PH 712 Probability and Statistical Inference

Part VIII: Hypothesis Testing

Zhiyang Zhou (zhou67@uwm.edu, zhiyanggeezhou.github.io)

2025/11/29 15:24:12

Recall the (two-sided) t-test

- Assumption: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$ with unknown μ and σ^2
- Hypotheses: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$
- Test statistic:
$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$
  - (Sample variance) $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$
- Level α rejection region:
$$\{(x_1, \ldots, x_n) : |T| \geq t_{n-1,1-\alpha/2}\},$$
  - $t_{n-1,1-\alpha/2}$: the $(1-\alpha/2)$ quantile of t distribution with $n-1$ degrees of freedom.
- p-value:
$$2\left\{1 - F_{t(n-1)}(|T|)\right\},$$
  - $F_{t(n-1)}(\cdot)$: cdf of t distribution with $n-1$ degrees of freedom
- Decision rule:
  - Reject H_0 if $|T| \geq t_{n-1,1-\alpha/2}$ or p-value $\leq \alpha$; otherwise, accept H_0.
- Hypothesis testing is a route to deciding between two classes based on observed data

A binary classification problem: Is it a squirrel?

Figure 1: Potential Squirrel (Photograph by Joel Sartore)

- Make a decision between two hypotheses H_0 : YES and H_1: NO.
  - Checking necessary conditions under H_0: e.g., size, color, tail, behavior, habitat, etc.

Problem formalization

- Assumptions
  - $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x \mid \theta)$
    - θ is fixed and unknown BUT is believed to be inside Θ
  - To make a decision on θ between two hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$
    - $\Theta_0 \cup \Theta_1 = \Theta$
    - $\Theta_0 \cap \Theta_1 = \emptyset$
- Four possible outcomes
  - True positive (TP): H_0 is wrong (i.e., H_1 is true) and we reject H_0 (i.e., accept H_1);
  - False positive (FP, type I error): H_0 is true (i.e., H_1 is wrong) but we reject H_0 (i.e., accept H_1);
  - True negative (TN): H_0 is true (i.e., H_1 is wrong) and we accept H_0 (i.e., reject H_1);
  - False negative (FN, type II error): H_0 is wrong (i.e., H_1 is true) but we accept H_0 (i.e., reject H_1).
  - E.g., in the context of identifying the animal,
    - TP: it is NOT a squirrel and is NOT identified as a squirrel
    - FP: it is a squirrel but is NOT identified as a squirrel
    - TN: it is a squirrel and is identified as a squirrel
    - FN: it is NOT a squirrel but is identified as a squirrel

	Accept H_0	Reject H_0
H_0 is true	True negative (TN)	False positive (FP, type I error)
H_0 is false	False negative (FN, type II error)	True positive (TP)

- Different objectives leading to different strategies:
  - Minimizing the misclassification rate: $\Pr(\text{FP}) + \Pr(\text{FN})$
    - Commonly adopted by binary classification techniques
  - Controlling the false discovery rate (FDR): $\Pr(\text{FP})/\{\Pr(\text{FP}) + \Pr(\text{TP})\}$
    - For sequential or simultaneous testing
  - Minimizing Pr(type II error) with Pr(type I error) $\leq \alpha$
    - Leading to the optimal hypothesis test

Formalizing the hypothesis test

- A test, say ϕ, is an indicator function

$$\phi(x_1, \ldots, x_n) = \mathbf{1}_R(x_1, \ldots, x_n) = \begin{cases} 0, & (x_1, \ldots, x_n) \notin R \\ 1, & (x_1, \ldots, x_n) \in R \end{cases}$$

  - Input: the sample or its realization
  - Output: the action after observing the input, i.e., 0 (accepting H_0) or 1 (rejecting H_0)
  - *Rejection region*: R, the set corresponding to the rejection of H_0
    - R is typically specified in terms of the realization of a *test statistic*; e.g., if $R = \{(x_1, \ldots, x_n) : \bar{x} \geq 3\}$, then $\bar{X}$ is a test statistic.
- Each test corresponds to a unique rejection region
  - Two tests are equivalent $\Leftrightarrow$ their rejection regions are identical

Uniformly most powerful (UMP) level α test

- *Power function*: given a test ϕ and its rejection region R, the power function $\beta_\phi(\theta)$ is the probability of rejecting H_0, i.e.,

$$\beta_\phi(\theta) = \Pr\{(X_1, \ldots, X_n) \in R\} = \Pr\{\phi(X_1, \ldots, X_n) = 1\}$$

