Chapter 23

Hypothesis Testing – Examples and Case Studies

Thought Question 1:

Recall study where difference in sample means for weight loss based on dieting only versus exercising only was 3.2 kg. Same study showed difference in ***average amount of fat weight lost was 1.8 kg*** and corresponding ***standard error was 0.83 kg***. Suppose means are actually equal, so mean difference in fat lost for populations is actually zero. What is the **standardized score** corresponding to observed difference of 1.8 kg? Would you expect to see a standardized score that large or larger very often?

Thought Question 2:

In journal article in Case Study 6.4, comparing IQs for children of smokers and nonsmokers, one of the statements made was, "*After control for confounding background variables, the average difference [in mean IQs] observed at 12 and 24 months was 2.59 points (95% CI: –3.03, 8.20; P = 0.37).*" (Olds et al., 1994, p. 223)

Reported value of 0.37 is the *p*-value. What are the **null and alternative hypotheses** being tested?

Thought Question 3:

In chi-squared tests for two categorical variables, introduced in Chapter 13, we were interested in whether a relationship observed in a sample reflected a real relationship in the population.

What are the **null and alternative hypotheses**?

Thought Question 4:

In Chapter 13, we found a statistically significant relationship between smoking (yes or no) and time to pregnancy (one cycle or more than one cycle).

Explain what the **type 1 and type 2 errors** would be for this situation, and the consequences of making each type of error.

23.1 How Hypothesis Tests Are Reported in the News

1. Determine the **null** hypothesis and the **alternative** hypothesis.
2. Collect and summarize the **data** into a **test statistic**.
3. Use the test statistic to determine the ***p*-value**.
4. The result is **statistically significant** if the *p*-value is less than or equal to the level of significance.

Often media only presents results of step 4.

Example 1: Cranberry Juice

CHICAGO (AP) A scientific study has proven what many women have long suspected: ***Cranberry juice helps protect against bladder infections****. Researchers found that* ***elderly women who drank 10 ounces of a juice drink containing cranberry juice each day had less than half as many urinary tract infections as those who consumed a look-alike drink without cranberry juice****. The study, which appeared today in the Journal of the American Medical Association, was funded by Ocean Spray Cranberries, Inc., but the company had no role in the study's design, analysis or interpretation, JAMA said. "This is the first demonstration that cranberry juice can reduce the presence of bacteria in the urine in humans," said lead researcher Dr. Jerry Avorn, a specialist in medication for the elderly at Harvard Medical School.*

(*Davis Enterprise*, Mar. 9, 1994, p. A9)

Example 1: Cranberry Juice

- Study to compare odds of getting an infection for population of elderly women following two regimes: 10 ounces of cranberry juice per day or 10 ounces of a placebo drink.
- **Null hypothesis** is odds ratio (juice/placebo) is 1.
- **Alternative hypothesis** is odds of infection are higher for the group drinking the placebo (juice/placebo < 1).
- The article indicates that the **odds ratio is under 50%**.
- Original article (Avorn et al., 1994) sets it at 42% and reports that the associated ***p*-value is 0.004**.
- Newspaper article captured most important aspect of research, but not clear that *p*-value was extremely low.

23.2 Testing Hypotheses About Proportions and Means

If the null and alternative hypotheses are expressed in terms of a **population proportion, mean, or difference between two means** and if the sample sizes are large …

… the **test statistic** is simply the corresponding **standardized score** computed assuming the null hypothesis is true; and the ***p*-value** is found from a table of percentiles for standardized scores.

Example 2: Weight Loss for Diet vs Exercise

Did dieters lose more fat than the exercisers?

Diet Only:

sample mean = 5.9 kg
sample standard deviation = 4.1 kg
sample size = n = 42
standard error = $\text{SEM}_1 = 4.1/\sqrt{42} = 0.633$

Exercise Only:

sample mean = 4.1 kg
sample standard deviation = 3.7 kg
sample size = n = 47
standard error = $\text{SEM}_2 = 3.7/\sqrt{47} = 0.540$

measure of variability $= \sqrt{[(0.633)^2 + (0.540)^2]} = 0.83$

Example 2: Weight Loss for Diet vs Exercise

Step 1. Determine the null and alternative hypotheses.

Null hypothesis: No difference in average fat lost in population for two methods. Population mean difference is ***zero***.

Alternative hypothesis: There is a difference in average fat lost in population for two methods. Population mean difference is not ***zero***.

Step 2. Collect and summarize data into a test statistic.

