Lecture 21

Significance, Importance, Undetected Differences

Thought Question 1:

Which do you think is **more informative** when you are given the results of a study: a **confidence interval** or a ***p*-value**?

Explain.

Thought Question 2:

Suppose you were to read that a new study based on 100 men had found that there was ***no difference*** in heart attack rates for men who exercised regularly and men who did not.

What would you suspect was the reason for that finding?

Do you think the **study found *exactly* the same rate** of heart attacks for the two groups of men?

Thought Question 3:

Example in Chapter 23 used results of public opinion poll to conclude that a majority of Americans did not think Bill Clinton had honesty and integrity they expected in a president.

Would it be fair reporting to claim that "significantly fewer than 50% of American adults in 1994 thought Bill Clinton had the honesty and integrity they expected in a president"? Explain.

Thought Question 4:

When reporting the results of a study, explain why a distinction should be made between "**statistical significance**" and "**significance**," as the term is used in ordinary language.

Thought Question 5:

Remember that a **type 2 error** is made when a study fails to find a relationship or difference when one actually exists in the population.

Is this kind of error **more likely** to occur in studies **with large samples or with small samples**?

Use your answer to explain why it is important to learn the size of a study that finds no relationship or difference.

24.1 Real Importance versus Statistical Significance

A **statistically significant** relationship or difference ***does not necessarily mean an important one***.

Whether results are statistically significant or not, it is helpful to examine a confidence interval so that you can determine the ***magnitude*** of the effect.

From **width** of the confidence interval, also learn how much **uncertainty** there was in sample results.

Example 1: Is the President that Bad?

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 **45% answered yes**, 238 answered no (46%), and rest unsure.

Test indicated proportion of population who would answer yes was statistically significantly less than half.

Can we report: significantly less than 50% of all American adults in 1994 thought Bill Clinton had the honesty and integrity they expected in a president?

Example 1: Is the President that Bad?

What the Word *Significant* Implies

95% confidence interval:
sample value ± 2 (standard deviation)
0.45 ± 2(0.022) => 0.45 ± 0.044 => 0.406 to 0.494

True proportion could be as high as 49.4%!
Although less than 50%, it is certainly not "significantly less" in the usual, nonstatistical meaning of the word.

Example 1: Is the President that Bad?

Consider testing about proportion who answer no:

Null hypothesis: The population proportion who would answer no is 0.50.

Alternative hypothesis: The population proportion who would answer no is **less than** 0.50.

Test statistic = –1.82 and *p*-value = 0.034

We would accept hypothesis that less than a majority would answer no. We have now found that less than a majority would answer yes and less than a majority would answer no.

The Importance of Learning the Exact Results

The problem is that only 91% of the respondents gave a definitive answer. The rest of them had no opinion.

Example 2: Is Aspirin Worth the Effort?

Relationship between taking aspirin and incidence of heart attack. Null (no relationship) vs alternative (yes relationship), chi-squared (test) statistic over 25 with p-value < 0.00001.

The Magnitude of the Effect

The test statistic and p-value do not provide information about the ***magnitude*** of the effect.

Representing the Size of the Effect

Rates of heart attack: 9.4 per 1000 for aspirin group and 17.1 per 1000 for placebo group, difference < 8 people per 1000, about 1 less heart attack for every 125 who took aspirin.

Relative risk: Aspirin group had half as many heart attacks; so could cut risk almost in half. Estimated relative risk as 0.53, with a 95% confidence interval extending from 0.42 to 0.67.

24.2 Role of Sample Size in Statistical Significance

If the sample size is **large enough**, almost **any null hypothesis can be rejected**.

There is almost always a ***slight*** relationship between two variables, or a difference between two groups, and if you ***collect enough data, you will find it***.

Example 3: How The Same Relative Risk Can Produce Different Conclusions

Study relationship between breast cancer and age at which women has first child.

		Developed Breast Cancer?		
		Yes	No	Total
First Child Before Age 25?	Yes	65	4475	4540
	No	31	1597	1628
	Total	96	6072	6168

Chi-squared test statistic = 1.746 and *p*-value = 0.19, so not statistically significant even though relative risk = 1.33.

