more costly, so we will use a very small α so the probability we send an innocent person to jail is very small

POWER OF HYPOTHESIS TESTS

The Power of a Hypothesis Test

- **Power:** the probability of rejecting H_0 when it is false

$$\text{Power} = 1 - P(\text{Type II error})$$
$$= 1 - P(\text{Fail to Reject } H_0 \mid H_0 \text{ False})$$

- A useful test should have large power (low probability of type II error)
  - Recall the tradeoff between type I and type II error

- Power is calculated to determine if the proposed test will be likely to reject H_0 in the event that it is false

The Power Calculation

- The power of a hypothesis test depends on many factors
  - usually calculated for a range of likely scenarios to get a realistic range of the power ***before the data is collected***

- For a HT of a population mean μ, power depends on:
  - Whether the test is one-sided or two-sided
  - The specific value of μ under H_1 (the alternative hypothesis): μ_1
  - The standard deviation σ
  - The sample size
  - The significance level α

- The calculation has five steps (see next slide)

5 Steps to Calculate Power

1. State the hypotheses

2. Work out the conditions to reject H_0 at significance level α:
   (a) Find the critical point(s) of the test statistic
   (b) Transform to the original scale

3. Choose the particular case of when H_0 is false (i.e. specify the value of the parameter under H_1)

4. Determine the distribution of the statistic in step 2b when H_0 is false

5. Calculate the probability of rejecting H_0 when H_0 is false (the power)

Example 6.28

Find the power of the 5% level test of $H_0: \mu \leq 80$ versus $H_1: \mu > 80$ for the mean yield of the new process under the alternative $\mu = 82$, assuming $n = 50$ and $\sigma = 5$.

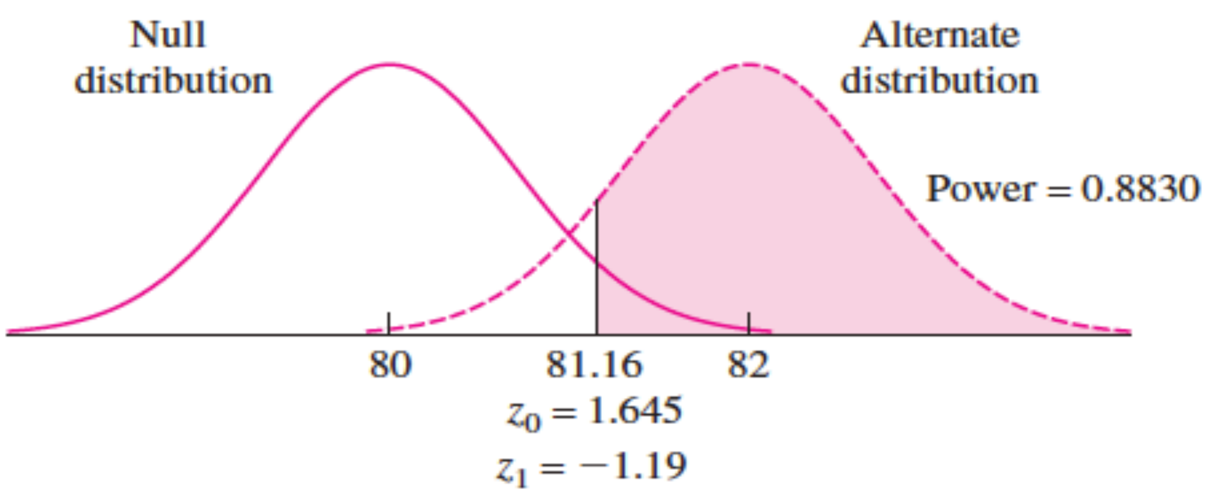
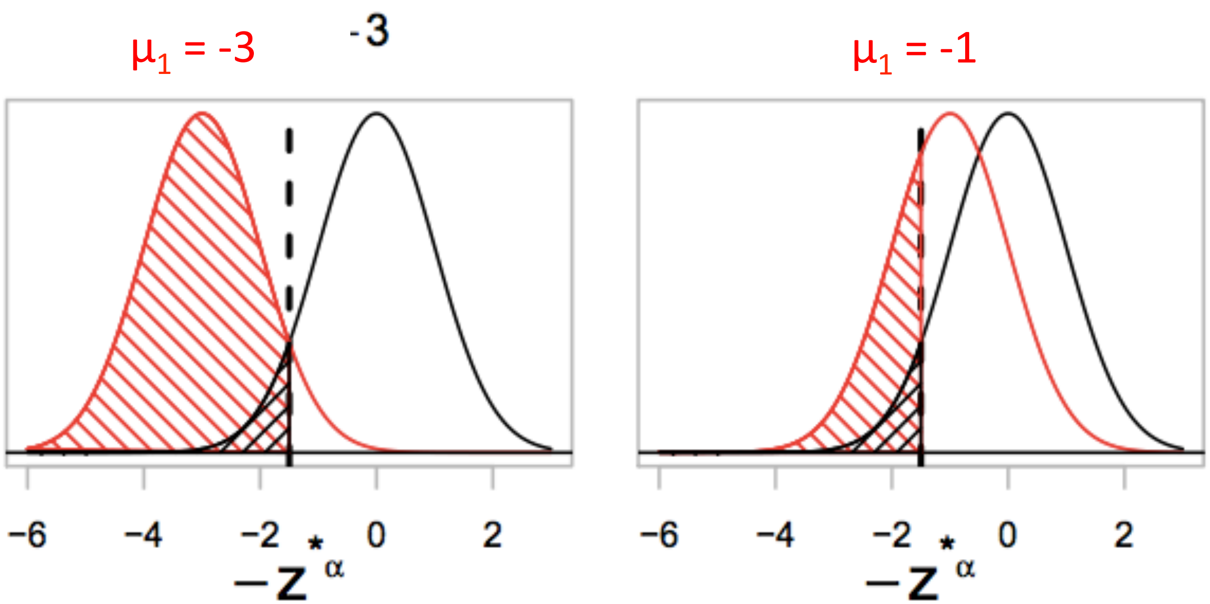


