Single sample hypothesis testing

9.07

3/02/2004

Statistical Inference

We generally make 2 kinds of statistical inference:

1. We estimate some population parameter using confidence intervals and margins of error.
2. We evaluate data to determine whether it provides evidence for some claim about the population. *Significance testing*.

Recall from our first class: Chance vs. systematic factors

- A *systematic* factor is an influence that has a predictable effect on (a subgroup of) our observations.
  - E.G. a longevity gain to elderly people who remain active.
  - E.G. a health benefit to people who take a new drug.
- A *chance* factor is an influence that contributes haphazardly (randomly) to each observation, and is unpredictable.
  - E.G. measurement error

Observed effects can be due to:

A. Systematic effects alone (no chance variation).

- We’re interested in systematic effects, but this almost never happens!

B. Chance effects alone (all chance variation).

- Often occurs. Often boring because it suggests the effects we’re seeing are just random.

C. Systematic effects plus chance.

- Often occurs. Interesting because there’s at least some systematic factor.

An important part of statistics is determining whether we’ve got case B or C.

Tests of significance

- Invented to deal with this question of whether there’s a systematic effect, or just noise (chance).

Example (from your book)

- A senator introduces a bill to simplify the tax code. He claims this bill is revenue-neutral, i.e. on balance, tax revenues for the government will stay the same.
- To evaluate his claim, the Treasury Department will compare, for 100,000 representative tax returns, the amount of tax paid under the new bill, vs. under the old tax law.

  d = tax under new bill – tax under the old rules

  (this d is going to be our random variable)

Evaluating the new tax bill

- However, first, just to get a hint of how the results might turn out (with less work) they run a pilot study, in which they just look at the difference between the old and new rules, d, for 100 returns randomly chosen from the 100,000 representative returns. Results from this sample of 100 returns were as follows:
- m(d) = -\$219
- s(d) = \$725.
  - This is a pretty big standard deviation for a mean of -\$219.
  - How much tax people pay is highly variable, and it's not surprising that a new bill would have a big effect on some returns, and very little on others.

Initial impressions

- If, under the new law, the government really loses an average of $219 per tax return, that could add up to a lot of money!
  - $200/return x 100 million returns = $20 billion!
- But, this was just a pilot with 100 returns. And there's a very large standard deviation.
  - Do we expect this result of m(d)=-$219 to generalize, or is it different from $0 just by chance?

Does the tax law have an effect on revenue?

- From the results of a sample of 100 differences, d (i.e., we've looked at the effect of the tax law on 100 tax returns), what can we say about the underlying population?
- I.E. is the difference we observe in the sample (-$219) a real difference, or did it occur just due to chance (we happened to randomly pick 100 returns on which the new tax law will bring in less revenue, on average)?

Hypotheses

- The Treasury Dept. in this story believes that the tax law will negatively affect revenue. They believe the sample mean of -$219 represents a real difference.
- The senator believes that the tax law will make no difference, on average. He believes that the -$219 was due to chance.
- These two alternatives, (a real difference, vs. due to chance) are two opposing *hypotheses*.