  - Pr(type I error) $= \beta_\phi(\theta)$ if $\theta \in \Theta_0$
  - Pr(type II error) $= 1 - \beta_\phi(\theta)$ if $\theta \in \Theta_1$

- Since the true θ is unknown, a good test requires small $\beta_\phi(\theta)$ for all $\theta \in \Theta_0$ AND large $\beta_\phi(\theta)$ for all $\theta \in \Theta_1$

- A test ϕ is of *level* $\alpha \Leftrightarrow \sup_{\theta\in\Theta_0} \beta_\phi(\theta) \le \alpha \Leftrightarrow$ the maximum of $\beta_\phi(\theta)$ in the closure of Θ_0
  - A test ϕ is of *level* $\alpha \Rightarrow$ its type I error rate $\le \alpha$
- Let ϕ be a level α test for $H_0 : \theta_0 \in \Theta_0$ vs $H_1 : \theta_0 \in \Theta_1$. If $\beta_\phi(\theta) \ge \beta_{\phi'}(\theta)$ for all $\theta \in \Theta_1$ and any other test ϕ' of level α, then ϕ is a UMP level α test.
  - That is, UMP level α test minimizes the type II error rate among all the level α tests.

Example Lec8.1

- (Calculating the sample size of a clinical trial) A pharmaceutical company is running a clinical trial of a new drug for lowering systolic blood pressure (SBP). For the ith enrolled patient, let X_i denote the change in SBP (in mm Hg) from baseline to 12 weeks. Specifically, X_i = baseline - the measure at week 12 (i.e., larger values mean more SBP reduction). Assume $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 100)$ with unknown θ. People want to test whether the drug achieves a prespecified target mean reduction $\theta_0 > 0$, i.e., $H_0 : \theta \le \theta_0$ vs $H_1 : \theta > \theta_0$. Consider the rejection region $\{(x_1, \ldots, x_n) : \sqrt{n}(\bar{x} - \theta_0)/10 > c\}$.
  1. Elaborate the power function.
  2. Find sample size n and threshold c if the desired type I error rate at θ_0 is 5% and the type II error rate at $\theta_0 + \sigma$ is at most 25%.

```
n_max = 100
c = qnorm(1-.025)
type2.err.rates = rep(NA, n_max)
for (n in 1:n_max) {
  type2.err.rates[n] = pnorm(c-n^.5)-pnorm(-c-n^.5)
  if (type2.err.rates[n] <= .25) {
    break
  }
}
type2.err.rates
```

Likelihood ratio test (LRT)

- Hypotheses: $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$
  - $\Theta = \Theta_0 \cup \Theta_1$
  - $\Theta_0 \cap \Theta_1 = \emptyset$
- Test statistic
$$\lambda(X_1, \ldots, X_n) = \frac{L(\hat{\theta}_{\text{ML},0})}{L(\hat{\theta}_{\text{ML}})}$$
  - $L(\cdot)$: the likelihood function
  - $\hat{\theta}_{\text{ML},0}$: MLE of θ under H_0
  - $\hat{\theta}_{\text{ML}}$: MLE of $\theta \in \Theta$
- Level α rejection region
$$R = \{(x_1, \ldots, x_n) : \lambda(x_1, \ldots, x_n) \le c_\alpha\},$$
where critical point c_α is chosen to make sure
$$\sup_{\theta\in\Theta_0} \beta_\phi(\theta) = \sup_{\theta\in\Theta_0} \Pr\{\lambda(X_1, \ldots, X_n) \le c_\alpha\} = \alpha.$$
  - Actually ensuring Pr(type I error) $\le \alpha$ since Pr(type I error) $\le \sup_{\theta\in\Theta_0} \beta_\phi(\theta)$
  - Essential but challenging to know the distribution of $\lambda(X_1, \ldots, X_n)$ under H_0
- Implementation
  1. Confirm the value of α;
  2. Figure out $\hat{\theta}_{\text{ML},0}$ and $\hat{\theta}_{\text{ML}}$.