FIGURE 6.28 The rejection region, consisting of the upper 5% of the null distribution, is shaded. The z-score of the critical point is $z_0 = 1.645$ under the null distribution and $z_1 = -1.19$ under the alternate. The power is the area of the rejection region under the alternate distribution, which is 0.8830.

Value of the Parameter Under H_1

The further away μ_1 is from the hypothesized value under H_0, the greater the power

H_0: $\mu = 0$



What if Power is Low?

- If the power is too low, it may not be worthwhile to collect the data
  - No rule of thumb threshold, but generally power values in the range of 0.8 or higher are considered 'good'

- Can change the study design to increase power:
  - **Increase the sample size**
  - Increase the significance level (if it's acceptable to increase type I error rate)

Calculating Power for Other Tests

- The power calculation for a test of proportion follows similarly, except
  - we need to know p_1 instead of μ_1
  - we will use the distribution of $\hat{p}$ instead of $\bar{X}$

- To calculate the power tests using the t or F distributions, will need to use R to get a precise estimate
  - Will only get a broad range using the tables

Example 6.30

A pollster will conduct a survey of a random sample of voters in a community to estimate the proportion who support a measure on school bonds.

Let p be the proportion of the population who support the measure. The pollster will test H_0: $p = 0.50$ versus H_1: $p \neq 0.50$ at the 0.05 level. If 200 voters are sampled, what is the power of the test if the true value of p is 0.55?

MULTIPLE TESTING

Multiple Tests

- Sometimes we need to perform many hypothesis tests simultaneously

- The basic rule is that **as more tests are performed, the confidence that we can place in our results decreases**

Why?

- Say you have 20 hypothesis tests you wish to perform simultaneously
- You might think to simply test each one separately, using a level of significance of 0.05
- Consider the case where actually all of the null hypotheses are correct (we should not reject any)
- Then what is the probability of observing **at least one significant result just due to chance?** (<- type I error)

P(at least 1 type I error) = 1 – P(no type I errors)

$= 1 - (1\text{-}P(\text{type I error in one test}))^{20}$

$= 1 - (1 - 0.05)^{20} = 0.6415$

Over 64% chance of getting at least one false positive in 20 HTs at the 0.05 level!

Multiple Testing Problem

- This problem just gets worse as we increase the number of tests
  - In genomics, we often desire to perform a hypothesis test on each gene when we have measurements for thousands of genes
  - In that case we are pretty much guaranteed to have many type I errors (false positives)

- How do we control the probability of having type I error for a collection of hypothesis tests?

One Solution – Bonferroni Method

- The **Bonferroni method** provides a way to adjust p-values upward when several hypothesis tests are performed
  - makes it more difficult to reject the null hypothesis

- If a p-value remains small even after the adjustment, then the null hypothesis may be rejected

- To make the Bonferroni adjustment, simply multiply the p-value by the number of tests performed N:

  Bonferroni-adjusted p-value = N x Original p-value

Exercise 6.14.3

Six different settings are tried on a machine to see if any of them will reduce the proportion of defective parts. For each setting, an appropriate null hypothesis is tested to see if the proportion of defective parts has been reduced. The six P-values are 0.34, 0.27, 0.002, 0.45, 0.03, and 0.19.

a. Find the Bonferroni-adjusted P-value for the setting whose P-value is 0.002. Can you conclude that this setting reduces the proportion of defective parts? Explain.

b. Find the Bonferroni-adjusted P-value for the setting whose P-value is 0.03. Can you conclude that this setting reduces the proportion of defective parts? Explain.

Notes

- The Bonferroni adjustment is conservative
  - Protects the overall probability of a *single* type I error at the expense of reducing power

- There are other (more complicated) methods of adjusting p-values for multiple tests that are effective at reducing the type I error rate but not penalizing power as much
  - e.g. the widely used Benjamini-Hochberg method

Next

- Using R to perform hypothesis tests

- Intro to correlations