Understanding Statistics & Experimental Design

Some trick(y) questions

Q1 “Statistically speaking: What is the correct interpretation of a p-value, what does it tell you?”

Tick all correct statements:

A. P is the probability that the result observed is due to chance.
B. A large value of P, say for a test that μ = 0, would suggest that the mean $\bar{x}$ actually recorded was due to chance, and could be assumed to be zero (Schmidt and Hunter 1997).
C. P is the chance of observing a sample mean this or more extreme if in fact H0 is true (assuming all assumptions of the test are met: normality, etc.)
D. 1-p is considered the reliability of the result; that is, the probability of getting the same result if the experiment were repeated.
E. P can be treated as the probability that the null hypothesis is true.
F. P is the Pr[observed or more extreme data I H0], the probability of the observed data or data more extreme, given that the null hypothesis is true, the assumed model is correct, and the sampling was done randomly. (Johnson, 1999)
G. All of the above.
H. None of the above.

Q1 Answer:

- C and F

- Source: Johnson 1999 http://www.jstor.org/stable/3802789;
- http://www.methodspace.com/profiles/blogs/should-i-buy-this-book

Q2

Imagine there were 10 studies (you can assume they are of a suitably high quality with no systematic differences between them). They have a measure of constipation as their outcome (let's assume it's a continuous measure). A positive difference between means indicates that the intervention was better than the control group at reducing constipation.

(Continues on next slide…)

Source: Andy Field

Q2

- Here are the results:

Study	Difference between Means	t	p
Study 1	4.193	3.229	0.002*
Study 2	2.082	1.743	0.086
Study 3	1.546	1.336	0.187
Study 4	1.509	0.890	0.384
Study 5	3.991	2.894	0.006*
Study 6	4.141	3.551	0.001*
Study 7	4.323	3.745	0.000*
Study 8	2.035	1.479	0.155
Study 9	6.246	4.889	0.000*
Study 10	0.863	0.565	0.577

(Continues on next slide…)

Q2

Mark the single correct statement:

A. The evidence is equivocal, we need more research.

B. All of the mean differences show a positive effect of the intervention, therefore, we have consistent evidence that the treatment works.

C. Five of the studies show a significant result ($p < .05$), but the other 5 do not. Therefore, the studies are inconclusive: some suggest that the intervention is better than TAU, but others suggest there's no difference. The fact that half of the studies showed no significant effect means that the treatment is not (on balance) more successful in reducing symptoms than the control.

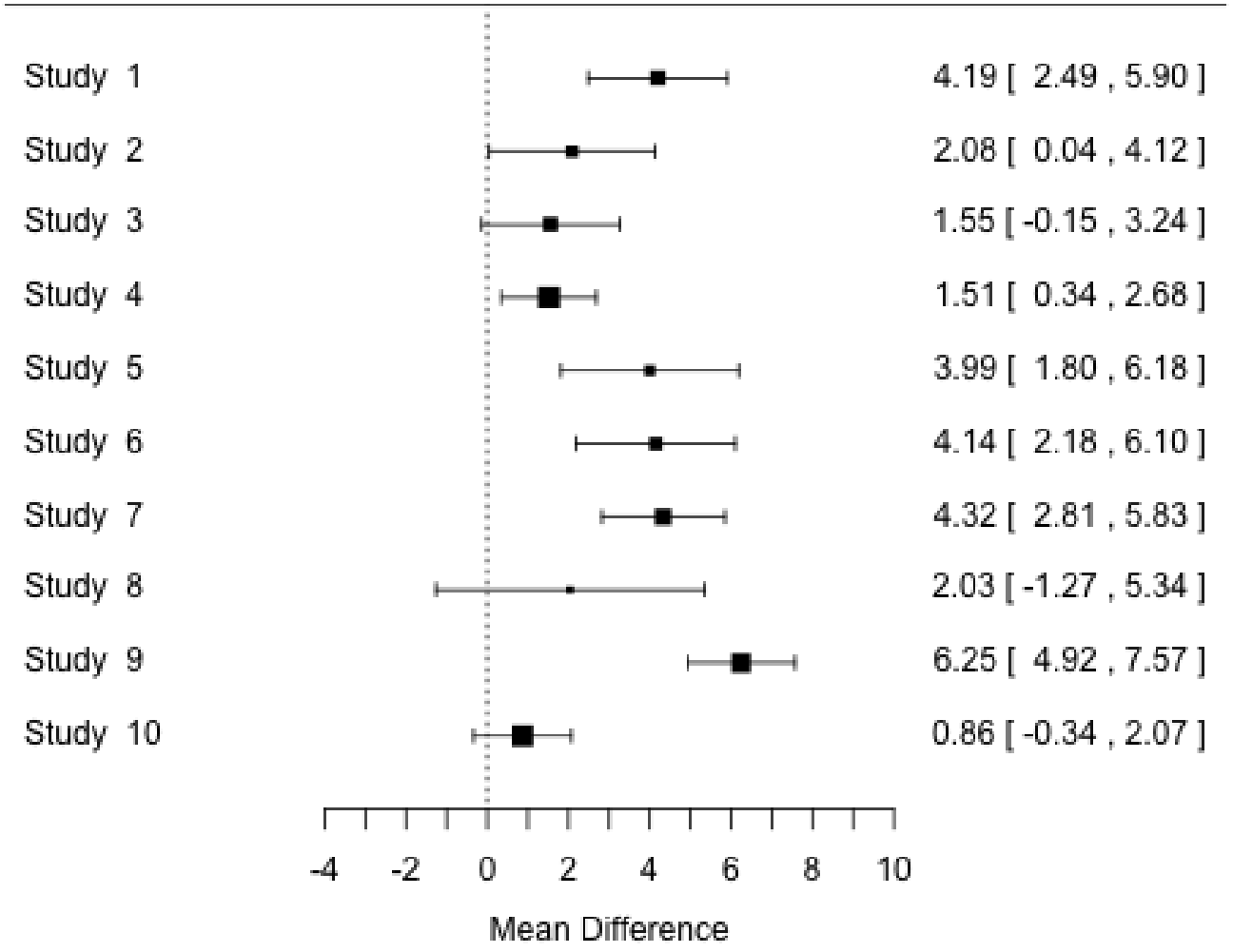
D. I want to go for C, but I have a feeling it's a tricky question.