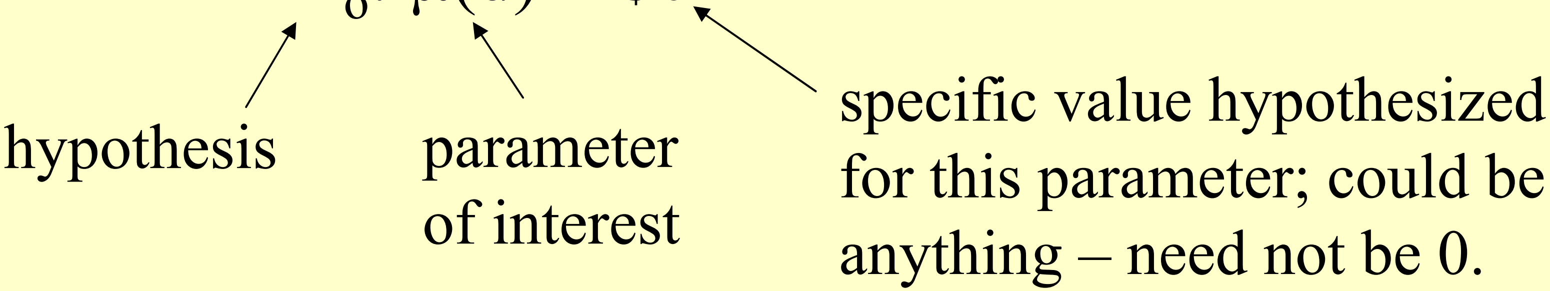
Hypotheses

- The terminology for these hypotheses in statistics is the “null hypothesis” and the “alternative hypothesis”.
- Null hypothesis, H_o:
  - There is no “real,” systematic effect. The observed effect was due to chance.
- Alternative hypothesis, H_a:
  - There is a real effect. In this case, the Treasury believes the effect is explicitly a negative one – a reduction in tax revenue. (They wouldn’t be so annoyed with the senator if they thought his law would *increase* tax revenues…)

Hypotheses

- Null hypothesis, H_o:
  - There is no “real,” systematic effect.

    $H_o: \mu(d) = \$0$



hypothesis

parameter of interest

specific value hypothesized for this parameter; could be anything – need not be 0.

- Alternative hypothesis, H_a:
  - There is a real effect, and it’s negative.

    $H_a: \mu(d) < \$0$

Hypotheses

- Note that the senator and the Treasury are arguing over what is true for the *population*. They agree over what happened for the 100 tax returns they looked at.
- Tests of significance only make sense when you want to know about the population. Put another way, you want to know whether to expect the result (observed in the sample) to generalize to other samples.

Comment on terminology

- The alternative hypothesis is often the interesting one – and often the one that someone sets out to prove.
  - E.G. The drug works – it has a real effect.
- The null hypothesis, in lay terminology, is the often more boring “alternative”.
  - The drug doesn’t work. Any effect you saw was due to chance.
- This terminology may seem a bit strange, but it’s standard. Think of “null” as “there’s no real effect,” and “alternative” as “other than null.”

Back to the tax law problem

- This problem is much like what you've seen in earlier examples.
- Recall:
  - 100 samples of d, the difference between the new and old tax laws.
  - m(d) = -\$219, SD(d) = \$725.
  - How likely is it that we would see a sample mean of -\$219 or less, if (H_o: $\mu(d)=\$0$) were true?

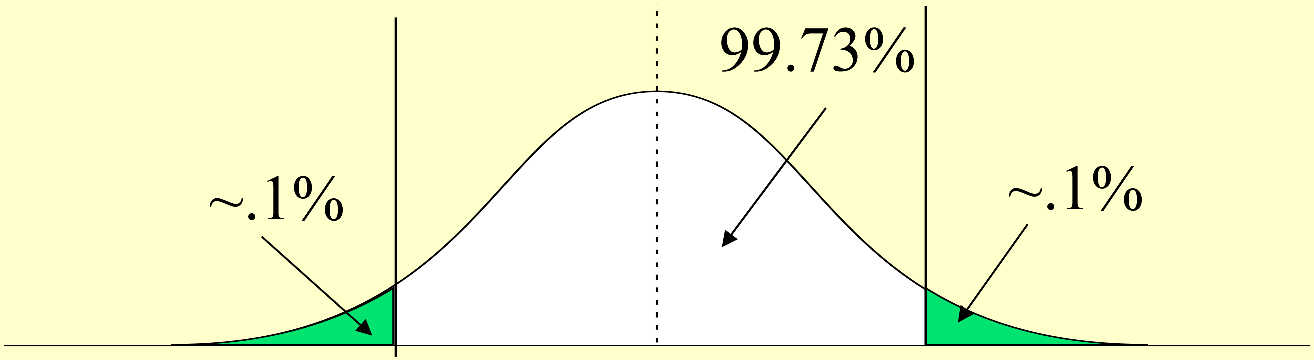
$P(m(d) \leq -\$219 \mid \mu(d) = \$0) = ?$