If ***three times as many women sampled***, but observed same pattern (relative risk still 1.33), then the chi-squared statistic would be 3(1.746) = 5.24 with a *p*-value of 0.02; so would declare a statistically significant relationship.

24.3 No Difference versus No Statistically Significant Difference

If the sample size is **too small**, an important relationship or difference can go **undetected**.

In that case, we would say that the ***power*** of the test is ***too low***.

Example 4: All that Aspirin Paid Off

Physician's Health Study (1988): 22,071 participants.
Chi-squared statistic = 25.01 highly statistically significant!

If ***only 3000 participants***, but observed same pattern. Results are shown below: Chi-squared statistic = 3.65 with *p*-value of 0.06; so result not statistically significant using the 0.05 level.

	Heart Attack	No Heart Attack	Total	Percent Heart Attacks	Rate per 1000
Aspirin	14	1486	1500	0.93	9.3
Placebo	26	1474	1500	1.73	17.3
Total	40	2960	3000		

Case Study 24.1: Seen a UFO? You May Be Healthier Than Your Friends

Are people who claim to have seen a UFO psychologically disturbed and prone to fantasy?

- Recruited 49 volunteers who claimed to have seen a UFO by placing newspaper ad, “Researcher seeks adults who have seen UFOs. Confidential”.
- Recruited 53 community members using ad that read “for personality study” instead of “seen UFOs.”
- Recruited 74 students who received course credit.

All given tests and questionnaires measuring psychological health, intelligence, imaginal propensities, and more.

Source: Spanos et al., 1993, p. 625.

Case Study 24.1: Seen a UFO? You May Be Healthier Than Your Friends

A study of 49 people who have reported encounters with unidentified flying objects, or UFOs, has found no tendency toward abnormality, apart from a previous belief that such visitations from beyond the earth do occur. . . . The tests [given to the participants] included standard psychological tests used to identify subjects with various mental disorders and assess their intelligence. The UFO group proved slightly more intelligent than the others.

(*NYT*, Sullivan, 29 November 1993, p. 37)

Might think: no statistically significant differences found in psychological health of groups, although perhaps a significant difference in intelligence. **Not So!**

Case Study 24.1: Seen a UFO? You May Be Healthier Than Your Friends

Alternative hypothesis was *one-sided:* UFO observers ***less*** healthy. Data indicated might be healthier – not consistent, so null hypothesis could not be rejected.

The most important findings indicate that neither of the UFO groups scored lower on any measures of psychological health than either of the comparison groups. Moreover, both UFO groups attained higher psychological health scores than either one or both of the comparison groups on five of the psychological health variables. In short, these findings provide ***no support whatsoever*** *for the hypothesis that UFO reporters are psychologically disturbed.*

(*Source:* Spanos et. al., 1993, p. 628)

24.4 A Summary of Warnings

1. If word ***significant*** used to convince you there is an important effect/relationship, determine if the word used in usual sense or statistical sense only.
2. If study based on very **large sample size**, relationships found to be ***statistically significant*** may not have much ***practical importance***.
3. If "**no difference**" or "no relationship" found in a study, determine sample size used. Unless sample size is large, an important relationship may well exist in population but not enough data collected to detect it, (i.e. test could have very ***low power***).

A Summary of Warnings

4. If possible, learn what **confidence interval** accompanies the hypothesis test, if any. Can be misled into concluding no effect when there really is, but will have more information about **magnitude** of possible difference or relationship.
5. Determine whether test was **one-sided or two-sided**. If one-sided, as in Case Study 24.1, and details aren't reported, could be misled into thinking no difference, when in fact there was one in direction opposite to that hypothesized.

A Summary of Warnings

6. Decision to do one-sided must be made ***before*** looking at data. Using same data to both generate and test the hypotheses is cheating!
7. Sometimes researchers perform a **multitude of tests**, and reports focus on those that achieved statistical significance. Remember if nothing interesting is happening and all null hypotheses tested are true, then 1 in 20 tests should achieve statistical significance just by chance. Beware of reports where many tests conducted, but results of only one or two are presented as “significant.”