Ch. 29: A Closer Look at Tests of Sifnificance

Always report the P-value of a test, don't just say " the result was statistically significant".

Also:

- Summarize data
- Say which test was used.
- Report *exact* P-value whenever possible (not just P-value $< 5\%$ or P-value $< 1\%$).

Is the Result Significant?

We reject the null hypothesis if the P-value is large / small (circle one)

If the P-value is less that _____ the result is "statistically significant".

If the P-value is 4.9% we reject / do not reject null hypothesis (circle one).

If the P-value is 5.1% we reject / do not reject null hypothesis (circle one).

Is this really such a difference????

Moral:
Do not take the 5% and 1% levels too seriously!

Doing Many Tests

If the null hypothesis is true, we have a ____ % chance of rejecting it.

If we do 100 tests and all the null hypotheses are true, we expect to reject _____ of them!

Therefore, some of these null hypotheses might be rejected just by chance!

Moral:
Report *all* tests, not just significant ones!

Data Snooping

Usually, researchers decide which hypotheses to test only after they have seen the data.

This is called _____.

Data snooping makes P-values hard to interpret.

Ex: Test whether a coin is fair. In 100 tosses, we get 61 H's.

Null hypothesis:

Alternative hypothesis:
The coin is biased: Chance of H is *over* 50%.

or

The coin is biased: Chance of H is *not* 50%.

Note:

- Researchers like one–tailed tests because it is easier to get "significant results".

- It is more correct to decide which test to use before we look at the data.

- If there is any doubt, a two–tailed test should be done.

Is the result important?

Ex: Differences between rural and urban children on a vocabulary test average 1 point out of 50. The sample sizes are large, so the SE's are both very small, and this result is highly *statistically* significant, but of no practical importance.

Ex: Factory components differ in average size by day of the week (Monday, Wednesday, or Friday) but the differences were well within the tolerances for the component, and hence of no practical significance.

For a **large sample**, even tiny differences can be statistically significant, but it does not mean they are important.

For a **small sample**, even an important difference may not be statistically significant.

In the End ...

Watch out for:

- Tests where the data are the whole population (especially two–sample tests, χ^2 – tests...)

  IF YOU CAN'T MAKE A BOX MODEL, DON'T DO A SIGNIFICANCE TEST!

- Non – random samples (especially convenience samples);

- Badly designed experiments.

— The End —