STAT 22000 Lecture Slides
The General Framework of Hypothesis Testing

Yibi Huang
Department of Statistics
University of Chicago

Textbook Coverage

This set of slide covers some of 4.3 in the text.

Can Dogs Smell Cancer?

https://www.youtube.com/watch?v=e0UK6kkS0_M

https://www.youtube.com/watch?v=jlrgtWWwqZo

Case Study: Can Dogs Smell Bladder Cancer?

- A study by M. Willis et al. (*British Medical Journal*, vol. 329, September 25, 2004.) considered whether dogs could be trained to detect if a person has bladder cancer by smelling certain compounds in the patient's urine.
- 6 dogs of varying breeds were trained to discriminate between urine from patients with bladder cancer and urine from control patients without it.
- The dogs were taught to indicate which among several specimens was from the bladder cancer patient by lying beside it.
- Once trained, the dogs's ability to distinguish cancer patients from controls was tested using urine samples from subjects not previously encountered by the dogs.

Case Study: Can Dogs Smell Bladder Cancer?

- The researchers blinded both dog handlers and experimental observers to the identity of urine samples.
- Each of the 6 dogs was tested with 9 trials. In each trial, one urine sample from a bladder cancer patient was randomly placed among 6 control urine samples.
- Outcome: In the total of 54 trials with the 6 dogs, the dogs made the correct selection 22 times.
    - The dogs were correct for $22/54 \approx 40.7\%$ of the time.
    - If the dogs just guessed at random, they were expected to be correct for $1/7 \approx 14.3\%$ of the time
    - Is this difference (40.7% v.s. 14.3%) surprising?

Two Competing Hypotheses

Let p be the proportion of time that dogs were correct.

1. *Null hypothesis (H_0)*: $p = 1/7$

   "There is nothing going on."

   The dogs just guessed at random.

   - "null" means "nothing surprising is going on".
   - The dogs were just lucky to make more correct selections than expected.

2. *Alternative hypothesis (H_A)*: $p > 1/7$

   "There is something going on."

   Dogs can do better than random guessing.

Weighing Evidence

The next step of hypothesis testing is to weigh the evidence — could these data plausibly have happened by chance if the H_0 was true?

- If the observed result was very unlikely to have occurred under the H_0, then the evidence raises more than a reasonable doubt in our minds about the H_0.

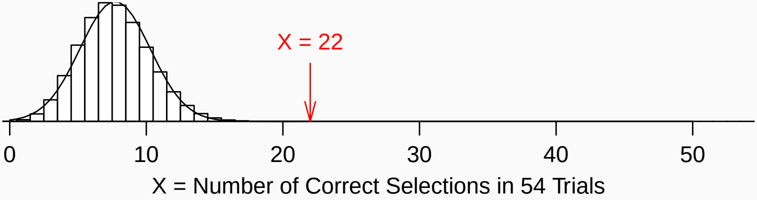
Test Statistic

The *test statistic* is a summary of the data that best reflect the evidence for for or against the hypotheses.

- For this study, we use $X =$ “the total number of correct selection in the 54 trials” as our test statistic.
- The larger X is, the stronger the evidence for H_A and against H_0
- The smaller X is, the stronger the evidence for H_0 and against H_A

If H_0 is true, then $X \sim Bin(n = 54,\ p = 1/7)$ (Why?)

$$P(X = k) = \binom{54}{k}(1/7)^k(6/7)^{54-k}, \quad k = 0, 1, 2, \ldots, 54.$$



Under H_0

$$P(X \geq 22) = \sum_{k=22}^{54} \binom{54}{k} (1/7)^k (6/7)^{54-k} \approx 1.86 \times 10^{-6}$$

```
> sum(dbinom(22:54,54,1/7))
[1] 1.861522e-06
```

If the dogs just guessed at random, they could be correct in 22 or more of the 54 trials no more than 2 out of 1 million of the time

The observed result was very unlikely to have occurred under the H_0 — strong evidence to disbelieve H_0.

P-values

The probability $P(X \geq 22) \approx 1.86 \times 10^{-6}$ is called the *p-value* of the test. What is a *p-value*?

The *p-value* of test is the **probability of observing data such that the evidence for the H_A is at least as strong as our current data set, assuming the H_0 is true**.

- The *p*-value for the dog study is $P(X \geq 22)$ not $P(X = 22)$.

