Large-sample Tests of Hypotheses
Chapter 9

Purpose of Hypothesis Testing

In the last chapter, we studied methods of *estimating a parameter* (μ, p or p_1-p_2) based on sample data:

- point estimation
- confidence intervals

In contrast, the goal of hypothesis testing is to make a **decision** about the *value of a population parameter* based on *sample data*

Hypothesis Testing Applications

- A sociologist wants to determine if the mean number of children in an American family is still 2.5
- A buyer wants to decide if the proportion of defective bolts in a shipment exceeds 3% (the manufacturer's specification).
- The California Dept of Conservation needs to decide if the mean weight of a recycled aluminum can has decreased from 0.034 lb
  - Why? You pay in CRV based on the number of aluminum cans you buy. When you recycle your cans the CRV is paid out based on weight. Thus, count = total weight/averge weight of one can
  - To save money, aluminum can manufacturers are constantly trying to make the cans lighter

3

The Logic of Hypothesis Testing

How do we use the sample data to make these types of decisions?

- First, the research question is phrased as a decision about the value of a population parameter. Formally, this is done by choosing a **null hypothesis**, denoted H_0, and an **alternative hypothesis**, H_a.
- Next, we determine whether the sample data provide "evidence" ***against*** the null hypothesis.

Example: Is the mean number of children per family still 2.5 or has it changed?

H_0: $\mu = 2.5$

H_a: $\mu \neq 2.5$

Suppose we collect data for n=100 families and $\bar{x} = 2.3$ with s = 0.50. We need to decide if this is evidence against H_0: $\mu = 2.5$. Recall that some sample-to-sample variation in xbar is typical and expected.

4

Formulating the Null and Alternative Hypotheses

Example: Is the proportion of defective bolts in this shipment more than 3% (the manufacturer's specification)?

H_0: $p = 0.03$

H_a: $p > 0.03$

What evidence will we collect from a sample of size n?

Example: The California Dept of Conservation needs to decide if the mean weight of a recycled aluminum can has decreased from 0.034 lb

H_0: $\mu = 0.034$

H_a: __________

What evidence will we collect from a sample of n alum cans?

Would you consider a sample mean of 0.035 support for the alternative hypothesis (or evidence against the null)?

5

Formulating the Null and Alternative Hypotheses

- The null and alternative hypothesis must be stated in terms of a population parameter (**<u>never</u>** a sample statistic)
- The null will always contain the equality sign (=). The alternative will contain one of: ≠, > or <.
- No overlap between the parameter values specified under H_0 and H_a
- The parameter value(s) specified under H_0 typically represents the "status quo" or currently accepted belief. H_a represents the "new" finding the researcher wishes to establish
  - Current "wisdom" is taken as H_0; departures from it are taken as H_a
    - Normal human body temp is 98.6 deg F so H_0: $\mu=98.6$
    - Mean IQ is 100 so H_0: _______________
  - Generally, new drugs, teaching methods, and procedures must be **<u>proven</u>** to work so the hypothesis corresponding to the new procedure is "better" will be H_a
    - Drug A increases IQ gives H_0: __________ and H_a: _________

6

Trial by Jury Analogy

The logic behind hypothesis testing is similar to a trial by jury where the defendant is "assumed innocent until proven guilty."

H_0: the defendant is innocent
H_a: the defendant is guilty

- The prosecutor must convince the jury that the defendant is guilty "beyond a reasonable doubt" in order to obtain a conviction.
- A preponderance of evidence is required to obtain a conviction because we don't want declare an innocent person guilty.
- In the language of hypothesis testing, we don't want to make the mistake of rejecting a true null hypothesis. So we must have a lot of evidence (from the sample data) against H_0 in order to reject it.

7

Trial by Jury Analogy

Example: Is the mean number of children per family still 2.5 or has it changed?

