Classical Inference

(Lecture)

Lucy Gao

January 9, 2024

About me

- Lucy L. Gao
- lucy.gao@stat.ubc.ca
- https://www.lucylgao.com/
- Assistant Professor, Department of Statistics

Course at a glance

Category	Contribution	Timing
Homework	40%	Week 3, Week 5
In-class Lab	20%	Week 1, Week 2, Week 3, Week 4
In-class Presentation	40%	Week 6

- Attendance and participation are important
- … But flexibility built in for special circumstances
- Assessment grades mostly based on effort

My approach to this class

Pre-requisites

- Comfortable with frequentist inference at “Casella-Berger” level
- Comfortable with coding, ideally in R

Treatment of content

- Problem forward, not solutions backwards
- Rigor/Theory Level: Moderate
- Breadth/depth of methods coverage: Low

Disclaimer

This is a newly developed course! Moreover, developed with little to no textbook support.

That means:

- Lots of parts reflect my current knowledge and opinions: these change over time
- Lots of experimental notation and presentation
- Course format and timing is untested!

Bottom line: Feedback very welcome for next iteration!

Statistics, science, and replicability

> **“Statistics is about science, and science is about proving things to people.”** - Scott S. Emerson, paraphrased

How do you prove something about the world?

- Show it once? *Not good enough.*
- Show twice? Ideally lots of times? *Better.*

Key idea: Random variation in experimentation and data collection should be accounted for when weighing the produced evidence.

Frequentist inference

(i.e. Hypothesis tests and confidence intervals)

Key Question: “To what extent does random variation inherent in experimentation affect our ability to draw conclusions?”

Answer: Summaries of what happens if we *repeat*, i.e. collect new data and repeat calculations.

Illustrations:

- **p-value**: “If the null hypothesis were true, then how often would I see such an extreme value of the test statistic, were I to *repeat* the experiment?”
- **Coverage**: “If I were to *repeat* the experiment, how often would I construct a confidence interval that contains the true parameter?”

Modern challenges to replicability

Experiments are now very complicated: lots of data, lots of scientific aims, lots of calculations

Now common settings:

- **Multiple testing**: Collect data and use it to make inference on more than one (non-data-adaptive!) thing
- **Selective inference**: Collect data, then use it to decide what to make inference on

In these settings:

- All ideas related to “collect new data and repeat calculations” get more complicated
- This complexity is not always taken into account in the statistical analysis

Agnostic statistics

> **"Everything should be as simple as possible, but no simpler."** - Albert Einstein (sort of)

Many textbooks/courses: We try to make inference about some parameters in a (usually simplified) finite-dimensional model.

- Eg. Use $X \sim N(\mu, \sigma^2)$ to learn about μ and σ^2.

Agnostic statistics: We try to make inference about some summaries of the model (*functionals*) for the data, and this model could be *infinite-dimensional*.

- Eg. Use $X \sim F$ to learn about $\mathbb{E}[X]$ and $\mathrm{Var}[X]$.

We will think more agnosticly in this class because (IMO) it's important to understanding selective inference.

Functionals

- Let F be a distribution, and $\mathcal{F}$ be a set of distributions.
- Let $T : \mathcal{F} \mapsto \mathbb{R}$ be a function.
- We call $T(F)$ a *functional*. Depending on context, we may also call $T(\cdot)$ a functional.

Example: The mean is a functional, with $T(F) = \int x \, dF(x)$. If we introduce $X \sim F$, then we can instead write $T(F) = \mathbb{E}[X]$.

***Informally*: a functional is a numerical summary of a distribution.**

Hypothesis testing

Let X be a random variable (could be scalar, vector, or matrix-valued) with distribution F.

The goal of a *hypothesis test* is to use a realization x from X to decide which of two complementary hypotheses is true:

- The null hypothesis, $H_0 : T(F) \in \Theta_0$
- The alternative hypothesis, $H_1 : T(F) \in \Theta_0^C$

We often expect a hypothesis test to output a **p-value**, quantifying the strength of evidence against the null hypothesis.

Evaluating a hypothesis test

What makes a test good?

