Lecture 21-22: Inference in Practice

Tomer Altman and Alan Hubbard

October 18, 2024

Learning objectives

- Learn how to make a confidence interval more narrow
- Define p-hacking and know some issues with relying on p-values when performing research
- Calculate what sample size is required for a given margin of error or confidence interval width
- Define statistical power and know how to calculate it
- Define type I and type II error

Readings

- Chapter 15 of Baldi and Moore
- Online resources:
  - Type I and Type II error; section 11.2, page 330); page 320,
  - Basics of power, including its relation to Type I/Type II error, and factors that affect power, section 10.3-10.3.1

Recall the conditions for inference relying on normal sampling distribution

1. SRS
2. Population has a Normal distribution (or n enough)
3. σ is known (or n large that one can use sample SD)

Condition 1: Where do the data come from?

Remember that statistical inference involves the concept of repeated experiments (e.g, SRS or randomized experiment).

Sometimes one still reports inference for non-random samples, treating them as if they were generated from a SRS, given reporting some measure of uncertainty is better not, though the inference might not reflect the true sampling distribution.

Condition 1: Where do the data come from?

- Example 1: A neurobiologist is interested in how visual perception can be fooled by optical illusions.They use their students as a convenience (i.e., non-random) sample.

- Example 2: A sociologist is interested in attitudes toward the use of human subjects in science. They use their students as a convenience (i.e., non-random) sample.

- For one of these examples, it matters less that the sample is non random. Which do you think it might be and why?

Condition 1: Where do the data come from?

Even if the data come from a randomized experiment, there could be issues that hinder the randomness:

- Non response: Some surveyed individuals may not complete a survey. What happens if they are different than the individuals who complete the survey?
- Dropout ("Lost to follow-up", in RCT jargon): In a randomized trial, you may follow enrolled individuals over time, but some people may stop participating in the study and drop out. When might these individuals be different from those individuals who stay in study until the end of the study period?

Condition 2: What is the shape of the population distribution?

- There is flexibility with this condition. We started by assuming Normality of the distribution, which **guarantees** that the sampling distribution for the mean is Normally distributed, no matter the sample size.
- However, because of the Central Limit Theorem, the sampling distribution for the mean – no matter the shape of the underlying distribution – will eventually become Normally distributed when the sample size n is large enough.
- The z-test is reasonably accurate for any symmetric distribution if the sample size is moderate
- If the distribution is skewed, then you need a large enough n for the z-test to work.
- Thus, we are loosening the condition from the past couple of lectures: consider both the shape of the distribution (is it symmetric?) and the sample size to determine if you can perform a z-test

Condition 2: What is the shape of the population distribution?

- Examine the shape of the sample's distribution using a qq plot (ideally) or a histogram. Use it to infer the shape of the population distribution
- Difficult to infer much if there are too few observations.

Condition 2: What is the shape of the population distribution?

- Outliers can affect tests of non-resistant measures like the mean.
- Double check if the outlier is "real" or an error. If it is real, can use other methods that aren't sensitive to outliers.

How confidence intervals behave

Recall the form of a CI:

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

Where $z^* \frac{\sigma}{\sqrt{n}}$ is the **margin of error**.

The margin of error gets smaller when:

- **z* is smaller** (i.e., you change to a lower confidence level). To obtain a narrower confidence interval using the same data (i.e., same σ and n) you need to accept a lower confidence level, implying a trade-off between the confidence level and the margin of error.
- σ **is smaller**. You might be able to reduce σ if there is measurement error of the variable that you can improve or eliminate. Often times, the σ can't be reduced, it is just a characteristic of the population.

- n **is larger**. You need to quadruple the sample size to halve the margin of error:
  - Recall that $SE = \frac{\sigma}{\sqrt{n}}$. If you increase n to $4 \times n$ then $SE = \frac{\sigma}{\sqrt{4 \times n}} = \frac{\sigma}{2 \times \sqrt{n}} = \frac{1}{2} \times \frac{\sigma}{\sqrt{n}}$, which is half of the original SE

The margin of error only accounts for sampling error

- This is one of the most important points!
- If you're taking epidemiology, you've likely learned about epidemiologic bias: confounding, measurement error, and selection bias. These are **systematic errors**. The confidence interval does not account for systematic errors.
- For example, if you were measuring birth weight in grams and the scale off by +50 grams for each baby, then estimate of the average would also be off by +50 grams and the confidence would be shifted to the right by +50 grams. Increasing sample size would not overcome this measurement error.
- There are methods for bias-adjusted CI's, but only in situations where the bias can be calculated (often it's unknowable).

How hypothesis tests behave

- The p-value of a test is dependent on whether H_a is one-sided or two-sided.
- The p-value for a two-sided test is double the p-value for the one-sided test of the same H_0
- It is important to determine whether you are conducting a one- or two-sided test before you see the data, and if one-sided, whether you are testing if $\mu < \mu_o$ or $\mu > \mu_0$.
- That is you should never use $\bar{x}$ as a basis for determining the direction of the alternative hypothesis. Why do you think that is?

