Economics 103 – Statistics for Economists

Minsu Chang

University of Pennsylvania

Lecture 19

Hypothesis Testing I

An excerpt from *The Lady Tasting Tea* by David Salsburg

It was a summer afternoon in Cambridge, England, in the late 1920s. A group of university dons, their wives, and some guests were sitting around an outdoor table for afternoon tea. One of the women was insisting that tea tasted different depending upon whether the tea was poured into the milk or whether the milk was poured into the tea. The scientific minds among the men scoffed at this as sheer nonsense. What could be the difference? They could not conceive of any difference in the chemistry of the mixtures that could exist. A thin, short man, with thick glasses and a Vandyke beard beginning to turn gray, pounced on the problem. "Let us test the proposition" he said excitedly. He began to outline an experiment in which the lady who insisted there was a difference would be presented with a sequence of cups of tea, in some of which the milk had been poured into the tea and in others of which the tea had been poured into the milk.

Continued...

And so it was that summer afternoon in Cambridge. The man with the Vandyke beard was Ronald Aylmer Fisher, who was in his late thirties at the time. He would later be knighted Sir Ronald Fisher. In 1935, he wrote a book entitled The Design of Experiments, and he described the experiment of the lady tasting tea in the second chapter of that book. In his book, Fisher discusses the lady and her belief as a hypothetical problem. He considers the various ways in which an experiment might be designed to determine if she could tell the difference.

The Pepsi Challenge

(1 “Expert,” 1 Skeptic)

The Pepsi Challenge

Our expert claims to be able to tell the difference between Coke and Pepsi. Let's put this to the test!

- Eight cups of soda
  - Four contain Coke
  - Four contain Pepsi
- The cups are randomly arranged
- How can we use this experiment to tell if our expert can *really* tell the difference?

The Results:

of Cokes Correctly Identified:

What do you think? Can our expert really tell the difference?

(a) Yes

(b) No

If you just guess randomly, what is the probability of identifying *all four cups of Coke correctly*?

If you just guess randomly, what is the probability of identifying *all four cups of Coke correctly*?

- $\binom{8}{4} = 70$ ways to choose four of the eight cups.

If you just guess randomly, what is the probability of identifying *all four cups of Coke correctly*?

- $\binom{8}{4} = 70$ ways to choose four of the eight cups.
- If guessing randomly, each of these is *equally likely*

If you just guess randomly, what is the probability of identifying *all four cups of Coke correctly*?

- $\binom{8}{4} = 70$ ways to choose four of the eight cups.
- If guessing randomly, each of these is *equally likely*
- Only *one* of the 70 possibilities corresponds to correctly identifying all four cups of Coke.

If you just guess randomly, what is the probability of identifying *all four cups of Coke correctly*?

- $\binom{8}{4} = 70$ ways to choose four of the eight cups.
- If guessing randomly, each of these is *equally likely*
- Only *one* of the 70 possibilities corresponds to correctly identifying all four cups of Coke.
- Thus, the probability is $1/70 \approx 0.014$

If you just guess randomly, what is the probability of identifying *all but one cup of Coke correctly*?

If you just guess randomly, what is the probability of identifying *all but one cup of Coke correctly*?

- $\binom{8}{4} = 70$ ways to choose four of the eight cups.

If you just guess randomly, what is the probability of identifying *all but one cup of Coke correctly*?

- $\binom{8}{4} = 70$ ways to choose four of the eight cups.
- If guessing randomly, each of these is *equally likely*

If you just guess randomly, what is the probability of identifying *all but one cup of Coke correctly*?

- $\binom{8}{4} = 70$ ways to choose four of the eight cups.
- If guessing randomly, each of these is *equally likely*
- There are 16 ways to mis-identify one Coke.
- Thus, the probability is $16/70 \approx 0.23$

Probabilities if Guessing Randomly

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If you're just guessing, what is the probability of identifying *at least* three Cokes correctly?

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If you're just guessing, what is the probability of identifying *at least* three Cokes correctly?

- Probabilities of mutually exclusive events sum.
- $P(\text{all four correct}) = 1/70$
- $P(\text{exactly 3 correct }) = 16/70$
- $P(\text{at least three correct}) = 17/70 \approx 0.24$

The Pepsi Challenge

- Even if you're just guessing randomly, the probability of correctly identifying three or more Cokes is around 24%
- In contrast, the probability of identifying *all four* Cokes correctly is only around 1.4% if you're guessing randomly.
- We should probably require the expert to get them all right.
- What if the expert gets them all wrong? This also has probability 1.4% if you're guessing randomly...

