

# 1-sample Hypothesis Tests

## Module 9

Statistics 251: Statistical Methods

Updated 2021

### Introduction

We have learned about estimating parameters by point estimation and interval estimation (specifically confidence intervals). More often than not, the objective of an investigation is not to estimate a parameter but to decide which of two (or more) contradictory claims about the parameter is correct.

This part of statistics is called *hypothesis testing*

### Terms

*Statistical hypotheses is a claim or assertion about*

- (1) The value of a single parameter
- (2) The values of several parameters
- (3) The form of an entire probability distribution

*Hypotheses*

- (1) Null hypothesis, denoted by  $H_0$ , is the claim that is initially assumed to be true (the “prior belief” or “historical” claim)
- (2) Alternative hypothesis, denoted by  $H_a$ , is the assertion that is contradictory to  $H_0$ ; it is a researcher’s claim, what they are trying to prove (thus the reason behind the study)

### Hypothesis Testing Checklist

All tests include the following four steps:

- (1) State hypotheses, check assumptions
- (2) Calculate the test statistic
- (3) Find the rejection region
- (4) Results and conclusion of the test
- (5) State possible error that could have been made and discuss it within the context

### Hypotheses

When stating the hypotheses, the notation used is always population parameter notation; inferences upon populations need population notation (the Greek letters)

$\mu$  for the mean and  $\pi$  for the proportion

### Hypotheses for $\mu$

*Hypotheses for inferences concerning means (regardless of whether or not  $\sigma$  is known*

$$H_0 : \mu = \mu_0 \text{ vs. } H_a : \mu \neq \mu_0$$

$$H_0 : \mu \geq \mu_0 \text{ vs. } H_a : \mu < \mu_0$$

$$H_0 : \mu \leq \mu_0 \text{ vs. } H_a : \mu > \mu_0$$

Most often the null hypothesis will have = while the alternative will be one of either  $\neq$ ,  $>$ , or  $<$ .  $\mu_0$  is a specified value (a number that is given in the problem)

## Hypotheses for $\pi$

*Hypotheses for inferences concerning proportions:*

$$H_0 : \pi = \pi_0 \text{ vs. } H_a : \pi \neq \pi_0$$

$$H_0 : \pi \geq \pi_0 \text{ vs. } H_a : \pi < \pi_0$$

$$H_0 : \pi \leq \pi_0 \text{ vs. } H_a : \pi > \pi_0$$

Most often the null hypothesis will have = while the alternative will be one of either  $\neq$ ,  $>$ , or  $<$ .  $\pi_0$  is a specified value (a number that is given in the problem)

## Assumptions

- (1) Independence: observations are independent from one another
- (2) Randomization: proper randomization was used
  - Takes care of independence issue if there is one
- (3) Normality
  - (a) Means need an *approximate* normal distribution ( $n \geq 30$  should take care of it)
  - (b) Proportions need  $n \geq 60$  (via CLT)

**If assumptions are violated, the results from the analyses are not valid nor reliable**

## Test Statistic

1-sample test of the mean  $\mu$  when  $\sigma$  is known: Use  $Z$

$$z = \frac{\bar{X} - \mu_0}{se_{mean}} ; se_{mean} = \frac{\sigma}{\sqrt{n}}$$

1-sample test of the proportion  $p$ : Use  $Z$

$$z = \frac{\hat{\pi} - \pi_0}{se_{\pi}} ; se_{\pi} = \sqrt{\frac{\pi_0(1 - \pi_0)}{n}}$$

1-sample test of the mean  $\mu$  when  $\sigma$  is unknown: Use  $t$

$$t = \frac{\bar{X} - \mu_0}{se_{mean}} ; se_{mean} = \frac{s}{\sqrt{n}}$$

## Rejection Region

Is based on significance level  $\alpha$ .  $\alpha = 1 - CL$  where CL is the confidence level

**Always** assume  $\alpha = 0.05$  unless specified otherwise)

Two methods for rejection:

- (1) Critical value approach (not learning)
- (2) *pvalue* approach

**The alternative hypothesis ( $H_a$ ) determines rejection based on where you are at on the curve**

### *pvalue* logistics I

The *pvalue* of a test is the probability that, *given* the null hypothesis ( $H_0$ ) is true, the results from another random sample will be as or more extreme as the results we observed from our sample.