Primary importance:

- Control the *Type 1 error rate*: the probability of rejecting the null if the null hypothesis were true
- Produce a *valid p-value*, i.e. one that is super-uniform under the null hypothesis.

Secondary importance:

- Have high *power*: the probablity of rejecting the null if the null hypothesis were not true

Why evaluate this way?

> **“I don’t know, it depends on the scientific context at hand.”** - My annoying partner, whenever I ask him ANY question involving statistics.

Crispest motivating example: clinical interventions and government regulation of these interventions, specifically in the USA.

Scientific challenge: How do we evaluate the efficacy and safety of a clinical intervention in a way that is timely, economical, and reliable?

- *Why economical?* Government cares about tax dollars, businesses care about profit.
- *Timely and reliable?* Medical ethics, government accountability.

Clinical trials

> **“The primary goal of clinical trials is to obtain a statistically reliable evaluation of whether the experimental intervention is safe and provides clinically meaningful benefit.” -** Thomas Fleming

Phase I: Basic dose range and basic safety (~70% move on)

- Small number of healthy volunteers, no randomization

Phase II: Dose finding and preliminary efficacy/safety (~30% move on)

- Enroll patients, usually randomized

Phase III: Definitive assessment of efficacy and safety (~25-30% move on)

- Enroll lots of patients, definitely randomize

Phase III clinical trials

Heavily simplified procedure:

- Recruit n patients from the population
- Randomize 50% of them to control, and randomize 50% of them to treatment
- Measure a single outcome, e.g. blood pressure. (Relax this next week!)
- Apply a two-tailed test of equality of some functional (e.g. mean, proportion, hazard rate) with $\alpha = 0.05$ to the outcome measurements
- Construct the 95% confidence interval as well to go with it

Everything in the analysis and design is decided before the study is conducted, and cannot be changed after the study is conducted.

Food and Drug Administration (FDA)

Heavily simplified policy: Approves a clinical intervention if applicant shows that they rejected the null hypothesis in a Phase III clinical trial.

Why? (According to me)

- If the FDA approves a clinical intervention, then it will be used to treat a lot of people, and will be assumed by the public to be safe and effective. If we assume wrong then it is a *catastrophe*.
- If every Phase III trial controls the Type I error rate at level 0.05, then the FDA can be pretty sure that among a very large amount of ineffective drugs, the proportion that the FDA approves (leading to catastrophe) is no more than 5%.
- In fact, in reality FDA often expects two independent Phase III trials before it approves, so that it can bound the catastrophe proportion by 0.25%.

What about power? What about CIs?

Role of a confidence interval: We do care about knowing *how much* benefit an intervention offers, not just about proving that it offers *nonzero* benefit. This is helpful information for clinicians to have.

Role of power: It is damaging to fail to approve an intervention that actually works:

- Intervention development is expensive and time-consuming
- Clinical trials are expensive and time-consuming
- Once the FDA rejects an application, the intervention is less likely to be studied again

So power is important. But it's **not a catastrophe** from a regulatory agency's POV to pass on an intervention that actually works. So priority 1 is the **type 1 error rate**, not power.

Type 1 error rate, with math

"The probability of rejecting the null hypothesis when it holds."

Let $X \sim F_X$. Suppose that we test $H_0 : T(F_X) \in \Theta_0$ using realizations from X.

We say that we control the type 1 error rate at level α if:

$$\mathbb{P}_{F_X}(\text{Reject } H_0 \text{ using } X) \leq \alpha, \quad \text{for all } F_X \in \mathcal{F}_0,$$

where $\mathcal{F}_0 = \{F : T(F) \in \Theta_0\}$. Equivalently,

$$\sup_{F_X \in \mathcal{F}_0} \mathbb{P}_{F_X}(\text{Reject } H_0 \text{ using } X) \leq \alpha.$$

It's often enough for the test to *asymptotically* control the type 1 error rate.

Type 1 error rate, more scientifically

"The probability of rejecting the null hypothesis when it holds."

$$\mathbb{P}_{F_X}(\text{Reject } H_0 \text{ using } X) \leq \alpha$$

Suppose that F_X satisfies H_0. (The data generating mechanism satisfies null hypothesis.) Then:

- Repeat the data collection N times to get N realizations ("replicate data sets") from $X \sim F_X$, for N very large
- Test H_0 using each realization

The type I error rate is the proportion of replicate data sets in which we reject the null hypothesis, when it holds.