That was a Hypothesis Test!

We'll go through the details in a moment, but first an analogy...

Hypothesis Testing is Similar to a Criminal Trial

Criminal Trial

- The person on trial is either innocent or guilty (but not both!)

Hypothesis Testing

- Either the null hypothesis H_0 or the alternative H_1 hypothesis is true.

Criminal Trial

- The person on trial is either innocent or guilty (but not both!)
- "Innocent Until Proven Guilty"

Hypothesis Testing

- Either the null hypothesis H_0 or the alternative H_1 hypothesis is true.
- Assume H_0 to start

Criminal Trial

- The person on trial is either innocent or guilty (but not both!)
- "Innocent Until Proven Guilty"
- Only convict if evidence is "beyond a shadow of a doubt"

Hypothesis Testing

- Either the null hypothesis H_0 or the alternative H_1 hypothesis is true.
- Assume H_0 to start
- Only reject H_0 in favor of H_1 if there is strong evidence.

Criminal Trial

- The person on trial is either innocent or guilty (but not both!)
- "Innocent Until Proven Guilty"
- Only convict if evidence is "beyond a shadow of a doubt"
- *Not Guilty* rather than Innocent

Hypothesis Testing

- Either the null hypothesis H_0 or the alternative H_1 hypothesis is true.
- Assume H_0 to start
- Only reject H_0 in favor of H_1 if there is strong evidence.
- *Fail to reject* rather than Accept H_0

Criminal Trial

- The person on trial is either innocent or guilty (but not both!)
- "Innocent Until Proven Guilty"
- Only convict if evidence is "beyond a shadow of a doubt"
- *Not Guilty* rather than Innocent
  - Acquit $\neq$ Innocent

Hypothesis Testing

- Either the null hypothesis H_0 or the alternative H_1 hypothesis is true.
- Assume H_0 to start
- Only reject H_0 in favor of H_1 if there is strong evidence.
- *Fail to reject* rather than Accept H_0
  - (Fail to reject H_0) $\neq$ (H_0 True)

Criminal Trial

- The person on trial is either innocent or guilty (but not both!)
- "Innocent Until Proven Guilty"
- Only convict if evidence is "beyond a shadow of a doubt"
- *Not Guilty* rather than Innocent
  - Acquit $\neq$ Innocent
- Two Kinds of Errors:
  - Convict the innocent
  - Acquit the guilty

Hypothesis Testing

- Either the null hypothesis H_0 or the alternative H_1 hypothesis is true.
- Assume H_0 to start
- Only reject H_0 in favor of H_1 if there is strong evidence.
- *Fail to reject* rather than Accept H_0
  - (Fail to reject H_0) $\neq$ (H_0 True)
- Two Kinds of Errors:
  - Reject true H_0 (Type I)
  - Don't reject false H_0 (Type II)

Criminal Trial

- The person on trial is either innocent or guilty (but not both!)
- "Innocent Until Proven Guilty"
- Only convict if evidence is "beyond a shadow of a doubt"
- *Not Guilty* rather than Innocent
  - Acquit $\neq$ Innocent
- Two Kinds of Errors:
  - Convict the innocent
  - Acquit the guilty
- Convicting the innocent is a worse error. Want this to be rare even if it means acquitting the guilty.

Hypothesis Testing

- Either the null hypothesis H_0 or the alternative H_1 hypothesis is true.
- Assume H_0 to start
- Only reject H_0 in favor of H_1 if there is strong evidence.
- *Fail to reject* rather than Accept H_0
  - (Fail to reject H_0) $\neq$ (H_0 True)
- Two Kinds of Errors:
  - Reject true H_0 (Type I)
  - Don't reject false H_0 (Type II)
- Type I errors (reject true H_0) are worse: make them rare even if that means more Type II errors.

How is the Pepsi Challenge a Hypothesis Test?

Null Hypothesis H_0

Can't tell the difference between Coke and Pepsi: just guessing.

How is the Pepsi Challenge a Hypothesis Test?

Null Hypothesis H_0

Can't tell the difference between Coke and Pepsi: just guessing.

Alternative Hypothesis H_1

Able to distinguish Coke from Pepsi.

How is the Pepsi Challenge a Hypothesis Test?

Null Hypothesis H_0

Can't tell the difference between Coke and Pepsi: just guessing.

Alternative Hypothesis H_1

Able to distinguish Coke from Pepsi.

Type I Error – Reject H_0 even though it's true

Decide expert can tell the difference when she's really just guessing.

