

Dec. 8 Statistic for the day:

In Britain, a man spends two years over the course of an average lifetime going to the toilet while a woman spends only 6 months.

Expected number of quizzes this semester, conditional on what's happened so far: 14

Most likely $P(13 \text{ quizzes}) = 29.6\%$;

possibilities: $P(14 \text{ quizzes}) = 44.4\%$;

$P(15 \text{ quizzes}) = 22.2\%$

Assignment:

Complete online practice problems for Chapters 23-25

Recall: Power of a test

The **power** of a statistical test refers to its ability to detect when something is happening – in other words, to reject the null hypothesis when it should be rejected.

Power as it relates to type 2 errors

If the null hypothesis is false:

- Rejecting it is a good thing.
- Failure to reject is a type 2 error.
- The probability of rejecting is the power.

Thus, high power corresponds to low probability of type 2 error.

Eliminating type 1 or type 2 error: A couple of really stupid tests

Test A: Always reject the null hypothesis, no matter what the data are.

This test will never make a type 2 error!
It has perfect power! (But it's stupid.)

Test B: Never reject the null hypothesis, no matter what the data are.

This test will never make a type 1 error!
But it has no power, and it's stupid.

Ways to increase power

- Increase the sample size
- Increase the probability of a type 1 error (set the p-value cutoff higher, making it easier to reject the null hypothesis).

You can control these

Also note: Power depends on how close the true population value is to the null value: The farther apart they are, the easier it is to detect the difference, so the greater the power.

You can think about this but not control it

Recall: Warning #7. Sometimes researchers perform a multitude of tests, and the reports focus on those that achieved statistical significance. If all of the null hypotheses tested are true, then 1 in 20 tests should achieve statistical significance just by chance.

Beware of reports where it is evident that many tests were conducted, but where results of only one or two are presented as "significant".

Beware multiple p-values computed simultaneously!

A commonly used approach to correcting p-values when multiple simultaneous tests are run is called a *Bonferroni* correction:

Just multiply each p-value times the number of tests. Then ask if they're still significant.

Meta-analysis

- A collection of statistical techniques for combining studies.
- By combining many studies, we may sometimes be able to obtain a large "meta-study" that helps to answer difficult questions that are not clear from smaller studies.

Vote-counting example (p. 436, #14)

Recall that vote-counting is a poor way to combine studies.

Suppose ten studies were done to assess the relationship between watching violence on television and subsequent violent behavior in children. Suppose that none of the ten studies detected a significant relationship.

No → Is it possible for a vote-counting procedure to detect a relationship? Is it possible for a meta-analysis to detect a relationship? Explain.

Yes → Using just vote-counting, we see 0 out of 10 significant results! However, a meta-analysis can combine these ten studies, giving in effect one large sample that might be enough to show a statistically significant effect.

Which studies should be included in a meta-analysis?

Different studies may differ widely in their quality of work. Often, many studies must be eliminated from a meta-analysis because it is not absolutely clear that what is being studied in them is the desired focus of the research.

A meta-analysis of the effect of behavior on blood pressure eliminated all but 26 out of 857 possible studies!

Should studies be compared or combined?

If one wishes to combine studies, make sure they're really measuring the same thing on the same population!

Consider two studies comparing surgery to relaxation for treating chronic back pain. One is conducted at a back-care specialty clinic, the other at a suburban medical center.

Where will the people with the most severe back pain go? The two studies are probably conducted on different populations. (We'll revisit this example later..)

Is smoking related to lower sperm count in men?

One study found a 22.8% reduction in sperm count for smokers, but it only used 88 subjects and the finding was not statistically significant.

An accompanying meta-analysis estimated a similar reduction, but with the power of the combined studies, the p-value was found to be less than 0.0001.

(Remember, these findings are based on observational studies and do not imply causation.)

Are mammograms an effective screening device for women aged 40-49?

A 1993 meta-analysis said NO.

This raises a potential problem with meta-analyses: The possibility of type 2 errors might be ignored because it seems unlikely that such a large study could miss any significant result!

The 1993 meta-analysis did not dissuade the American Cancer Society from recommending mammograms for women 40-49. The ACS and others have pointed to various potential flaws with the meta-analysis.

Potential problems with meta-analysis

- Simpson's paradox
- Confounding variables
- Subtle differences in treatments of the same name
- File drawer problem
- Biased or flawed original studies
- Statistical significance vs. practical significance
- False findings of "no difference"

Simpson's paradox

Simpson's paradox can result when groups are inappropriately combined (obviously an issue in meta-analysis).

Recall the law-school/business-school example we viewed in class a while back, where higher proportions of WOMEN were admitted to BOTH schools. However, because men tended to apply to the easier law school and women to the more difficult business school, when the admissions data were combined a higher proportion of MEN was admitted overall (!)

Incidentally, there is really no "paradox" here once you understand it!

Simpson's paradox cont'd

Consider the hypothetical example discussed on p. 425. I'll invent some numbers:

At the back-care specialty clinic, 100 people had surgery (20 improved) and 100 used relaxation (15 improved).

Improvement rates: 20% for surgery, 15% for relaxation

At the suburban medical center, 20 people had surgery (16 improved) and 100 used relaxation (75 improved).

Improvement rates: 80% for surgery, 75% for relaxation

Overall: 36/120 improved with surgery (that's 30%)
90/200 improved with relaxation (that's 45%)

COMBINING THESE TWO STUDIES IS DANGEROUS!

The file drawer problem

A meta-analysis can ONLY rely on those studies that have been published. But this might be a biased sample of studies!

Why? Because those studies that indicate a significant finding are more likely to be published.

(Just imagine all of those unpublished studies sitting in **file drawers** that do not indicate a significant finding!)

Statistical vs. practical significance

We have already discussed the difference between the two. Just because a difference is found to be statistically significant, that doesn't mean that the difference is meaningful from a practical point of view.

Suppose that we observe in a really large study of smokers in the U.S. that men smoke on average 0.35 more cigarettes per month than women, and this difference is statistically significant.

WHO CARES?

Because it combines the power of many studies of the same phenomenon, a meta-analysis is especially susceptible to finding differences statistically significant that are not actually of any interest practically.

Quiz we didn't have

Suppose that in a representative sample of size 100 students, we observe a mean weight of 150 pounds with a standard deviation of 40 pounds.

Derive an interval of "reasonable values" for the true population mean, equal to the sample mean plus or minus two times the standard deviation of the sample mean.

Quiz we didn't have

The player in a game of craps has a 0.493 probability of winning the game.

Suppose that a player plays 100 games of craps. What does the "rule of sample proportions" say about the proportion of times this player will win?

(Your answer should include something about the shape, center, and spread of the distribution of this proportion.)

Quiz we didn't have

Research question: Do males and females at PSU differ with respect to average credits?

Variable	Sex	N	Mean	StDev	SE Mean
Credits	Female	138	15.82	1.76	0.15
	Male	81	15.57	2.16	0.24

Show how to compute the test statistic here. You don't have to find the value, just set up the calculation.