Valid p-values

A p-value is *valid* if it is (super-)uniform under the null hypothesis.

Let $X \sim F_X$. Let $p(X)$ be a p-value computed on X. We say that the p-value is valid for $H_0 : T(F_X) \in \Theta_0$ if:

$$\mathbb{P}_{F_X}(p(X) \leq \alpha) \leq \alpha, \quad \text{for all } F_X \in \mathcal{F}_0 \text{ and all } 0 \leq \alpha \leq 1.$$

Equivalently,

$$\sup_{F_X \in \mathcal{F}_0} \mathbb{P}_{F_X}(p(X) \leq \alpha) \leq \alpha, \quad \text{for all } 0 \leq \alpha \leq 1.$$

Note that for $U \sim \text{Uniform}(0, 1)$, $P(U \leq \alpha) = \alpha$.

Valid p-values and type I error rate

Valid p-values are an “easy” way to construct a test that controls the type 1 error rate at level α:

- Reject H_0 on the basis of x whenever $p(x) \leq \alpha$.

To see this, compare the equations:

$$\text{Type I error rate:} \quad \sup_{F_X \in \mathcal{F}_0} \mathbb{P}_{F_X}(\text{Reject } H_0 \text{ using } X) \leq \alpha,$$

$$\text{Valid p-value:} \quad \sup_{F_X \in \mathcal{F}_0} \mathbb{P}_{F_X}(p(X) \leq \alpha) \leq \alpha.$$

Again, we are often happy enough if the p-value is *asymptotically* valid.

“Easy” valid p-value construction

Let $S(x)$ be a test statistic computed on a realization x from X, where large values give evidence against $H_0 : T(F_X) \in \Theta_0$.

Consider the p-value

$$p(x) = \sup_{F_X \in \mathcal{F}_0} \mathbb{P}_{F_X}(S(X) \geq S(x)).$$

Then, $p(X)$ is a valid p-value. (Prove for hwk. Hint: probability integral transform theorem.)

To make calculating the $\sup$ easier, try to pick test statistics $S(X)$ that are (asymptotically) *pivots*, i.e. the distribution of $S(X)$ is the same for any $F_X \in \mathcal{F}_0$.

Example 1: Two-sample testing

Let $X_1, \ldots, X_m$ be i.i.d. copies of $X^* \sim \mathbb{F}_X$. Let $Y_1, \ldots, Y_n$ be i.i.d. copies of $Y^* \sim \mathbb{F}_Y$. Let all X's be independent from all Y's. Using realizations $x_1, \ldots, x_m, y_1, \ldots, y_n$, test:

- $H_0 : \int x \, dF_X - \int y \, dF_Y = 0$
- $H_1 : \int x \, dF_X - \int y \, dF_Y \neq 0$

Here, the model is infinite-dimensional, and the functional we care about is the difference in means.

Take a simple random sample from each of two large populations, and use the data to test the null hypothesis that the means of the two populations are equal.

Example 1, continued

When $\min\{m, n\}$ is large, the two-sample t-test is practically equivalent to the following:

- **Test statistic**: $S(x, y) = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/m + s_y^2/n}}$
- **p-value**: $p(x, y) = \mathbb{P}(|Z| \geq |S(x, y)|)$ for $Z \sim N(0, 1)$.

Let $\mathcal{F}_0 = \{(F, F') : \int x \, dF = \int x \, dF'\}$. The p-value from the two-sample t-test can be viewed as approximating the following valid p-value,

$$\sup_{(F_X, F_Y) \in \mathcal{F}_0} \mathbb{P}_{F_X, F_Y}(|S(X, Y)| \geq |S(x, y)|)$$

When is this approximation warranted?

Example 1, continued

Conveniently, approximation is reasonable when $\min\{m, n\}$ is large!