The sample mean difference = 5.9 – 4.1 = 1.8 kg
and the standard error of the difference is 0.83.

So the *test statistic:* $z = \dfrac{1.8 - 0}{0.83} = 2.17$

Example 2: Weight Loss for Diet vs Exercise

Step 3. Determine the *p*-value.

Recall the alternative hypothesis was two-sided.

p-value = 2 × [proportion of bell-shaped curve above 2.17]

Table 8.1 => proportion is about 2 × 0.015 = 0.03.

Step 4. Make a decision.

The ***p*-value of 0.03 is less than or equal to 0.05**, so …

- If really no difference between dieting and exercise as fat loss methods, would see such an extreme result only 3% of the time, or 3 times out of 100.
- Prefer to believe truth does not lie with null hypothesis. We conclude that there is a ***statistically significant difference between average fat loss for the two methods***.

Example 3: Public Opinion About President

On May 16, 1994, Newsweek reported the results of a public opinion poll that asked: *"From everything you know about Bill Clinton, does he have the honesty and integrity you expect in a president?"* (p. 23).

Poll **surveyed 518 adults and 233, or 0.45** of them (clearly less than half), answered yes.

Could Clinton's adversaries conclude from this that **only a minority (less than half) of the population** of Americans thought Clinton had the honesty and integrity to be president?

Example 3: Public Opinion About President

Step 1. Determine the null and alternative hypotheses.

Null hypothesis: There is no clear winning opinion on this issue; the proportions who would answer yes or no are each 0.50.

Alternative hypothesis: Fewer than 0.50, or 50%, of the population would answer yes to this question. The majority do not think Clinton has the honesty and integrity to be president.

Step 2. Collect and summarize data into a test statistic.

Sample proportion is: 233/518 = 0.45.

The *standard deviation* = $\sqrt{\frac{(0.50)\times(1-0.50)}{518}} = 0.022.$

Test statistic: $z = (0.45 - 0.50)/0.022 = -2.27$

Example 3: Public Opinion About President

Step 3. Determine the *p*-value.

Recall the alternative hypothesis was one-sided.

p-value = proportion of bell-shaped curve below –2.27

Exact *p*-value = 0.0116.

Step 4. Make a decision.

The ***p*-value of 0.0116 is less than 0.05**, so we conclude that the proportion of American adults in 1994 who believed Bill Clinton had the honesty and integrity they expected in a president was **significantly less** than a majority.

23.3 Revisiting Case Studies: How Journals Present Tests

Whereas newspapers and magazines tend to simply report the decision from hypothesis testing, **journals tend to report *p*-values as well**.

This allows you to make your own decision, based on the severity of a type 1 error and the magnitude of the *p*-value.

Case Study 6.1: Mozart, Relaxation, and Performance on Spatial Tasks

Three listening conditions— Mozart, a relaxation tape, and silence—and all subjects participated in all three conditions.

Null hypothesis: **No differences** in population mean spatial reasoning IQ scores after each of three listening conditions.

Alternative hypothesis: Population mean spatial reasoning IQ scores **do differ for at least one** of the conditions compared with the others.

Case Study 6.1: Mozart, Relaxation, and Performance on Spatial Tasks

*A one-factor (listening condition) repeated measures analysis of variance … revealed that subjects performed better on the abstract/spatial reasoning tests after listening to Mozart than after listening to either the relaxation tape or to nothing ($F[2,35] = 7.08$, **$p = 0.002$**).*

(Rauscher et al., 14 October 1993, p. 611)

Conclusion: At least one of the means differs from the others. If there were no population differences, sample mean results would vary as much as the ones in this sample did, or more, only 2 times in 1000 (0.002).

Case Study 6.1: Mozart, Relaxation, and Performance on Spatial Tasks

> *The music condition differed significantly from both the relaxation and silence conditions (Scheffé's $t = 3.41$, $p = 0.002$; $t = 3.67$, $p = 0.0008$, two-tailed, respectively). The relaxation and silence conditions did not differ ($t = 0.795$, $p = 0.432$, two-tailed).*
>
> *(Rauscher et al., 14 October 1993, p. 611)*

Significant differences were found between the music and relaxation conditions (*p*-value = 0.002) and between the music and silence conditions (*p*-value = 0.0008). The difference between the relaxation and silence conditions, however, was not statistically significant (*p*-value = 0.432).