How hypothesis tests behave

- Statistical significance depends on sample size (since sample size determines the standard error of the sampling mean)
- Recall the form of the z-test:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- We can call $\bar{x} - \mu_0$ the "magnitude of the observed effect" since it is our measure of how much the population average we're looking at differs from some null hypothesized mean.
- We can also call $\sigma/\sqrt{n}$ the "size of chance variation" because it quantifies how much "chance" variations (only due to randomness) we expect from sample to sample.
- Then:

$$\frac{\text{magnitude of observed effect}}{\text{size of chance variation}} = \frac{signal}{noise}$$

- The "signal to noise ratio" is something you might have heard mentioned in articles you read online and this is what that means.
- Statistical significance depends on:
  - The size of the observed effect ($\bar{x} - \mu$)
  - The variability of individuals in the population (σ)
  - The sample size (n)

How hypothesis tests behave

- Statistical significance depends on:
  - The size of the observed effect ($\bar{x} - \mu_0$)
  - The variability of individuals in the population (σ)
  - The sample size (n)
- If you obtain a small p-value it is not necessarily because the effect size is large.
- Very tiny effects can be deemed statistically significant when you have enough data. This is a big problem in the age of big data, because you're almost guaranteed to obtain statistically significant results, no matter the effect size.
- On the other hand, if your sample size is too small, you might not obtain statistical significance even if your effect size is large.
- This means: **An absence of evidence is not evidence of absence**. Said another way: Failing to reject the null hypothesis does not imply that the null hypothesis is true.

How hypothesis tests behave

- Statistical significance is different from clinical/practical significance.
- A statistically significant result might not be large enough to be important in context of the question.

P-values

- Watch this 11-min YouTube video on "P-hacking": https://www.youtube.com/watch?v=Gx0fAjNHb1M
- Read this Vox article about a Cornell food researcher how engaged in p-hacking: https://www.vox.com/science-and-health/2018/9/19/17879102/brian-wansink-cornell-food-brand-lab-retractions-jama
- Read this two-page ASA brief on statistical significance and p-values: https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf
- We don't have time to discuss these today, but the material contained in these sources is testable!

Selecting an appropriate sample size

How big does the sample size need to be?

Suppose you want your margin of error to equal m. What sample size do you need to obtain a margin of error of m?

You can rearrange the formula for the margin of error (moe) to solve for n:

$$moe = \frac{z^* \sigma}{\sqrt{n}}$$

$$n = \left(\frac{z^* \sigma}{moe}\right)^2$$

Example of calculation sample size

Body temperature has a known $\sigma = 0.6$ degrees F. We want to estimate the mean body temperature μ for healthy adults within ±0.05 F with 95% confidence. How many healthy adults must we measure?

$$n = \left(\frac{z^* \sigma}{moe}\right)^2$$

$$n = \left(\frac{1.96 \times 0.6}{0.05}\right)^2 = 553.2$$

- We must recruit 554 healthy adults for this study.
- Note that we always round up when calculating sample size, because if we rounded to the nearest whole digit (553) then the margin or error would be smaller than ±5.

Example of calculation sample size

- Be careful! Rather than giving you the required margin of error, the question may give you the required **width** of the confidence interval.
- E.g., the previous question could have said: We want to estimate a 95% confidence interval for the mean that has a width of 0.1 units. How many healthy adults must we measure?
- The width of the confidence interval is $2 \times moe$, so here $moe = 0.05$, just like in the previous question.

Sample size when conducting a hypothesis test

To think about sample size for a z-test, three things matter:

- **Significance level:** How much protection do we want against saying there is "evidence against H_0 in favor of H_A" when there really is no effect in the population? The significance level we choose fixes this conditional probability to a fixed level, often 5%: $P(\text{Reject } H_0 | H_0 \text{ is true}]) = 0.05$
- **Effect size:** How large of an effect in the population is important in practice?
- **Power:** How confident do we want that our study will detect an effect of the size we think is important? I.e., what is the probability of rejecting H_0 when the alternative hypothesis is true? That is, what is $P(\text{Reject } H_0 | H_A \text{ is true})$?

Example of calculating power

- Suppose you know that $\sigma = 1$.
- $H_0 : \mu = 0$
- $H_a : \mu > 0$
- Set $\alpha = 0.05$.
- Additional condition: You need to be 90% confident that the test will reject H_0 when the true underlying mean equals $\mu = 0.8$. That is, you want 90% **power** when $\mu = 0.8$.
- What sample size will give you 90% power?

Example of calculating power

Begin by assuming $n = 10$ and calculate the minimum z-value required to reject H_0:

```
qnorm(p = 0.05, mean = 0, sd = 1/sqrt(10), lower.tail = F)
```

```
## [1] 0.5201484
```

So for any z-test with this value or higher, you will reject H_0 in favor of H_a.