When $(F_X, F_Y) \in \mathcal{F}_0$, we have $\mathbb{E}[X^*] - \mathbb{E}[Y^*] = 0$, and as $\min\{m, n\} \to \infty$,

$$\frac{\bar{X} - \bar{Y}}{\sqrt{s_X^2/m + s_Y^2/n}} \xrightarrow{d} N(0, 1).$$

i.e. $S(X, Y)$ asymptotically pivotal under the null, so the distribution of $S(X, Y)$ is basically close enough to the distribution Z for any $(F_X, F_Y) \in \mathcal{F}_0$.

(*Proof sketch*: Lindeberg-Feller CLT to show that $\frac{\bar{X}-\bar{Y}}{\mathrm{Var}(\bar{X}-\bar{Y})} \xrightarrow{d} N(0, 1)$, show $\frac{\mathrm{Var}(\bar{X}-\bar{Y})}{s_X^2/m+s_Y^2/n} \xrightarrow{p} 1$, continuous mapping.)

Example 1, wrapping up

We can say that the two-sample t-test (with unequal variance) asymptotically:

- Produces valid p-values … and in fact, converges to a Uniform(0, 1) distribution
- Controls the Type 1 error rate at nominal level α … and in fact, asymptotically achieves an exact error rate of α.

Note: Similarly agnostic treatment is available for linear regression, despite its heavily parametric reputation; we will see this in the selective inference part of the class.

Power, with math

"The probability of rejecting the null if the null hypothesis were false".

There are often lots of ways for the null hypothesis to not be true.

- Consider hypotheses of the form $H_0 : T(\mathbb{F}) = 0$ vs $H_1 : T(\mathbb{F}) \neq 0$.
- Rejection probability when $T(\mathbb{F}) = 10$ is naturally very different from rejection probability $\mathbb{F}$ satisfies $T(\mathbb{F}) = 0.1$!

For that reason, we care about a whole *power function*. Let $\mathcal{F}(c) = \{\mathcal{F} : T(\mathcal{F}) = c\}$, and assume that for any $c \in \mathbb{R}$, the test statistic's distribution is the same for all $F_X \in \mathcal{F}(c)$.

$$\text{Power}(c) = \mathbb{P}_{F_X \in \mathcal{F}(c)}(\text{Reject } H_0 \text{ based on } X).$$

Note that $\text{Power}(0)$ is the type I error rate!

Power, more scientifically

"The probability of rejecting the null if the null hypothesis were false".

$$\mathbb{P}_{F_X \in \mathcal{F}(c)}(\text{Reject } H_0 \text{ based on } X)$$

Suppose that $F_X \in \mathcal{F}(c)$. (The null hypothesis is false, and scalar functional of interest takes value c.) Then:

- Repeat the data collection N times to get N realizations ("replicate data sets") from $X \sim F_X$, for N very large
- Test H_0 using each realization

The power is the proportion of replicate data sets in which we reject the null hypothesis for a particular effect size.

Confidence sets/intervals

The goal of a *confidence set* is to use a draw from X to compute a *set-valued* estimate of the functional $T(F_X)$.

What's the point of this?

- Make a less strong assertion about the value of $T(F_X)$ than the point estimate.
- … in return, gain *confidence* about how the interval estimate relates to the value of $T(F_X)$

For simplicity:

- We usually only compute confidence sets for scalar-valued functionals, and …
- we additionally try to design them to guarantee that they are *intervals*.

Evaluating confidence intervals

Primary importance:

- Attain nominal coverage: this is how we define “gain confidence” about how the interval estimate relates to the value of $T(F_X)$
- If we don’t have nominal coverage, then we don’t gain any confidence in reporting an interval estimate, so what’s the point?

Secondary importance:

- Have small length, i.e. be short
- If the confidence interval is valid (has nominal coverage properties), then we can begin to care about making more precise assertions.

Coverage

Let $L(x)$ and $U(x)$ be functions that map data to a lower and upper limit for a confidence interval intended to cover $T(F_X)$, respectively.