- As when we did sampling theory, we need to know the mean and standard error of the sampling distribution of the mean.
  - Assuming H_o is true gives us a theory about what the sampling distribution of the mean looks like.
  - H_o true -> $\mu(d) = \$0$.
  - H_o true doesn't tell us the standard error, σ/sqrt(N). However, we can approximate it by s/sqrt(N).

$P(m(d) \leq -\$219 \mid \mu(d) = \$0) = ?$

- N=100 is pretty large, so we can use the z-tables for this.
- $$\begin{aligned} z(-\$219) &= (-\$219 - \mu(d))/SE(d) \\ &= (-\$219 - \$0)(\$725/\mathrm{sqrt}(100)) \\ &\approx -3 \end{aligned}$$
- Looking up in our z-tables, what is the probability that $z \leq -3$?

$P(z \leq -3) = ?$



- From the tables:

z	Height	Area
3	0.443	99.73

- So, the probability that we would see a mean difference on 100 tax returns of -$219 if the population mean difference were $0 is about 1 in 1000.

What do we report in a research paper?

- Summarize the data, say what test is used, report the p-value.
  - "The difference in revenue between the new tax proposal and the old tax law (M=-$219, SD=$725) was highly significant (z = -3, p<0.01)"

summary

statistic

its observed value

p-value.
Alt: p=0.0014

Will the tax law negatively affect tax revenue?

- Well, we assumed that it would not (the null hypothesis, H_o).
- And, following that assumption to its logical conclusion, we found a “contradiction,” of sorts. An “absurdity”. An unlikely event.
  - If H_o is true, it is highly unlikely that we would have observed such a low value for m(d) = -$219.
- Therefore we “reject” the hypothesis, H_o, and decide to “accept” the alternative hypothesis, H_a, that there’s a real systematic effect. It’s likely that the tax law has a real negative impact on tax revenues.

Argument by contradiction

- You've probably seen proof by contradiction.
- E.G. Prove that there is no largest prime number.
  - **Assume the contradiction: there is a largest prime number, p_M.**
  - Let $N = p_1 \cdot p_2 \cdots p_M + 1$
  - $N>p_M$, so by our assumption, N is not prime.
  - But if it's not prime, it must be divisible by one of our M primes, p_1 through p_M (with remainder 0).
  - But, N/p_i always has remainder 1, by construction.
  - So N must be prime. **A contradiction.**
  - **Therefore, our original assumption must be wrong – there is no largest prime number – there are an infinite number!**

Argument by contradiction

- Statistical tests are also based on argument by contradiction.
  - It’s not quite a “proof,” because we’re never 100% sure of our decision.
  - To test whether or not the alternative hypothesis is true, assume it is not – assume the null hypothesis is true.
  - If you can show that this leads to a highly unlikely event, i.e. that you would observe the data you observed, then
  - You can reject the null hypothesis, and accept the alternative hypothesis, that there is a real systematic effect.

Why argument by contradiction?

- Why test for significance in this convoluted way?
- In many cases, it's difficult to set up the alternative hypothesis so it can be tested directly.
  - Often don't know what the alternative mean is, for instance.

Another example: Racial Bias in Jury Panels?

- Panels of jurors are theoretically drawn at random from a list of eligible citizens

- In the South in the ‘50s and ‘60s few African Americans were found on jury panels, so some defendants challenged their verdicts.

Racial bias on juries: a composite of several cases argued in the South between 1960 & 1980

- On appeal, one expert statistical witness gave this evidence
  - 1. 50% of eligible citizens in the county were African American
  - 2. On the 80-person panel of potential jurors, only ***four*** were African American

- Could this be the result of pure chance?