The *pvalue* of the test is dependent on the type of test you are doing, as in one-tail upper, one-tail lower, or two-tail. The sign of the alternative hypothesis is the determining factor in calculation of the *pvalue*.

### *pvalue* logistics II

The *pvalue* approach; the null hypothesis can be rejected *iff* (if and only if)  $pvalue \leq \alpha$  (with  $\alpha = 0.05$  most often). This does not change, regardless of the sign of the alternative hypothesis. However, the calculation of the *pvalue* is dependent on the sign of the alternative hypothesis. The *pvalue* will be the  $P$ ( the results of the test |  $H_0$  is correct), in other words, it is the probability that the results would occur by random chance if the null hypothesis is actually correct.

Assume that  $\alpha = 0.05$  unless specified; any rejection of  $H_0$  means that the results (of experiment, survey, etc.) are significant.

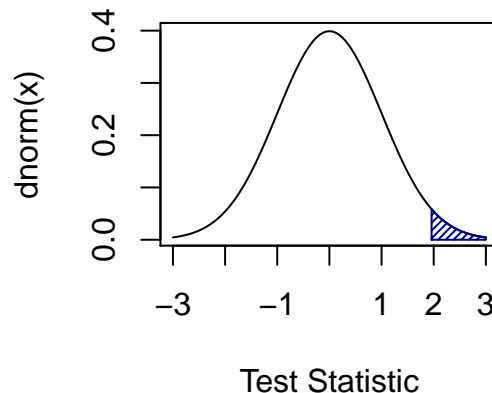
$$pvalue \leq \alpha \Rightarrow \text{Reject } H_0$$

$H_a : >$  **upper tail test**

**Note that while all examples are with  $z$ , it is interchangeable with  $t$  ( $df$  is needed).** In this case, *pvalue* represents the rejection region in the right tail of the distribution.

$$pvalue = P(Z \geq z_{calc}) = 1 - P(Z \leq z_{calc})$$

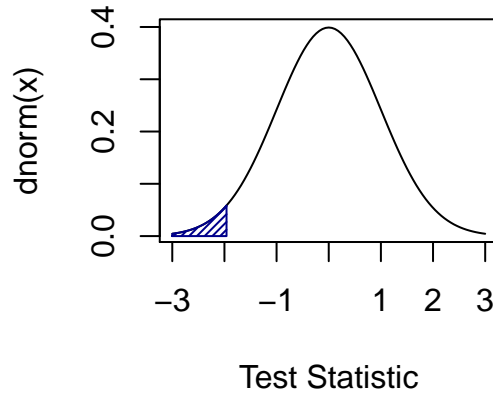
### **pvalue for upper tail test**



$H_a : <$  lower tail test

$$pvalue = P(Z \leq z_{calc})$$

### pvalue for lower tail test

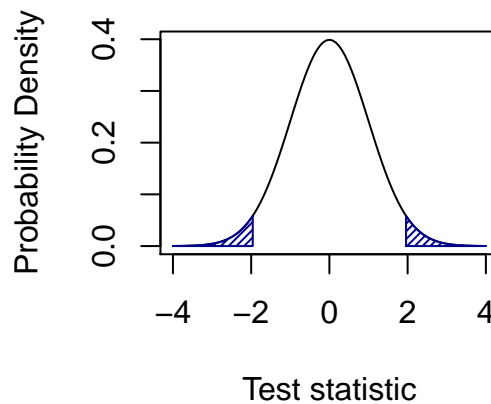


$H_a : \neq$  two tail test

$$pvalue = 2[P(Z \leq z_{calc})] \text{ or } 2[1 - P(Z \leq z_{calc})]$$

$$= 2[1 - P(Z \leq |z_{calc}|)]$$

### pvalue for 2-tailed test



## Results and Conclusion

- Results: we either
  - Reject  $H_0$  (rejecting the null hypothesis in favor of the alternative)
  - Fail to reject  $H_0$  (we are not rejecting the null hypothesis so that means that the null hypothesis gives a reasonable explanation of the question at hand) Conclusion: explain what the results did in relation to the actual data