Answer Q2:

B is the correct answer, which becomes immediately clear when you plot the confidence intervals of the studies.

This is the reason we should always report CIs with out p values.

Source: Andy Field
http://www.methodspace.com/profiles/blogs/the-joy-of-confidence-intervals



Q3:
“What can you conclude from a p-value of p = .2?”

Mark the correct answer(s)

A) Your sample size was not big enough to find the hypothesized effect.

B) The effect you are trying to find is too small to be detected with the given sample size.

C) The effect you are trying to find does not exist.

D) You will have a hard time getting this published.

Q3 answer:

All might be correct. Or none. In fact non-significant p-values don’t tell us much, especially when we do not know the sample-size.

Q4

You have access to a huge database containing 10 years of GPS data of 1,103,086 drivers, alongside personality tests. You test the hypothesis that drivers who fall into the category of extroverts drive faster on average than introverts. You find that the effect is highly significant ($p < .00001$). What do you conclude?

A. Extroverts are indeed driving faster than introverts. The effect is large, since p is very far below the alpha threshold of 0.05.
B. Extroverts are indeed driving faster than introverts. The effect is small, since p is very far below the alpha threshold of 0.05.
C. Extroverts are indeed driving faster than introverts. We don’t know the effect size.
D. With a sample this big, even the tiniest effect will be highly significant. Therefore, extroverts are indeed driving faster than introverts, but only very little (small effect size).

Q4: Answer

- C is right – the effect size is not given. The difference in driving speed might or might not be big, or too tiny to see when looking at individual drivers. As long as we don't know the effect size we don't even know if we should be interested in this finding.
- While the former statement of D is correct, the latter is not: We cannot conclude that the effect is small just because the sample is large.
- For A and B the former part is likely true. But again, we don't know the effect size, so we don't know if the difference has any practical importance. If extraverts drive 0.1km/h faster on average, I'm pretty sure nobody gives a damn.

- See e.g.,
  Kalinowski & Fidler 2010 Interpreting Significance - The Differences Between Statistical Significance, Effect Size and Practical Importance

Q5:

“How high should the p value be before it is reasonable to assume that H0 (there is no effect) is true?”

Q5 answer:

None. You're trying to prove the null hypothesis, which is impossible with a p value.

A non-significant p value tells you that there is *(too) little evidence* to reject H0 in favor of H1. It does **NOT** tell you that there is *evidence for* H0.

Sentences in papers like "Transgenic and wild type mice did not behave differently, as the t-test comparison showed a non-significant p value" are utter nonsense!

Q6:

You compared the effect of a new drug on two groups of people (men and women). 30 men and 30 women participate in the trial and are assigned randomly to a group taking the drug or a group taking placebo pills. A ‘blind’ doctor rates their health after a few weeks of being treated with the new drug or placebo pills (control groups).

Women taking the drug are rated healthier than women taking the placebos ($p = .001$). Men taking the drug are rated healthier than men taking the placebos ($p = .12$).