H_0: $\mu = 2.5$

H_a: $\mu \neq 2.5$

- The evidence is our sample mean. This evidence should be compelling against H_0 before we are willing to reject H_0.
- Is $\bar{x} = 2.3$ compelling evidence against H_0? We consider whether $\bar{x} = 2.3$ is likely or unlikely under H_0. We will define a **rejection region** which defines values of $\bar{x}$ that are unlikely under H_0

8

Rejection Region for Example

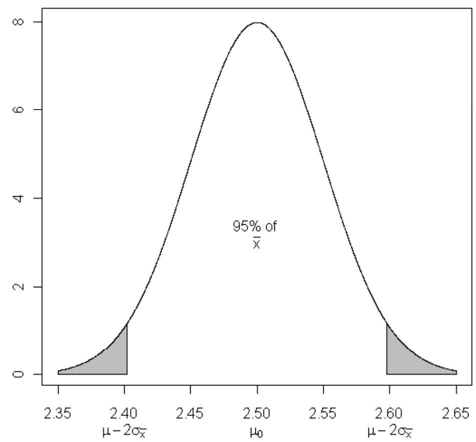
To determine the rejection region, we consider the distribution of xbar for n=100, ASSUMING H_0 IS TRUE

Since n is large, xbar :

- Is approximately normally distributed
- has mean = μ_0 (generic notation for the mean specified under H_0. Here, μ_0 is 2.5.
- has standard deviation ≈ s/sqrt(n) =0.5/sqrt(100) = 0.05

If H_0 is true, about 95% of the xbars will fall within 2 standard deviations of μ_0 or in the interval 2.5 ±2(0.05)= (2.40, 2.60). The unlikely values of xbar are outside this interval, beneath the gray shaded sections of the graph. Thus, the rejection region is (-∞,2.40) and (2.60, ∞).

Since xbar = 2.3 is below 2.40, we reject H_0.
Would we reject H_0 if xbar is 2.53?
2.64?



9

Alternate Rejection Region for Example

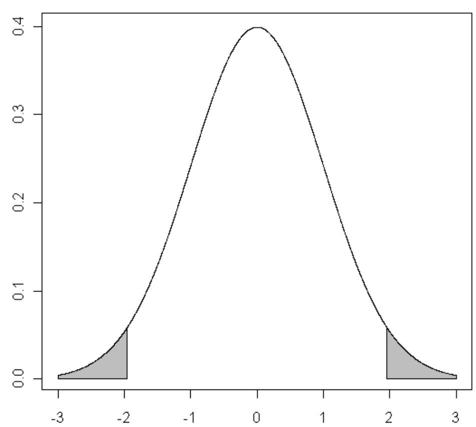
Alternately, we can standardize xbar, again assuming H_0 is true. Then compare it to the standard normal distribution to determine if it is unlikely under H_0.

Under the Empirical Rule, any z-score below -2 and above 2 is "unusual."

$$\frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{2.3-2.5}{0.5/\sqrt{100}} = -4.00$$

So the standardized xbar is unusual under H_0. Note that it falls in the gray rejection region of the left tail. We reject H_0.

Would we reject H_0 if the standardized test statistic was -0.97? 5.55?



10

Five Steps in a Hypothesis Test

All hypothesis tests share a common format:

- State the null hypothesis (statement about the value of a relevant parameter)
- State the alternative hypothesis
- Calculate the test statistic (the evidence from the sample data)
- Determine the Rejection Region
- State the conclusion in non-technical language

11

Summarize the Steps for the Example

1. H_0: $\mu = 2.5$
2. H_a: $\mu \neq 2.5$
3. Test statistic:

$$\frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{2.3-2.5}{0.5/\sqrt{100}} = -4.00$$

4. Reject H_0 if the standardized test statistics is below z = -2.00 or above z = 2.00. Since -4.00 < -2.00, we reject H_0
5. There is sufficient evidence to conclude that the population mean number of children in a family is no longer 2.5.