How is the Pepsi Challenge a Hypothesis Test?

Null Hypothesis H_0

Can't tell the difference between Coke and Pepsi: just guessing.

Alternative Hypothesis H_1

Able to distinguish Coke from Pepsi.

Type I Error – Reject H_0 even though it's true

Decide expert can tell the difference when she's really just guessing.

Type II Error – Fail to reject H_0 even though it's false

Decide expert just guessing when she really can tell the difference.

How do we find evidence to reject H_0?

- Choose a significance level α maximum probability of Type I error that we are willing to tolerate.
  - Measures how often we will reject a true null, i.e. convict an innocent person

How do we find evidence to reject H_0?

- Choose a significance level α maximum probability of Type I error that we are willing to tolerate.
  - Measures how often we will reject a true null, i.e. convict an innocent person
- Test Statistic T_n uses sample to measure plausibility of H_0

How do we find evidence to reject H_0?

- Choose a significance level α maximum probability of Type I error that we are willing to tolerate.
  - Measures how often we will reject a true null, i.e. convict an innocent person
- Test Statistic T_n uses sample to measure plausibility of H_0
- Null Hypothesis $H_0 \Rightarrow$ Sampling Distribution for T_n
  - "Under the null" means "assuming the H_0 is true"

How do we find evidence to reject H_0?

- Choose a significance level α maximum probability of Type I error that we are willing to tolerate.
  - Measures how often we will reject a true null, i.e. convict an innocent person
- Test Statistic T_n uses sample to measure plausibility of H_0
- Null Hypothesis $H_0 \Rightarrow$ Sampling Distribution for T_n
  - "Under the null" means "assuming the H_0 is true"
- Using α and the sampling distribution of T_n under the null, we construct a decision rule in terms of a critical value c_α

How do we find evidence to reject H_0?

- Choose a significance level α maximum probability of Type I error that we are willing to tolerate.
  - Measures how often we will reject a true null, i.e. convict an innocent person
- Test Statistic T_n uses sample to measure plausibility of H_0
- Null Hypothesis $H_0 \Rightarrow$ Sampling Distribution for T_n
  - "Under the null" means "assuming the H_0 is true"
- Using α and the sampling distribution of T_n under the null, we construct a decision rule in terms of a critical value c_α
  - Reject H_0 if $T_n > c_\alpha$

Example: Pepsi Challenge

Test Statistic T_n

$T_n =$ Number of Cokes correctly identified

H_0: No skill, just guessing randomly

Under this null hypothesis, the sampling distribution of T_n is:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

Example: Pepsi Challenge

T_n: # of Cokes correctly identified. Sampling Dist. under H_0:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If I choose a significance level of $\alpha = 0.05$, what critical value should I use?

(Remember that α is the probability of rejecting H_0 when it is actually true.)

Example: Pepsi Challenge

T_n: # of Cokes correctly identified. Sampling Dist. under H_0:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If I choose a significance level of $\alpha = 0.05$, what critical value should I use?

(Remember that α is the probability of rejecting H_0 when it is actually true.)

Want $P(\text{Reject } H_0 | H_0 \text{ True}) \leq 0.05$

$P(T_n \geq 3 | \text{Just Guessing}) = 17/70 \approx 0.23 > 0.05$

$P(T_n \geq 4 | \text{Just Guessing}) = 1/70 \approx 0.014 \leq 0.05$

Example: Pepsi Challenge

T_n: # of Cokes correctly identified. Sampling Dist. under H_0:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If I choose a significance level of $\alpha = 0.25$, what critical value should I use?

Example: Pepsi Challenge

T_n: # of Cokes correctly identified. Sampling Dist. under H_0:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If I choose a significance level of $\alpha = 0.25$, what critical value should I use?

Want $P(\text{Reject } H_0 | H_0 \text{ True}) \leq 0.25$

$P(T_n \geq 2|\text{Just Guessing}) = 53/70 \approx 0.76 > 0.25$

$P(T_n \geq 3|\text{Just Guessing}) = 17/70 \approx 0.23 \leq 0.25$

Example: Pepsi Challenge

H_0: Expert is just guessing randomly.

H_1: Expert can distinguish Coke from Pepsi.

T_n: # of Cokes correctly identified. Has following sampling under the null:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If I choose $\alpha = 0.05$, what decision rule should I use?