### pvalue rejection Examples

- (1)  $pvalue = 0.4$  with  $\alpha = 0.05$ . Since  $pvalue = 0.4 \not\leq \alpha(0.05)$ ,  $H_0$  is not rejected (fail to reject  $H_0$ ). There is a 40% chance that we would see these results due to random chance (dumb luck) if the null

hypothesis is correct; results are not significant.

- (2)  $pvalue = 0.04$  with  $\alpha = 0.05$ . Since  $pvalue = 0.04 \leq \alpha(0.05)$ ,  $H_0$  is rejected. There is a 4% chance that we would see these results due to random chance (dumb luck) if the null hypothesis is correct; results are significant.
- (3)  $pvalue = 0.04$  with  $\alpha = 0.01$ . Since  $pvalue = 0.04 \not\leq \alpha(0.01)$ ,  $H_0$  is not rejected. There is a 4% chance that we would see these results due to random chance (dumb luck) if the null hypothesis is correct; results are not significant.

## Errors

Type I= $\alpha = P(\text{reject } H_0 | H_0 \text{ true})$ . This is a conditional probability statement that reads as “the probability of rejecting the null given that the null is true.”

TLDR; we rejected a true null hypothesis (that’s a bad thing).

**Type I can only happen when  $H_0$  is rejected**

Type II= $\beta = P(\text{Fail to reject } H_0 | H_0 \text{ false})$ . This is a conditional probability statement that reads as “the probability of not rejecting the null given that the null is false.”


TLDR; we kept a false hypothesis (again, a bad thing).

**Type II can only happen when  $H_0$  is not rejected**

Power= $1 - \beta = P(\text{reject } H_0 | H_0 \text{ false})$ . This is a conditional probability statement that reads as “the probability that the null is rejected given that it is false.”

TLDR; we correctly rejected  $H_0$  when  $H_0$  is false (a good thing. Finally!)

## Error table



		The truth	
		$H_0$ true	$H_0$ false
My decision	Reject $H_0$	Type I ( $\alpha$ )	:-)
	Fail to reject $H_0$	:-)	Type II ( $\beta$ )

Figure 1: Errors

## Checklist

- (1) State hypotheses, check assumptions if requested
- (2) State  $t$  statistic,  $df$ , and  $pvalue$  from output
- (3) State test results
- (4) Make conclusion in context from results
- (5) State possible error that could have been made and discuss it within the context

Again, we will be using `t.test()` in R for the tests

## Example test of $\mu$

A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is 130°F; it is known from previous studies that the temperatures are normally distributed. A random sample of  $n = 9$  systems was taken. Is there sufficient evidence that the true mean activation temperature is more than what the manufacturer claims?

## Sprinklers setup

$$H_0 : \mu = 130 \text{ vs. } H_a : \mu > 130$$

Assumptions:

- (1) Independence: random so yes
- (2) Randomization: yes
- (3) Normality: stated temps were normal so yes

Organization of information:

$\mu_0 = 130$  (claimed mean)

$n = 9$  (acceptable because temps are normal or  $n \geq 30$ ; either way we will use  $t$ )

$H_a : >$  (upper tail test)

$\alpha = 0.05$  (assumed because not specifically stated otherwise)

## Sprinklers analysis output

```
t.test(sprinklers,mu=130,alternative='g')
```

One Sample t-test

data: sprinklers

$t = 3.3293$ ,  $df = 8$ ,  $p\text{-value} = 0.005197$

alternative hypothesis: true mean is greater than 130

95 percent confidence interval:

130.5965          Inf

sample estimates:

mean of x

131.3512

## Sprinklers conclusion

$t = 3.3293$ ,  $df = 8$ ,  $pvalue = 0.005197$

Results:  $pvalue = 0.005197 \leq \alpha(0.05) \therefore$  (therefore)  $H_0$  is rejected

Conclusion: since the null is rejected, that means that there is evidence that the sprinkler activation temperature is higher than the manufacturer's claim of  $130^\circ\text{F}$

Error: since  $H_0$  was rejected, a Type I error (reject null when null is true) could have been made; we think the activation temperature is higher than the claim but it is not higher. Why do we care?