12

All 5 Steps for Another Example

Are aluminum cans **decreasing** in weight? Currently, they are assumed to have mean weight 0.034 lb.

A random sample of 64 cans has sample mean weight 0.033 lb with standard deviation 0.008 lb.

1. H_0: $\mu = 0.034$ lb
2. H_a: $\mu < 0.034$ lb
3. Sketch the distribution of xbar under H_0. Use it to calculate the standardized test statistic (z-score).
4. Reject H_0 if the standardized test statistics is below z = -2.00 or above z = 2.00. What is our decision?___________________
5. Conclusion:____
   ____

13

Types of Errors

Since we are basing our conclusion on sample data, we can make an error in our conclusion.

	H_0 is true	H_a is true
Reject H_0	Type I Error	Correct
"Accept" H_0	Correct	Type II Error

- α = Pr(Type I Error)=Pr(Reject H_0 when H_0 is true)
- β = Pr(Type II Error)=Pr(Accept H_0, when H_0 is false)
- However, we can control the probabilities of these errors.
- The only way to have $\alpha = 0$ and $\beta = 0$ is to sample the entire population.

14

Level of Significance

α is also called the **level of significance** (l.o.s) of the hypothesis test

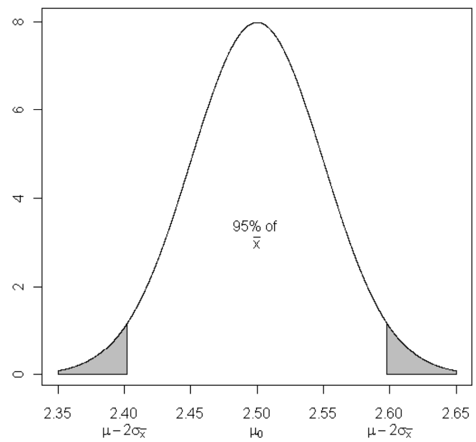
We can calculate α for the mean number of children in a family example

The distribution of xbar for samples of size 100 is shown again. Recall this distribution assumes the population mean is 2.5 (as specified under H_0). The values under the gray sections are the rejection region.

If the pop mean is really 2.5, is it possible to get an xbar in the rejection region? _____. In fact, it will happen about _____ of the time.

In other words, if H_0 is true, we will incorrectly reject H_0 5% of the time (over many repetitions of the hyp test). Since the l.o.s. is the probability of rejecting a true H_0, the l.o.s. is 0.05 .

How can we decrease α, the l.o.s.? Increase α?



15

Choosing the Rejection Region

It is standard to choose α first, ***then*** choose a rejection region so that the probability of rejecting a true H_0 is α.

Let's revisit the mean number of children per family example and choose a rejection region to obtain α=0.10:

1. H_0: μ = 2.5
2. H_a: μ ≠ 2.5
3. Test statistic: $z^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2.3 - 2.5}{0.5/\sqrt{100}} = 4.00$
4. We wish to reject H_0 for the 0.10=10% most unusual values of z^* (under H_0) that support H_a. Here, support for H_a will occur if the z* is ***too small or too large***. (two-tailed test).

We reject H_0 if z* is among the α/2=0.10/2=0.05=5% highest or the 5% lowest z*'s expected under H_0. (as shown in the graph on the next slide)

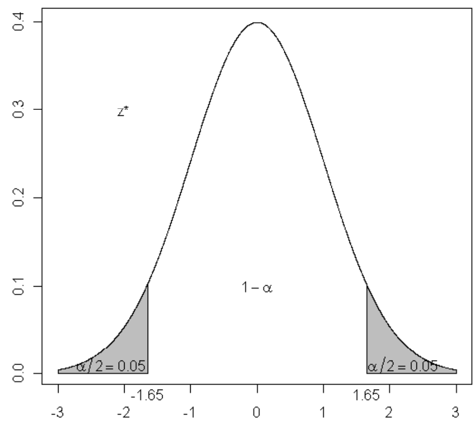
16

Rejection Region

If the null hypothesis true, z*, the standardized test stat will be distributed as shown in the graph.