(a) Reject H_0 if $T_n \geq 0$

(b) Reject H_0 if $T_n \geq 1$

(c) Reject H_0 if $T_n \geq 2$

(d) Reject H_0 if $T_n \geq 3$

(e) Reject H_0 if $T_n \geq 4$

Example: Pepsi Challenge

H_0: Expert is just guessing randomly.

H_1: Expert can distinguish Coke from Pepsi.

T_n: # of Cokes correctly identified. Has following sampling under the null:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If I choose $\alpha = 0.05$, what decision rule should I use?

Need $P(\text{Reject } H_0 | H_0 \text{ True}) \leq \alpha = 0.05$

$$\begin{aligned} P(T_n \geq 3 | \text{Just Guessing}) &= 17/70 \approx 0.23 > 0.05 \\ P(T_n \geq 4 | \text{Just Guessing}) &= 1/70 \approx 0.014 \leq 0.05 \end{aligned}$$

Critical value for $\alpha = 0.05$ is 4

Example: Pepsi Challenge

H_0: Expert is just guessing randomly.

H_1: Expert can distinguish Coke from Pepsi.

T_n: # of Cokes correctly identified. Has following sampling under the null:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If I choose $\alpha = 0.25$, what critical value should I use?

(a) 0

(b) 1

(c) 2

(d) 3

(e) 4

So far

Simple Example of Hypothesis Testing: the Pepsi Challenge

What's next?

Hypothesis Testing More Generally

Hypothesis: Assertion about Population(s)

- A Big Mac contains, on average, 550 kcal: $\mu = 550$
- Midterm 2 was harder than Midterm 1: $\mu_1 > \mu_2$
- Equal proportions of Republicans and Democrats know that John Roberts is the chief justice of SCOTUS: $p = q$
- Google stock is riskier than IBM stock: $\sigma_X^2 > \sigma_Y^2$
- There is no correlation between height and income: $\rho = 0$

Hypothesis Testing: Try to Find Evidence *Against* H_0

Null Hypothesis: H_0

- Start off assuming H_0 is true – "innocent until proven guilty"
- "Under the Null" = Assuming the null is true
- $H_0 \Rightarrow$ know something about population, can calculate probs.

This Course: *Simple* Null Hypotheses

H_0: f(Parameters) = Known Constant, for example

- $\mu_1 - \mu_2 = 0$
- $p = 0.5$
- $\mu = 0$
- $\sigma_X^2/\sigma_Y^2 = 1$

How do I know what my null hypothesis is?

There is no rule I can give you for this: it depends on the problem.

Here are some guidelines:

- It will take the form $f(\text{Parameters}) = \text{Known Constant}$
- Nulls are typically things like "there is no effect," "these two groups are not different," i.e. the *status quo*.
- Nulls are *very specific*: we need to be able to do probability calculations under the null – c.f. the Pepsi Challenge.

Example: How many calories in a Big Mac?

- According to McDonald's: 550 kcal on average
- Measure calories in random sample of 9 Big Macs: $X_1, \ldots, X_9 \sim \text{iid } N(\mu, \sigma^2)$

If we wanted to test McDonald's claim, what would be H_0?

(a) $\sigma^2 = 1$

(b) $\mu = 0$

(c) $\mu > 550$

(d) $\mu = 550$

(e) $\mu \neq 550$

Example: How many calories in a Big Mac?

- According to McDonald's: 550 kcal on average
- Measure calories in random sample of 9 Big Macs: $X_1, \ldots, X_9 \sim$ iid $N(\mu, \sigma^2)$

If McDonald's is telling the truth, approximately what value should we get for the sample mean caloric content of the 9 Big Macs?

Example: How many calories in a Big Mac?

- According to McDonald's: 550 kcal on average
- Measure calories in random sample of 9 Big Macs:
  $X_1, \ldots, X_9 \sim$ iid $N(\mu, \sigma^2)$

If the sample mean does not equal 550, does this prove that McDonald's is lying?

(a) Yes

(b) No

(c) Not Sure

How to find evidence against H_0? Test Statistic!

Test Statistic: T_n

A statistic that gives us information about the parameter we are testing and has a *known* sampling distribution *under* H_0.

Example: How many calories in a Big Mac?

- Measure calories in random sample of n Big Macs: $X_1, \ldots, X_9 \sim$ iid $N(\mu, \sigma^2)$
- $H_0 \colon \mu = 550$

If McDonald's is telling the truth, i.e. under the null, what is *exact* sampling distribution of $(\bar{X} - 550)/(S/3)$?

(a) χ^2_9

(b) $N(550, 1)$

(c) $F(9, 1)$

(d) $N(0, 1)$

(e) t_8