We want the following to be true for any θ in Θ, the range of $T(\cdot)$:

$$\mathbb{P}_{F_X}(L(X) \geq \theta \text{ and } U(X) \leq \theta) \geq 1 - \alpha, \quad \text{for all } F_X \in \mathcal{F}_\theta \equiv \{F : T(F) = \theta\}.$$

Equivalently,

$$\inf_{\theta \in \Theta} \inf_{F_X \in \mathcal{F}_\theta} \mathbb{P}_{F_X}(L(X) \geq \theta \text{ and } U(X) \leq \theta) \geq 1 - \alpha.$$

"Across many repetitions of the study, the proportion of the intervals containing $T(F_X)$ is at least $1 - \alpha$, regardless of what F_X is, and what the value of $T(F_X)$ is."

Length

The length of an interval constructed with a realization x (e.g. the data from a specific experiment) is $[L(x), U(x)]$ is $U(x) - L(x)$.

The length of the intervals you construct when *repeating* the experiment is a *random variable*, $U(X) - L(X)$.

This random variable has a distribution to be summarized like any other. Commonly seen summaries include:

- Expected length, $\mathbb{E}[U(X) - L(X)]$
- Median length, $\text{median}(U(X) - L(X))$.

Note: Not everyone agrees that these are good summaries. See e.g. Pratt (JASA 1961).

Construction: “Inverting a valid test”

Recall the following result, an “agnostification” of Theorem 9.2.2 in Casella and Berger:

Every confidence set corresponds to a family of tests and vice versa

Let $X \sim F_X$. Let $T(\cdot)$ be a functional, and let Θ be the range of T. For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : T(F_X) = \theta_0$. Furthermore, define

$$C(x) = \{\theta_0 : x \in A(\theta_0)\}.$$

Then, $C(X)$ is a confidence set for $T(F_X)$ with $100(1-\alpha)\%$ coverage. Conversely, suppose that $C(X)$ is a confidence set for $T(F_X)$ with $(1-\alpha)\%$ coverage. Then, for any $\theta_0 \in \Theta$, define

$$A(\theta_0) = \{x : \theta_0 \in C(x)\}.$$

Then, using $A(\theta_0)$ to construct the acceptance region of a test of $H_0 : T(F_X) = \theta_0$ yields a test of level α.

Result implications

1. As long as we can make valid tests of $H_0 : T(F_X) = \theta_0$ for any θ_0, we can make a valid confidence set.

2. The following interpretation of a $100(1-\alpha)\%$ confidence interval $[L(x), U(x)]$ is valid:

> "[L(x), U(x)] contains all values of θ_0 such that I would not reject the null hypothesis $H_0 : \theta = \theta_0$ based on x at level α."

Informally:

> "[L(x), U(x)] contains all values of θ_0 that (to my best knowledge) could plausibly have generated the data x."

Example 2: Two-sample inference

Let $X_1, \ldots, X_m$ be i.i.d. copies of $X^* \sim F_X$. Let $Y_1, \ldots, Y_n$ be i.i.d. copies of $Y^* \sim F_Y$. Let all X's be independent from all Y's. Using realizations $x_1, \ldots, x_m, y_1, \ldots, y_n$, construct an interval estimate of the quantity $\int x \, dF_X - \int y \, dF_Y = \mathbb{E}[X^*] - \mathbb{E}[Y^*]$.

Inverting the two-sample t-test (essentially) yields the following interval bounds:

- $L(x) = \bar{x} - \bar{y} - Z_{1-\alpha/2}\sqrt{s_x^2/m + s_y^2/n}$
- $U(x) = \bar{x} - \bar{y} + Z_{1-\alpha/2}\sqrt{s_x^2/m + s_y^2/n}$

Here, $Z_{1-\alpha/2}$ is the $100(1-\alpha/2)$th quantile of the standard normal distribution.

Example 2, continued

We know from earlier that the two-sample t-test is asymptotically valid as $\min\{m, n\} \to \infty$.

Technically we derived it for testing $H_0 : \mathbb{E}[X^*] - \mathbb{E}[Y^*] = 0$, but we could have adapted it to $H_0 : \mathbb{E}[X^*] - \mathbb{E}[Y^*] = c$ for any $c \in \mathbb{R}$....

Since we got our interval by inverting this family of asymptotically two-sample t-tests, it follows that the interval has **asymptotically correct coverage**.

This justifies the use of a t-interval when $\min\{m, n\}$ is sufficiently large.