It will be "too high" or "too low" as judged by the shaded rejection regions 0.10 or 10% of the time (assuming the null is true).

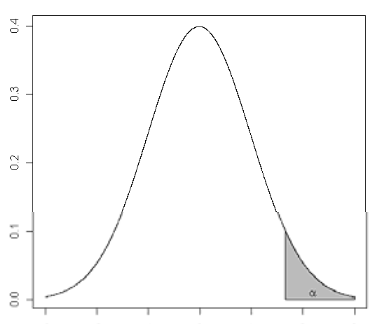
Thus, this rejection region will cause us to reject a true null hypothesis 10% of the time, and we attain the desired 0.10 probability of Type I error.



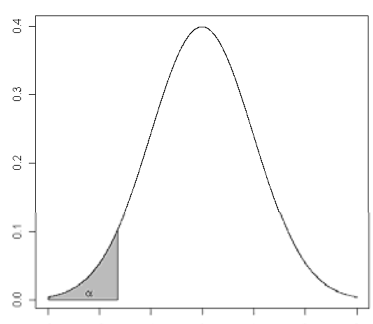
17

Rejection Regions for one-tailed Hypotheses

H_0: $\mu=\mu_0$ vs. Ha: $\mu>\mu_0$
We use a rejection region having right-tail area α



H_0: $\mu=\mu_0$ vs. Ha: $\mu<\mu_0$
We use a rejection region having left-tail area α



18

Calculate the Rejection Region

Assuming a standardized test statistic (z), determine the rejection region

- H_0: $\mu = 100$ vs. Ha: $\mu > 100$, $\alpha = 0.01$

What do you decide if $z^*=2.58$? $z^*=-2.58$?

- H_0: $\mu = 100$ vs. Ha: $\mu \neq 100$, $\alpha = 0.01$

- H_0: $\mu = 100$ vs. Ha: $\mu > 100$, $\alpha = 0.01$

19

Example: Hypothesis test at l.o.s. α (one-tailed)

Test if Drug A increases IQ at 0.05 l.o.s. A sample of 64 subjects on Drug A have sample mean IQ 105 and standard deviation 16.

1. H_0: $\mu = 100$
2. H_a: $\mu > 100$
3. Calculate standardized test statistic:______
4. Determine rejection region (depends on Ha and α). We will reject H0 if z* is __________________.
5. Conclusion: _______________________________

20

Example: Hypothesis test at l.o.s. α (two-tailed)

At Degrees-R-Us University, the true mean cost of a student's textbooks last semester was $503. A random sample of 60 students this semester have average textbook cost of $558 with standard deviation $70. At the 0.05 l.o.s., conduct a hypothesis test to determine if the true mean textbook cost has changed from last semester.

1. H_0: $\mu = 503$
2. H_a: $\mu \neq 503$
3. Calculate standardized test statistic:______
4. Determine rejection region (depends on Ha and α). We will reject H0 if z* is _________________.
5. Conclusion: ______________________________

21

What about Type II Error?

- Trade-off between α and β. For a fixed sample size, the smaller you make α, the larger β becomes and vice versa.
- Usually, sample size fixed due to amount of funding and α is fixed at 0.05 (or some other common value). Then, β is taken to be whatever these other two constraints dictate.
- Problem with the above approach: ____________________________________
- Sometimes however, it's the best we can do.

22

Exercise in Interpreting α and β

Non-statistical hypothesis test: You are on the road and notice your gas gauge nearing empty as you pass a gas station, the next gas station is 25 miles ahead.

H_0: I don't have enough gas to make it to the next gas station

H_a: I do have enough gas to make the next station

What are the Type I and II errors and the consequences of each? Do you prefer $\alpha = 0.01$ and $\beta=0.05$ or $\alpha=0.05$ and $\beta=0.01$?