Now suppose that $\mu = 0.8$ is true. The test will reject H_0 what percent of the time when H_a is true? To calculate this probability, we take the value from the previous calculation and calculate the *probability* above its value under H_A:

```
pnorm(q = 0.5201484, mean = 0.8, 1/sqrt(10), lower.tail = F)
```

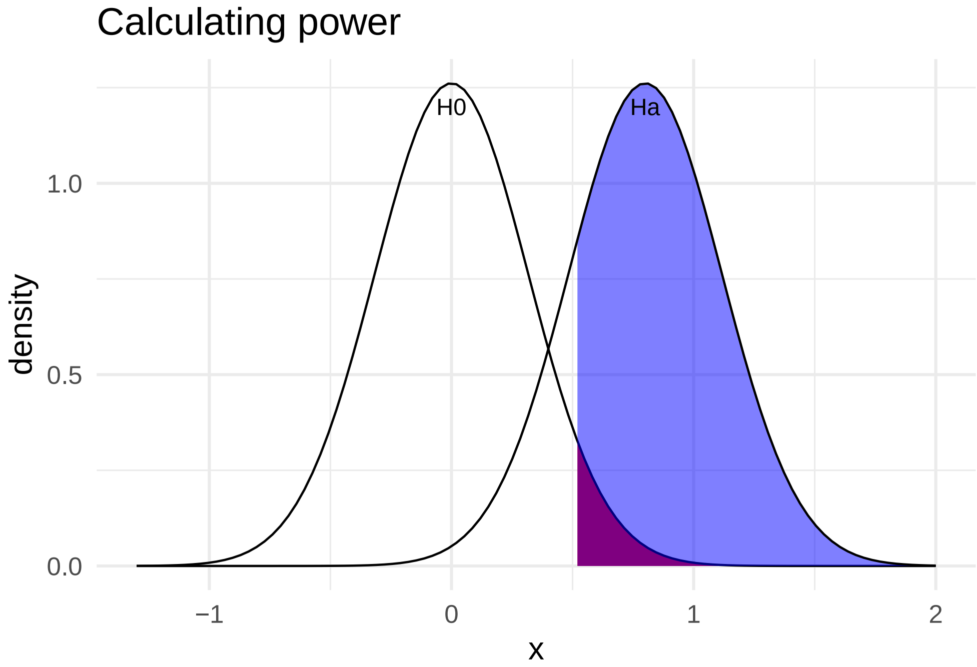
```
## [1] 0.8119132
```

Thus, you have a 82% chance of obtaining evidence in favor of H_A when $\mu = 0.8$ if $n = 10$. To obtain power of 90%, you will need to increase the sample size.

Example of calculating power, illustrated

```
## Warning in geom_text(aes(x = 0, y = 1.2), label = "H0", check_overlap = T): All aesthetics have leng
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```

```
## Warning in geom_text(aes(x = 0.8, y = 1.2), label = "Ha", check_overlap = T): All aesthetics have le
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```



- There is 5% of the area under H_0 that is shaded here in purple. This is the chance of rejecting the null hypothesis even when it is true. We found that x = 0.5201484 denotes the value where an $\bar{x} >= 0.5201484$ would provide evidence against H_0 at $\alpha = 0.05$
- Then, we sketch the distribution of the alternative if $\mu = 0.8$ and we calculate the probability of rejecting H_0 when $\mu = 0.8$ is true.

Power

Thus, for this example, if you want power of 90%, you need to increase $n > 10$.

Based on the calculations below, $n = 14$ reaches 90% power:

```r
n <- 13
z <- qnorm(p = 0.05, mean = 0, sd = 1/sqrt(n), lower.tail = F)
pnorm(q = z, mean = 0.8, sd = 1/sqrt(n), lower.tail = F)
```

```
## [1] 0.892436
```

```r
n <- 14
z <- qnorm(p = 0.05, mean = 0, sd = 1/sqrt(n), lower.tail = F)
pnorm(q = z, mean = 0.8, sd = 1/sqrt(n), lower.tail = F)
```

```
## [1] 0.9112467
```

What affects the sample size required for a hypothesis test?

- The difference between the null and true values being the same, smaller p-values occure with larger n.
- Higher power requires a larger n
- For a given α and power, two-sided alternatives require a larger n than one-sided tests
- Aiming to detect a smaller difference between $\bar{x}$ and μ requires a larger n

Power, Type I error, and Type II error in hypothesis tests

- The significance level α is the chance of making a wrong decision when the null hypothesis is true. This is known as a **type I error**
- The power is the chance of making the right decision when the alternative is true.
  - Thus, its complement, 1-power, is the chance of making a wrong decision when the alternative hypothesis is true. This is known as a **type II error.**
  - We let β denote the chance of a Type II error. Power $= 1 - \beta$

Power, Type I error, and Type II error in hypothesis tests

	H_0 is true	H_a is true
Reject H_0	Type I error $(prob = \alpha)$	Correct decision
Fail to reject H_0	Correct decision	Type II error $(prob = \beta)$