Chances aren't...

- If the selection of potential jurors was, in fact, random, then the number of African American jurors on a panel would be the binomial random variable X with n=80 trials and p=.5
- The chances of getting a panel with only 4 African Americans is $\Pr(X \leq 4) = 1.4 \times 10^{-18}$
- (or .0000000000000000014)

A Fair Deal?

- Since the probability is so small, this is strong evidence against the hypothesis of random selection
- To emphasize the point, the witness points out that this probability is less than the chances of getting three consecutive royal flushes in a poker game (3.6×10^{-18})
- The judge upholds the appeal

- Now, let’s go through this again, and talk more explicitly about what the steps are.

Step 1: Formulate Hypotheses

- H_0, the null hypothesis, is usually that the observations are the result of pure chance
  - – Each selection for the jury pool is 50% likely to be African American
  - – H_o: $p=0.5$

Step 1: Formulate Hypotheses

- H_a, the alternative hypothesis, is that the observations are the result of a real effect (plus some chance variation)
  - African Americans are under-represented, i.e. the probability of them being selected for the jury pool is lower than expected, given their representation in the population.
  - H_a: $p<0.5$

Step 2:
The Test Statistic

- Identify a statistic that will assess the evidence against the null hypothesis
  - In the court case, the test statistic is the binomial variable X with p=.50 and n=80

Step 3: Determine p-value

- A probability statement answering the question “if the null hypothesis were true, then what is the probability of observing a test statistic at least as extreme as the one we observed?”
  - $\Pr(X \leq 4 \mid p=.50 \text{ and } n=80) = 1.4 \times 10^{-18}$

Step 4:
Check significance

- Compare the p-value to a fixed significance level, α
- α acts as a cut-off point, below which we agree that an effect is statistically significant
- If $p \leq \alpha$ then we rule out H_0 and decide that something else is going on

When to Reject H_0

- In scientific work, we usually choose a fixed α of .05 or .01
  - $p<0.05$ -> "statistically significant"
  - $p<0.01$ -> "highly significant"
- This is arbitrary, to some extent varies from field to field, and is a holdover from the pre-computer days. But many scientific journals still only publish results when $p \leq 0.05$

- Different situations require different α levels.
  - What is the cost of being wrong?
  - Which do you want to avoid more: saying something is significant, when it's not, or saying it's not significant when it is? (More on this later.)
  - Does H_a seem really unlikely? In which case perhaps be conservative.

Step 5: Summarize the data, say what test is used, report the p-value

- If p<α:
  - "The difference between the proportion of African-American jurors selected (0.05), and the proportion predicted by their presence in the population of eligible citizens (0.5) was highly significant ($p=1.4x10^{-18}$, computed using the binomial formula)."
  - Old style: "… was highly significant (p<0.01…)"
    - Results were reported simply in terms of whether they were > or < the fixed value of α.
    - Nowadays, we can be more informative, report the value of p, and let people make their own judgments of how significant the results are.

Step 5: Report the results

- If $p>\alpha$:
  - “The significance of the difference between the observed proportion of African-American jurors (0.40) and the proportion expected by chance (0.50) was tested. This difference was not significant ($p = 0.22$, ns).”

Minding your p's and α's

- P-value: The probability of a result at least as extreme as the one we have obtained assuming H_0 is true. The smaller the p-value, the more surprising the result and the stronger the evidence against the null hypothesis
- Alpha (α): How much evidence against H_0 do we need in order to reject it? Lower α means we need more evidence.

Other statistical tests

- Many tests and choices of statistic:
  - One-sample z-test/z-statistic
  - t-tests/t-statistic
  - χ^2 test/ χ^2 statistic
  - F-tests/F-statistic
  - and so on.
- All tests follow the steps outlined above. And their p-values can be interpreted in the same way.