Which statements are true:

A. The drug seems to be effective for women, but not for men.
B. The drug seems to be effective for women, evidence is not clear for men.
C. The drug is more effective for women than for men.

Answer Q6:

A is the common conclusion from this kind of results, based on the fact that one effect is significant (women) and one is non-significant (men). However, non-significance does not prove that there is no effect…

B could be argued to be true, since the difference for men approaches significance. There *might* be an effect that was to small to be detected with the given sample size (lack of power), or the p value here is large by chance.

C is wrong, as it is an untested statement. First, a p value says nothing about effect size. Second, the reported p values stem from tests of treatment vs control group within each sex – what you would need to statement C is a direct comparison of the effect in men vs the effect in women.

For example, that the comparison with the control group passes the significance threshold for women, but not for men, might be solely due to higher variance in men, while the effect size might be the same or even higher.

Q7:

You have found a new drug for an old disease. You test your new drug against a placebo control group and find that it is very effective:

The treated people are rated way healthier than the placebo people, $p < .0001$, Cohen's $d = 0.8$ (large effect).

Should you take this drug to the market?

Answer Q7

Partial information!!

What is most interesting is not whether the drug is successful compared to placebo (no drug), but when compared to the drugs that are already on the market!

Unless your drug is way cheaper, there is no justification in developing an equally good / worse drug.

Would you invest in development of a new Aspirin?

See e.g., http://www.ted.com/talks/ben_goldacre_battling_bad_science?language=en

Q8

You are charged with improving road safety on the streets of canton Vaud. The canton gives you a list of accident foci. It turns out that with 103 deaths per year, the Bourdonette intersect is *the* traffic accident hotspot of the canton. Being super smart, you go look at the place and realize that most accidents might be due to people driving too fast when coming from EPFL/Unil and the highway. You install speed-traps to enforce the legal speed limits and lean back in your chair to await the new accident reports. A year later, the number of accidents at Bourdonette have decreased by 25%!

What conclusions are correct?

A. You have done a good job decreasing traffic accidents.
B. You have not done a good job, accident rate is still high.
C. You don’t know if you have done a good job until you run a t-test to test the effect.
D. More research is needed.

Q8 Answer

A. Maybe true
B. True
C. Bullshit. What is the kind of question you could ask in statistical terms? You can't just compare 103 deaths to 103 – 25% deaths, as these are single numbers, no sample means, no variation.
D. True: You have selected to study the Bourdonette intersections due to its *extremity*, i.e., because accident rates were already super high. If you had done *nothing*, the death-rate would have very likely decreased as well due to *regression to the mean*. What you need is one or more control groups/conditions/intersections.

See, e.g., https://en.wikipedia.org/wiki/Regression_toward_the_mean

Q9

You are planning an fMRI study into effect A. You delve into the literature to find out about the effect size of A that is usually reported.

You use this effect size to calculate the number of participants (samples) you will need in your experiment.

You arrive at N=28 to have a decent 80% chance to detect the effect if it has the same size as previously reported.

You take this to your supervisor and he starts laughing, because the scanner operator charges him 200CHF/hour and you just asked him for 8400CHF to run your experiment (N * 1.5h/subj * 200CHF = 8400CHF).

He says you can maybe have 8-10 subjects because it's too expensive otherwise and usually 8-10 is enough (it's what other people do).

What are your options?

Q9: Answer

- Convince him to give you more money.
- Get more money yourself.
- Seek collaboration with other scanner centers and pool subjects.
- **Do not run the study, as the outlook of success is low.**

- Some scientific questions are just not worth the money it would cost to test them *thoroughly.*

To be fair, in practice people *do* run this sort of experiments without any bad consequences for their careers (quite to the contrary). But for the field it's detrimental, as it leads to an accumulation of uncertain knowledge / potentially "unreal" effects like Bem's precognition (seeing into the future).

Q10

Same scenario:

You are planning an fMRI study into effect A. You delve into the literature to find out about the effect size of A that is usually reported.

You use this effect size to calculate the number of participants (samples) you will need in your experiment.

You arrive at N=28 to have a decent 80% chance to detect the effect if it has the same size as previously reported.

You take this to your supervisor and he starts laughing, because the scanner operator charges him 200CHF/hour and you just asked him for 8400CHF to run your experiment (N * 1.5h/subj * 200CHF = 8400CHF).

He says you should start measuring 8 subjects, as you will not be able to publish with less.
To be cost-effective, he suggests that if the effect is not yet significant after the first 8, you add subjects one by one until your power is high enough and you detect the effect.

How do you react? What are the problems?

Q10: Answer

Your advisor is asking you to “data-peek” / perform “optional stopping”.