The smaller the *p*-value, the stronger the evidence against the H_0:

- A *p*-value of 0.25 says that if the H_0 was true, then we would obtain a sample that looks like the observed sample 1 in 4 of the time; ⇒ the data seems to be consistent with H_0
- A *p*-value of 0.001 says that if the H_0 was true, then only 1 out of every 1,000 samples would resemble the observed sample; ⇒ the H_0 looks doubtful

Significance Level

- As remarked earlier, the smaller the *p*-value is, the stronger the evidence against the null.
- In some studies, we can simply report the *p*-value and let people judge whether the evidence is strong enough
- In other studies, we need to make a decision about which hypothesis to trust
  - We then select a cut-off value α, call the *significance level*
  - If the *P*-value $< \alpha$, we reject H_0
  - If the *P*-value $> \alpha$, we don't reject H_0
- Commonly used significance levels: 0.05 and 0.01
  - A test with *P*-value $<$ 0.05 is said to be *(statistically) significant*
  - A test with *P*-value $<$ 0.01 is said to be *highly significant*

Type 1 and Type 2 Errors

In a hypothesis test, we make a decision about which of H_0 or H_A might be true, but our decision might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the H_0 when it is true.
- A *Type 2 Error* is failing to reject the H_0 when it is false.
- We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

Consequences of Type 1 and 2 errors

Type 1 and type 2 errors are different sorts of mistakes and have different consequences

- Usually H_0 is the status quo, a thing we generally believe to be true
- If H_0 is not rejected, usually it means the status quo is fine. No action needs to be taken
- Rejecting H_0 means something we use to believe is overturned. It might be a scientific breakthrough (e.g., discovery of a new drug).
- A type 1 error introduces a false conclusion into the scientific community and can lead to a tremendous waste of resources before further research invalidates the original finding

Consequences of Type 1 and 2 errors

- A type 2 error — failing to recognize a scientific breakthrough — represents a missed opportunity for scientific progress
- Type 2 errors can be costly as well, but generally go unnoticed
- So it's more important to control the Type 1 error rate than the Type 2 error rate.

Significance Level = Type 1 Error Rate

- When the H_0 is true, there is only 5% chance to obtain a p-value $< 5\%$
- This means that, for those cases where H_0 is actually true, we won't incorrectly reject it more than 5% of those times in the long run
- In other words, when using a 5% significance level, there is about 5% chance of making a Type 1 error if the H_0 is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

- This is why we prefer small values of α — increasing α increases the Type 1 error rate.
- However, significance level doesn't control Type 2 error rate

True or False

If H_0 is rejected, then we can be certain that H_0 is false.

- __False__. Even if H_0 is true, 5% of the time the experiment will give a result with a *p*-value $<$ 5% so that H_0 is rejected.

If H_0 is rejected at 5% level, there is less than a 5% chance for H_0 to be true.

- __False__. A *P*-value does not give the chance of H_0 being true. In fact, the *P*-value is computed assuming H_0 is true.

Reporting the *P*-Value

Don't simply report the conclusion of whether H_0 is rejected. Attach the *p*-value.

- A *p*-value of 0.04 and a *p*-value of 0.000001 are not at all the same thing, even though H_0 will be rejected in both cases, but the strength of evidence are very different
- Simply reporting whether H_0 is rejected without *p*-value is like reporting the temperature as "cold" or "hot"
- It's much better to report the *p*-value and let people choose their own significance level, just like telling someone the temperature and let them decide for themselves whether they want to wear a coat

Conclusion of the Dogs Smell Bladder Cancer Study

- There is strong evidence that dogs have some ability to smell bladder cancer,
- However, the dogs were only correct 40% of the time, too low for practical application
- Another study (M. McCulloch et al., Integrative Cancer Therapies, vol 5, p. 30, 2006.) considered whether dogs could be trained to detect whether a person has lung cancer by smelling the subjects' breath. In one test with 83 Stage I lung cancer samples, the dogs correctly identified the cancer sample 81 times.

Remarks on the Design of the Dogs Smell Cancer Study

Q1. From the Youtube video, we see the urine samples were placed circularly on a carousal. Why didn't the investigators line up the sample in a row?

Q2. Why didn't the investigators let the dogs smell the subjects directly rather than the urine sample?

Q3. Why were the dogs tested using urine samples from a new set of subjects, not those used for training the dogs?

Recap: Hypothesis Testing Framework

- We start with a *null hypothesis (H_0)* that represents the status quo.
- We also have an *alternative hypothesis (H_A)* that represents our research question, i.e. what we're testing for.
- We then collect data and often summarize the data as a *test statistic*, which is usually a measure gauging whether H_0 or H_A are more plausible
- We then predict what the *test statistic* would be around under the assumption that the H_0 is true.
- If the *test statistic* is too far away from what the H_0 predicts, we then reject the H_0 in favor of the H_A.
  - We often computed a *p-value* based on the test statistic, which is the probability to obtain a test statistic at least as extreme as the one actually observed, assuming the H_0 is true
  - If the *p*-value is too small, we then reject the H_0 in favor of the H_A.

The Scientific Method: Proof and Disproof

- There is a subtle but very fundamental truth to the scientific method, which is that one can never really prove a hypothesis with it — only *disprove* hypotheses
- In the words of Albert Einstein,
  “*No amount of experimentation can ever prove me right;*
  a single experiment can prove me wrong”
- Therefore, we never declare the null hypothesis to be true
- When the evidence is not strong enough to reject the null, we don’t say “we accept the null hypothesis”, but say “we fail to reject the null hypothesis.”

This lecture introduces the general framework of hypotheses testing.

In the second half of STAT 220, we will introduce several hypotheses tests dealing with different types of problems.

In the next lecture, we will talk about hypotheses test about the population mean.