3. Solve the following equation for c_α

$$\sup_{\theta\in\Theta_0} \beta_\phi(\theta) = \sup_{\theta\in\Theta_0} \Pr\{\lambda(X_1,\ldots,X_n) \le c_\alpha\} = \alpha;$$

4. Reject H_0 if $\lambda(x_1,\ldots,x_n) \le c_\alpha$.

- LRT is promoted by math theorems
  - (Neyman-Pearson Lemma) LRT is the UMP level α test for simple hypotheses ($H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$)
  - (Karlin-Rubin theorem) under certain conditions, LRT is the UMP level α test for one-sided hypotheses ($H_0 : \theta \le \theta_0$ (or $\theta = \theta_0$) vs $H_1 : \theta > \theta_0$ OR $H_0 : \theta \ge \theta_0$ (or $\theta = \theta_0$) vs $H_1 : \theta < \theta_0$)
  - There is No UMP test for two-sided hypotheses ($H_0 : \theta = \theta_0$ vs $H_1 : \theta \ne \theta_0$) but LRT is UMP unbiased test for this scenario.
- Special cases
  - Equivalent to the Z-test if 1) the sample is iid normal with known variance and 2) the mean is to be tested
  - Equivalent to the t-test if 1) the sample is iid normal with unknown variance and 2) the mean is to be tested
  - Equivalent to the F-test if 1) the sample is iid normal with the mean and variance both unknown and 2) the variance is to be tested

LRT (asymptotic)

- Asymptotic level α rejection region

$$\{(x_1,\ldots,x_n) : \lambda(x_1,\ldots,x_n) \le \exp(-\chi^2_{\nu,1-\alpha}/2)\},$$

  where $\chi^2_{\nu,1-\alpha}$ is the $(1-\alpha)$ quantile of $\chi^2(\nu)$, i.e., $F_{\chi^2(\nu)}(\chi^2_{\nu,1-\alpha}) = 1-\alpha$.
  - I.e., asymptotically, $c_\alpha \approx \exp(-\chi^2_{\nu,1-\alpha}/2)$
  - Because, as $n\to\infty$, under H_0,

$$-2\ln\lambda(X_1,\ldots,X_n) \approx \chi^2(\nu),$$

    where $\nu =$ the difference of numbers of free parameters between Θ_0 and Θ.
- Implementation (asymptotic)
  1. Confirm the value of α;
  2. Figure out $\hat\theta_{\mathrm{ML},0}$ and $\hat\theta_{\mathrm{ML}}$;
  3. Check ν, the difference of numbers of free parameters between Θ_0 and Θ;
  4. The rejection region is $\{(x_1,\ldots,x_n) : \lambda(x_1,\ldots,x_n) \le \exp(-\chi^2_{\nu,1-\alpha}/2)\}$ $\Leftrightarrow$ reject H_0 if $\lambda(x_1,\ldots,x_n) \le \exp(-\chi^2_{\nu,1-\alpha}/2)$.
- Numerical illustration of the chi-square approximation of $-2\ln\lambda(X_1,\ldots,X_n)$ under H_0: Collecting sample $X_1,\ldots,X_{1000} \overset{\text{iid}}{\sim} f_X(x \mid p=1/4) = (1/4)^x(3/4)^{1-x}$, $x=0,1$, we test $H_0 : p = 1/4$ vs. $H_1 : p \ne 1/4$.
  1. Generate $B = 10,000$ bootstrap samples of $-2\ln\lambda(X_1,\ldots,X_{1000})$, where $\lambda(X_1,\ldots,X_{1000})$ is the likelihood ratio.
  2. Generate a figure to compare the simulated distribution of $-2\ln\lambda(X_1,\ldots,X_{1000})$ and the chi-square approximation.

```
options(digits = 4)
set.seed(712)
n = 1e3L # size of each bootstrap sample
B = 1e4L # number of bootstrap samples
test_stats = numeric(B)
p0 = 1/4
ell = function(p, Xs){
  log(p)*sum(Xs)+log(1-p)*(n-sum(Xs))
```

```
}
for (i in 1:B){
  Xs = rbinom(n, 1, p0)
  p_ml_0 = p0
  p_ml = optim(
      par = .5, lower = .00001, upper = .99999,
      fn = ell, Xs = Xs,
      method="L-BFGS-B",
      control=list(fnscale=-1))$par
  test_stats[i] = -2*(
    ell(p_ml_0, Xs) -
    ell(p_ml, Xs)
  )
}
seg = seq(0, 10, length.out=100)
pdfchi2 = dchisq(seg, 1)
hist(test_stats, breaks=100, xlim=c(0,10),
     freq=F, xlab = expression(paste('-2ln', lambda, '(x)')), main = '')
lines(seg, pdfchi2, col = "red")
```

Example Lec8.2

- For a given city in a given year, assume that the number of automobile accidents follows a Poisson distribution. In past years the average number of accidents per year was 15, and this year it was 10. Is it justified to claim that the accident rate has dropped?
- Demo report: Testing hypotheses H_0 : ___ vs. H_1 : ___ , we carried out the ___ test and obtained ___ as the value of test statistic. Since the critical point is ___ , there was/wasn't a strong statistical evidence against H_0 at the ___ (significance) level, i.e., we believed that ___ .