## Example test of $\mu$

New York City, NY is known as the "city that never sleeps." A random sample of 25 New Yorkers was taken and they were asked how much sleep they get per night. Hours of sleep follow an approximate normal distribution. Is there sufficient evidence that New Yorkers get a different amount of sleep from the "norm"; a full 8 hours of sleep?

## New Yorkers setup

$$H_0 : \mu = 8 \text{ vs. } H_a : \mu \neq 8$$

Assumptions:

- (1) Independence: random so yes

- (2) Randomization: yes
- (3) Normality: stated hours of sleep were normal so yes

Organization of information:

$\mu_0 = 8$  (claimed mean)

$n = 25 < 30$  (hours of sleep is normal)

$df = n - 1 = 25 - 1 = 24$

$H_a : \neq$  (2 tail test)

$\alpha = 0.05$  (assumed because not specifically stated otherwise)

## New Yorkers analysis output

```
t.test(ny, mu=8)
```

One Sample t-test

data: ny

t = 0.45284, df = 24, p-value = 0.6547

alternative hypothesis: true mean is not equal to 8

95 percent confidence interval:

7.527738 8.737749

sample estimates:

mean of x

8.132743

## New Yorkers conclusion

$t = 0.45284$ ,  $df = 24$ ,  $pvalue = 0.6547$

Results:  $pvalue = 0.6547 \not\leq \alpha(0.05) \therefore H_0$  is not rejected

Conclusion: since the null is not rejected, that means that there is not enough evidence to say that New Yorkers sleep on average is different than the usual 8.

Error: since  $H_0$  was not rejected, a Type II error (not rejecting the null when the null is false) could have been made; we think New Yorkers get around 8 hours of sleep when they do not. Why do we care?

## Example test of $\pi$

Ingots are huge pieces of metal often weighing more than 10 tons (20,000 lbs.). They must be cast in one large piece for use in fabricating large structural parts for cars and planes. If they crack while being made, the crack can propagate into the zone required for the part, compromising its integrity; metal manufacturers would like to avoid cracking if at all possible. In one plant, only about 80% of the ingots have been defective-free. In an attempt to reduce the cracking, the plant engineers and chemists have tried some new methods for casting the ingots and from a random sample of 500 ingot cast in the new method, 16% of the casts were found to be defective (cracked). Is there sufficient evidence that the defective rate has decreased?

## Ingots setup

$$H_0 : \pi = 0.2 \text{ vs. } H_a : \pi < 0.2$$

Assumptions:

- (1) Independence: random so yes
- (2) Randomization: yes
- (3) Normality:  $n = 500 \geq 60$  so yes

Organization of information:  
 $\hat{\pi} = 0.16$  (sample proportion)  
 $\pi_0 = 0.2$  (old method proportion)  
 $n = 500 \geq 60$   
 $H_a : <$  (lower tail test)  
 $\alpha = 0.05$  (assumed because not specifically stated otherwise)

## Ingots analysis output

```
t.test(ingots,mu=.2,alternative='l')
```

One Sample t-test

```
data: ingots
t = -2.4373, df = 499, p-value = 0.007573
alternative hypothesis: true mean is less than 0.2
95 percent confidence interval:
  -Inf 0.1870448
sample estimates:
mean of x
  0.16
```

## Ingots conclusion

$t = -2.4373$ ,  $df = 499$ ,  $pvalue = 0.007573$

Results:  $pvalue = 0.007573 \leq \alpha(0.05) \therefore H_0$  is rejected

Conclusion: since the null is rejected, that means that there is evidence that the defect rate of the new method is significantly less than the current method.

Error: since  $H_0$  was rejected, a Type I error (reject null when null is true) could have been made; we think the defect rate of the new method decreased but it did not. Why do we care?