Meaning of the p-value

- p = probability of seeing a value equal to the observed value, or more extreme than the observed value, if the null hypothesis is true.
- Since the null hypothesis is typically the hypothesis that there is no real systematic effect, and any difference is due to chance alone, p = probability of seeing the observed value, or more extreme values, due to chance alone.
- Put another way: p = probability that another investigator running the same experiment would get a difference at least as big as our observed value, if the null hypothesis were true.

What p does *not* mean

- p is *not* equal to the probability that the null hypothesis is true, given the data!
- p is computed *assuming* the null hypothesis.
- $p = P(x \leq 4 \mid H_o) \neq P(H_o \mid x \leq 4)$

Furthermore, we can’t easily determine the probability that the null hypothesis is true

- Could we get the probability by running the test a number of times?
  - (Frequency theory)
- According to frequency theory, there is no way to define the probability of the null hypothesis being right. The distribution is what it is – if you run the experiment many times, the null hypothesis is always right, or always false. You can’t just run the test lots of times and find the probability that it’s right.

The null hypothesis

- So, don’t talk about the probability that the null hypothesis is true – we don’t know this probability, and p does *not* equal this probability!

The null hypothesis

- In general, we can *reject* or *discredit* the null hypothesis with a fair degree of confidence, if our p-value is sufficiently low.
- But we can’t really *prove* the null hypothesis.
- If we do not reject the null hypothesis, we may say we *accept*, or *retain the null hypothesis*, or *treat the null hypothesis as viable*.

The null hypothesis

- Furthermore, common sense tells us that the null hypothesis is virtually never literally true to the last decimal place.
  - H_o: $\mu = 0.00000000\ldots$
  - Most sensible experimental manipulations ("does this drug have an effect?") cause at least *some* difference.
  - Retaining a null hypothesis of no mean difference is like saying that we're insufficiently confident whether the mean difference is > 0 or < 0.

Significance and multiple tests

- The point of testing is to distinguish between real differences and chance variation.
- Does statistical significance mean that the result cannot be explained by chance variation?
  - No. Once in a while, an event that is unlikely to occur due to chance can actually occur.
  - We talked about this with confidence intervals – roughly 1 in 20 times, the true mean fell outside of the 95% confidence interval.

Significance and multiple tests

- Put another way, a researcher who runs 100 tests can expect to get 5 results which are “statistically significant” ($p<0.05$), and one which is “highly significant” ($p<0.01$), even if the null hypothesis is correct in every case.
- You cannot tell, for sure, whether a difference is real or just coincidence.
  - This is why science requires replicable results. If n independent tests all show a statistically significant result, the probability of this happening due to chance is very small.

Multiple tests, looked another way

- Suppose we run a family of k experiments, specifying $\alpha=0.05$ for each experiment. What is the probability of at least one error (incorrectly rejecting the null hypothesis) in the family of studies?
- Bonferroni inequality:
  - p(one or more errors) $\leq k\alpha$
  - Holds regardless of whether the results of the k experiments are independent
- If we want to ensure that p(one or more errors in the *family* of experiments) < 0.05, we should use a criterion of $\alpha=0.05/k$ for each experiment.

Multiple tests

- Be wary of studies that run lots of tests, and use a liberal criterion like $\alpha=0.05$.
- "The Nurture Assumption" by Judith Rich Harris.
  - Studies on the effect of birth order will look for effects of birth order on sociability, extraversion, aggressiveness, excitability, nervousness, neuroticism, depression, inhibition, calmness, masculinity, dominance, and openness.
  - They found effects of birth order for families of 3 or more, where the lastborn was slightly less masculine.
  - But that's running something like 36 tests for differences – no wonder they found one significant result! It's likely to occur just by chance.

Another situation with this issue of correcting for multiple tests

- fMRI
- Each highlighted region is made up of voxels.
- Researchers must determine whether, for each voxel, there's a significant difference from the baseline response levels.
- I'm not sure how to do this. There are still new research papers on it every year – people are still working it out.

fMRI image removed due to copyright reasons