Historically, In a certain county, the annual number of new Lyme disease cases has historically followed a Poisson distribution with a mean of 42 cases per year. This year, only 25 cases were reported. Is it justified to claim that the incidence of Lyme disease has significantly decreased?

- Demo report: Testing hypotheses H_0 : ___ vs. H_1 : ___ , we carried out the ___ test and obtained ___ as the value of test statistic. Since the critical point is ___ , there was/wasn't a strong statistical evidence against H_0 at the ___ (significance) level, i.e., we believed that ___ .

Example Lec8.3 (A/B Testing)

At a large social media company, the historical (pre-experiment) daily active user (DAU) login success rate across the entire platform has been extremely stable at 99.30% for the past two years. The authentication team wants to roll out a new biometric login flow (fingerprint + face ID) that they believe will be faster and more reliable than the current password + 2FA flow. They run a properly randomized A/B test for 7 days:

Control (old login flow): 2,000,000 login attempts with 1,986,000 successes Treatment (new biometric login flow): 2,000,000 login attempts → 1,990,200 successes

The product manager comes to you and says: "Historically, our login success rate has been 99.3%. With the new biometric flow we observed 99.51%. That's a 0.21 percentage point increase! Can we declare that the new biometric login system has significantly improved the login success rate and start rolling it out to 100% of users?"

p-value

- Motivation
  - Recall that the level α rejection region R consists of a test statistic (e.g., $\lambda(X_1, \ldots, X_n)$ for LRT) and a critical point (e.g., c_α for LRT)
  - Would like to fix the critical point to be α by defining a test statistic $p(X_1, \ldots, X_n)$ (i.e., p-value) such that
$$R = \{(x_1, \ldots, x_n) : p(x_1, \ldots, x_n) \le \alpha\}$$
    * More convenient in communication because the critical point is α by default

- NOT always well-defined but working well for LRT
  - For LRT, asymptotically,
$$p\text{-value} = 1 - F_{\chi^2(\nu)}(-2 \ln \lambda(x_1, \ldots, x_n)).$$
    * $F_{\chi^2(\nu)}(\cdot)$: the cdf of $\chi^2(\nu)$

Revisit Example Lec8.2

- For a given city in a given year, assume that the number of automobile accidents follows a Poisson distribution. In past years the average number of accidents per year was 15, and this year it was 10. Is it justified to claim that the accident rate has dropped?

- Demo report: Testing hypotheses H_0 : ___ vs. H_1 : ___ , we carried out the ___ test and obtained ___ as the p-value. So, at the ___ (significance) level, there was/wasn't a strong statistical evidence against H_0, i.e., we believed that ___ .

Wald test

- Testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \ne \theta_0$

- Test statistic: $(\hat{\theta}_{\text{ML}} - \theta_0)/\sqrt{\widehat{\text{var}}(\hat{\theta}_{\text{ML}})}$
  - Asymptotically equivalent to LRT for hypotheses $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \ne \theta_0$
  - Refer to the previous part for how to obtain $\widehat{\text{var}}(\hat{\theta}_{\text{ML}})$ (via the observed Fisher information or delta method)

- Level α Wald rejection region: $\{(x_1, \ldots, x_n) : |\hat{\theta}_{\text{ML}} - \theta_0|/\sqrt{\widehat{\text{var}}(\hat{\theta}_{\text{ML}})} \ge \Phi^{-1}_{1-\alpha/2}\}$
  - $\Phi^{-1}_{1-\alpha/2}$: the $(1 - \alpha/2)$ quantile of $\mathcal{N}(0, 1)$

- p-value $= 2\Phi\left(-|\hat{\theta}_{\text{ML}} - \theta_0|/\sqrt{\widehat{\text{var}}(\hat{\theta}_{\text{ML}})}\right)$
  - $\Phi(\cdot)$: cdf of $\mathcal{N}(0, 1)$

Revisit Example Lec8.2

- For a given city in a given year, assume that the number of automobile accidents follows a Poisson distribution. In past years the average number of accidents per year was 15, and this year it was 10. Is it justified to claim that the accident rate has be changed?

- Demo report: Testing hypotheses H_0 : ___ vs. H_1 : ___ , we carried out the ___ test and obtained ___ as the p-value. So, at the ___ (significance) level, there was/wasn't a strong statistical evidence against H_0, i.e., we believed that ___ .