Chapter 29: A Closer Look at Tests of Significance

When performing a test of significance, we should:

- summarize the data
- say which test was used, 1-tailed or 2-tailed
- report a P-value

The 1% and 5% rules are guidelines only – a P-value of 4.9% is not very different from one of 5.1% and should not be treated differently.

The **P-value** tells you how likely the result is, under the null hypothesis. It does NOT tell you about the :

- importance of the result
- strength of a relationship
- reliability of the design of an experiment

To decide how important a result is, or how strong a relationship is, think of it applying to the whole population and think about the real-world consequences. e.g. if increasing exposure to a toxin 100-fold increases cancer rates by only 1%, it is not as important as a treatment that can cut fatalities by 30%.

To decide how good an experiment is, ask questions about how the study was conducted.

Data Snooping!

If we do a LOT of statistical tests, we expect to find some “statistically significant” results due to chance error.

e.g. testing at the 5% level, we expect 5% of our tests to show up as “statistically significant” just due to chance, even if ALL null hypotheses are true!

- if we do 100 tests, we expect 5 “false positives”
- if we do 500 tests, we expect 25 “false positives”

Always report how many tests you do, not just the ones that are “statistically significant”. There are ways to adjust the p-values to take into account how many tests you have done (beyond the scope of this class).

Replication

If we do a lot of tests, and think we have found something important, we can replicate the study to convincingly show the result.

Important studies that are not controlled experiments (e.g., studies on the relationship between smoking and lung cancer, heart disease, etc.) become convincing when

1. they are replicated and they show consistent effects,
2. The effects respond appropriately to dose (e.g. higher doses show higher rates of disease), and
3. Whenever possible, they are confirmed with clinical trials and lab experiments.

Sample Size

If we have a LARGE SAMPLE, even tiny differences can show up as “statistically significant” – they might not be important.

If we have a SMALL SAMPLE, even an important difference might not show up as “statistically significant” (we say the test lacks “power”).

Singling out the Worst

If we notice something unusual has happened, it does not make much sense to figure out the chances – rare things happen all the time.

e.g. In a town of 60,000 people someone notices that the rate of a rare form of cancer is 3.5 times the national average. They perform a statistical test, and find P-value = 0.02%. They find high-voltage power lines and conclude these caused the cancer.

What's wrong with this?

1-tailed vs 2-tailed

Researchers like 1-tailed tests because they are more likely to get “statistically significant” results. HOWEVER, the decision should be made BEFORE looking at the data and if there is any doubt about which direction to go, it should be a 2-tailed test.

e.g. the HIV example from the homework was designed as a 1-tailed test, but the data went the opposite way to what they expected.

Box Models

You MUST HAVE a box model. Watch out for:

- tests where the data are the whole population (especially 2-sample z tests and chi-square independence tests)
- samples of convenience (it is not valid to do a test)
- badly designed experiments (you might not be testing what you think you are testing)