Case Study 5.1: Quitting Smoking with Nicotine Patches

Compared the smoking cessation rates for smokers randomly assigned to use a nicotine patch versus a placebo patch.

Null hypothesis: The proportion of smokers in the population who would quit smoking using a nicotine patch and a placebo patch are the **same**.

Alternative hypothesis: The proportion of smokers in the population who would quit smoking using a **nicotine patch is higher** than the proportion who would quit using a placebo patch.

Case Study 5.1: Quitting Smoking with Nicotine Patches

Higher smoking cessation rates were observed in the active nicotine patch group at 8 weeks (46.7% vs 20%) ($P < .001$) and at 1 year (27.5% vs 14.2%) ($P = .011$).

(Hurt et al., 1994, p. 595)

Conclusion: *p*-values are quite small: less than 0.001 for difference after 8 weeks and equal to 0.011 for difference after a year. Therefore, rates of quitting are significantly higher using a nicotine patch than using a placebo patch after 8 weeks and after 1 year.

Case Study 6.4: Smoking During Pregnancy and Child's IQ

Study investigated impact of maternal smoking on subsequent IQ of child at ages 1, 2, 3, and 4 years of age.

Null hypothesis: Mean IQ scores for children whose mothers smoke 10 or more cigarettes a day during pregnancy **are same** as mean for those whose mothers do not smoke, in populations similar to one from which this sample was drawn.

Alternative hypothesis: Mean IQ scores for children whose mothers smoke 10 or more cigarettes a day during pregnancy are **not the same** as mean for those whose mothers do not smoke, in populations similar to one from which this sample was drawn.

Case Study 6.4: Smoking During Pregnancy and Child's IQ

Children born to women who smoked 10+ cigarettes per day during pregnancy had developmental quotients at 12 and 24 months of age that were 6.97 points lower (averaged across these two time points) than children born to women who did not smoke during pregnancy (95% CI: 1.62,12.31, P = .01); at 36 and 48 months they were 9.44 points lower (95% CI: 4.52, 14.35, P = .0002). *(Olds et al., 1994, p. 223)*

Researchers conducted *two-tailed tests* for possibility the mean IQ score could actually be *higher* for those whose mothers smoke. The CI provides evidence of the direction in which the difference falls. The *p*-value simply tells us there is a statistically significant difference.

For Those Who Like Formulas

Some Notation for Hypothesis Tests

The null hypothesis is denoted by H_0, and the alternative hypothesis is denoted by H_1 or H_a.

"alpha" = α = desired probability of making a type 1 error when H_0 is true; we reject H_0 if p-value $\leq \alpha$.

"beta" = β = probability of making a type 2 error when H_1 is true; power = $1 - \beta$

Steps for Testing the Mean of a Single Population

Denote the population mean by μ and the sample mean and standard deviation by $\overline{X}$ and s, respectively.

Step 1. H_0: $\mu = \mu_0$, where μ_0 is the *chance* or *status quo* value.

H_1: $\mu \neq \mu_0$ for a two-sided test; H_1: $\mu < \mu_0$ or H_1: $\mu > \mu_0$ for a one-sided test, with the direction determined by the research hypothesis of interest.

Step 2. This test statistic applies only if the sample is large. The test statistic is

$$z = \frac{\overline{X} - \mu_0}{s / \sqrt{n}}$$

For Those Who Like Formulas

Step 3. The p-value depends on the form of H_1. In each case, we refer to the proportion of the standard normal curve above (or below) a value as the "area" above (or below) that value. Then we list the p-values as follows:

Alternative Hypothesis	***p*-Value**
H_1: $\mu \neq \mu_0$	$2 \times$ area above $\lvert z \rvert$
H_1: $\mu > \mu_0$	area above z
H_1: $\mu < \mu_0$	area below z

Step 4. You must specify the desired α; it is commonly 0.05. Reject H_0 if p-value $\leq \alpha$.

For Those Who Like Formulas

Steps for Testing a Proportion for a Single Population

Steps 1, 3, and 4 are the same, except replace μ with the population proportion p and μ_0 with the hypothesized proportion p_0. The test statistic (step 2) is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

Steps for Testing for Equality of Two Population Means Using Large Independent Samples

Steps 1, 3, and 4 are the same, except replace μ with $(\mu_1 - \mu_2)$ and μ_0 with 0.

Use previous notation for sample sizes, means, and standard deviations; the test statistic (step 2) is:

$$z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s_1^{\,2}}{n_1} + \dfrac{s_2^{\,2}}{n_2}}}$$