The consequences were treated during the last few lectures by Prof. Francis.

This is a tricky situation: Your advisor with all his experience will not like you lecturing him/her on stuff he/she (and everybody else) has been doing for years. Yet, since you took this course, you *know* that it’s wrong. So how can you keep on doing your work with clean conscience without jeopardizing the relationship with your advisor? We have no advice on this one I’m afraid…

Q11:

Suppose I set up a study to see if one group (e.g. men) differs from another (women) on brain response to auditory stimuli (e.g. standard sounds vs deviant sounds – a classic mismatch negativity paradigm). I measure the brain response at frontal and central electrodes located on two sides of the head. The nerds among my readers will see that I have here a four-way ANOVA, with one between-subjects factor (sex) and three within-subjects factors (stimulus, hemisphere, electrode location). My hypothesis is that women have bigger mismatch effects than men, so I predict an interaction between sex and stimulus, but the only result significant at $p < .05$ is a three-way interaction between sex, stimulus and electrode location. What should I do?

A. Describe this as my main effect of interest, revising my hypothesis to argue for a site-specific sex effect
B. Describe the result as an exploratory finding in need of replication
C. Ignore the result as it was not predicted and is likely to be a false positive

Source: Deborah Bishop

Q11: Answer

C is the correct answer, while I believe that you wouldn't be shot doing B either. It's just A that's really evil.

We have learned that an ANOVA is a solution to the multiple testing problem (instead of running 5 t-tests with each a 5% risk of a type I error, you just run one ANOVA). However, ANOVAs only do this for the number of "levels" within one "factor" - for the type I error rate it does not matter if you run a one-way ANOVA with 3 or 300 levels. However, there is no correction for the number of factors: The chance that *something* is significant just by chance is higher in a two-way ANOVA than a one-way ANOVA, etc.

In the words of Deborah Bishop http://deevybee.blogspot.ch/2013/06/interpreting-unexpected-significant.html

- ANOVA adjusts for the number of **levels** within a factor, so, for instance, the probability of finding a significant effect of group is the same regardless of how many groups you have. ANOVA makes **no** adjustment to p-values for the number of factors and interactions in your design. The more of these you have, the greater the chance of turning up a "significant" result.
- So, for the example given above, the probability of finding **something** significant at .05, is as follows:
- For the four-way ANOVA example above, we have 15 terms (four main effects, six 2-way interactions, four 3-way interactions and one 4-way interaction) and the probability of finding no significant effect is .95^15 = .46. It follows that the probability of finding **something** significant is .54.
- And for a three-way ANOVA there are seven terms (three main effects, three 2-way interactions and one 3-way interaction), and p (something significant) = .30.

Q12

What does it mean if a result is reported as significant at $p < 0.05$?

A. If we were to repeat the analysis many times, using new data each time, and if the null hypothesis were really true, then on only 5% of those occasions would we (falsely) reject it.
B. Without knowing the statistical power of the experiment, and not knowing the prior probability of the hypothesis, I cannot estimate the probability whether a significant research finding ($p < 0.05$) reflects a true effect.
C. The probability that the result is a fluke (the hypothesis was wrong, the drug doesn't work, etc.), is below 5 %.

Q 12: Answer

- A and B

- Source: http://dirnagl.com/2014/09/22/p-value-vs-positive-predictive-value/#more-537

Q13

You run a study and find a significant negative correlation between brain size and intelligence. What do you conclude?

A. Brain size does not influence intelligence (as indicated by the absence of positive correlation).
B. Brain size negatively affects intelligence (as indicated by the presence of a negative correlation).
C. Brain size positively affects intelligence (correlation is significant).
D. None of the above.
E. All of the above.

Q13: Answer

The correct answer is D. Here's why: A negative correlation means that either smaller brains usually go together with higher intelligence and/or vice versa. It does NOT tell us if …

1) brain size affected intelligence
2) intelligence affected brain size
3) a third ("hidden") factor is the reason for what appears to be a relation between brain size and intelligence
4) If the effect is even interesting → effect size is missing here (i.e., the strength/slope of the correlation)

(Continues on next slide …)

Q13: Answer

Answers A, B and C were all *directional*, stating that “A influences B”, instead of “here is some kind of link between A and B”. Also:

A Brain size does not influence intelligence (as indicated by the absence of positive correlation).

→ a negative correlation is a correlation, too

B Brain size negatively affects intelligence (as indicated by the presence of a negative correlation).

→directional statement and hence false

C Brain size positively affects intelligence (correlation is significant).

→ Plain wrong, and who cares about significance anyway?

For some funny accidental (“spurious”) correlations check out http://www.tylervigen.com/