23

Type I and Type II error for a Stat hypothesis test

Recall Example: Test if Drug A increases IQ at 0.05 l.o.s. A sample of 64 subjects on Drug A have sample mean IQ 105 and standard deviation 16.

1. H_0: $\mu = 100$ (drug doesn't work)
2. H_a: $\mu > 100$ (drug does work)

- Type I Error: reject H_0 when H_0 is true. Specifically, ______________________________
- Type II Error: accept H_0 when H_0 is false. Specifically, ______________________________

What is the probability Type I error?________

Type II error?_________

24

P-values

- Problem with reporting study results using the level of significance approach: ________________________________
- Consequently, pretty much every study reported in a professional journal will report the p-value
- The p-value allows the reader to decide if the study results present enough of a contradiction to H_0, (thus, "support" of H_a), based on his/her personal favorite l.o.s.
- The p-value is the probability of obtaining a test statistic as extreme or more extreme than what was observed -- assuming H_0 is true.
  - "Extreme" means values in the direction (too high or too low or both) that would support H_a.

25

Calculate p-values for one-tail test

Recall Example: Test if Drug A increases IQ at 0.05 l.o.s. A sample of 64 subjects on Drug A have sample mean IQ 105 and standard deviation 16.

1. H_0: $\mu = 100$ (drug doesn't work)
2. H_a: $\mu > 100$ (drug does work)
3. Test statistic: $$z^* = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{105 - 100}{16/\sqrt{64}} = 2.50$$
4. P-value = Pr(z*>2.50)=1-0.9938=0.0062 (Draw a picture here)
5. Thus, if the drug doesn't work and the true mean IQ is still 100 with the drug, the probability of seeing a sample mean as large or larger than we observed is 0.0062 (not likely!). We reject the null. There is sufficient evidence that the true mean IQ increases with the drug.

Small p-values call for rejection of H_0. If p-value ____ α, we reject H_0.

26

P-value for a two-tailed test

1. H_0: $\mu = 100$
2. H_a: $\mu \neq 100$
3. Test statistic: $z^* = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{105 - 100}{16/\sqrt{64}} = 2.50$
4. P-value = Pr(z*>2.50)+Pr(z*<-2.50)=2*(1-0.9938)=2(0.0062)=0.0124 ________
5. Conclusion: There is sufficient evidence to conclude the _______ mean differs from 100.

27

Is there scientific evidence for ESP?

- Ganzfeld study
  - One of four different pictures is randomly chosen as the target
  - "sender" sees the picture and tries to send the image to the receiver (in another room)
  - The receiver then sees the 4 pictures and attempts to identify the target picture
  - If the receiver does ***not*** have ESP, what is the probability he/she will guess the target picture?____
- Data from many subjects over many ganzfeld experiments yielded 122 "hits" out of 355 trials (Utts, 1991). The sample proportion is ________ = 34%, instead of 25% which would be expected in the absence of ESP. Are these data sufficient to conclude that ESP exists?

28

Does ESP Exist?

1. H_0: p = _____ (no ESP)
2. H_a: p > _____ (ESP)
3. Test statistic:

$$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.34 - 0.25}{\sqrt{\frac{0.25(1-0.25)}{355}}} = 3.92$$

4. P-value = Pr(z*>3.92)=0.000044.
   Decision______________
5. Conclusion:

29

What if the sample size is small?

All our inferential procedures for μ (confidence intervals and hypothesis tests) assume ***a large sample size***. If the sample size is small,

- the CI becomes:

$$\bar{x} \pm t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

- For Hypothesis test: the rejection region comes from the tails of the T-distribution with n-1 degrees of freedom (everything else is the same)
  - T-distribution is actually a family of distributions. Specifying the degrees of freedom as 1,2,3,… indicates which T-distn you refer to (picture here)
  - T-distn looks like the normal only with fatter tails
  - Table 4 gives probabilities for T-distn

30