

# PH 712 Probability and Statistical Inference

## Part VIII: Hypothesis Testing

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

2025/12/01 13:12:03

---

### Recall the (two-sided) $t$ -test

- Assumption:  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} (\mu, \sigma^2)$  with unknown  $\mu$  and  $\sigma^2$

- 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  $\alpha$  rejection region:

$$\{(x_1, \dots, x_n) : |T| \geq t_{n-1, 1-\alpha/2}\},$$

- $t_{n-1, 1-\alpha/2}$ : the  $(1 - \alpha/2)$  quantile of  $t$  distribution with  $n - 1$  degrees of freedom.

- $p$ -value:

$$2 \{1 - F_{t(n-1)}(|T|)\},$$

- $F_{t(n-1)}(\cdot)$ : 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
  - $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x | \theta)$ 
    - \*  $\theta$  is fixed and unknown BUT is believed to be inside  $\Theta$
  - To make a decision on  $\theta$  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  $\Pr(\text{type II error})$  with  $\Pr(\text{type I error}) \leq \alpha$ 
    - \* Leading to the optimal hypothesis test

## Formalizing the hypothesis test

- A test, say  $\phi$ , is an indicator function

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

- 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  $R = \{(x_1, \dots, x_n) : \bar{x} \geq 3\}$ , 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 $\alpha$ test

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

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

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

- Since the true  $\theta$  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  $\phi$  is of *level*  $\alpha \Leftrightarrow \sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \alpha \Leftrightarrow$  the maximum of  $\beta_\phi(\theta)$  in the closure of  $\Theta_0$ 
  - A test  $\phi$  is of *level*  $\alpha \Rightarrow$  its type I error rate  $\leq \alpha$
- Let  $\phi$  be a level  $\alpha$  test for  $H_0 : \theta_0 \in \Theta_0$  vs  $H_1 : \theta_0 \in \Theta_1$ . If  $\beta_\phi(\theta) \geq \beta_{\phi'}(\theta)$  for all  $\theta \in \Theta_1$  and any other test  $\phi'$  of level  $\alpha$ , then  $\phi$  is a UMP level  $\alpha$  test.
  - That is, UMP level  $\alpha$  test minimizes the type II error rate among all the level  $\alpha$  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  $i$ th enrolled patient, let  $X_i$  denote the change in SBP (in mm Hg) from baseline to 12 weeks. Specifically,  $X_i = \text{baseline} - \text{measure at week 12}$  (i.e., larger values mean more SBP reduction). Assume  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\theta, 100)$  with unknown  $\theta$ . People want to test whether the drug achieves a prespecified target mean reduction  $\theta_0 > 0$ , i.e.,  $H_0 : \theta \leq \theta_0$  vs  $H_1 : \theta > \theta_0$ . Consider the rejection region  $\{(x_1, \dots, x_n) : \sqrt{n}(\bar{x} - \theta_0)/10 > c\}$ .
  1. Elaborate the power function.
  2. Find sample size  $n$  and threshold  $c$  if the desired type I error rate at  $\theta_0$  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

$$\lambda(X_1, \dots, X_n) = \frac{L(\hat{\theta}_{ML,0})}{L(\hat{\theta}_{ML})}$$

- $L(\cdot)$ : the likelihood function
- $\hat{\theta}_{ML,0}$ : MLE of  $\theta$  under  $H_0$
- $\hat{\theta}_{ML}$ : MLE of  $\theta \in \Theta$
- Level  $\alpha$  rejection region

$$R = \{(x_1, \dots, x_n) : \lambda(x_1, \dots, x_n) \leq c_\alpha\},$$

where critical point  $c_\alpha$  is chosen to make sure

$$\sup_{\theta \in \Theta_0} \beta_\phi(\theta) = \sup_{\theta \in \Theta_0} \Pr\{\lambda(X_1, \dots, X_n) \leq c_\alpha\} = \alpha.$$

- Actually ensuring  $\Pr(\text{type I error}) \leq \alpha$  since  $\Pr(\text{type I error}) \leq \sup_{\theta \in \Theta_0} \beta_\phi(\theta)$
- Essential but challenging to know the distribution of  $\lambda(X_1, \dots, X_n)$  under  $H_0$
- Implementation
  1. Confirm the value of  $\alpha$ ;
  2. Figure out  $\hat{\theta}_{ML,0}$  and  $\hat{\theta}_{ML}$ .

3. Solve the following equation for  $c_\alpha$

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

4. Reject  $H_0$  if  $\lambda(x_1, \dots, x_n) \leq c_\alpha$ .

- LRT is promoted by math theorems
  - (Neyman-Pearson Lemma) LRT is the UMP level  $\alpha$  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  $\alpha$  test for one-sided hypotheses ( $H_0 : \theta \leq \theta_0$  (or  $\theta = \theta_0$ ) vs  $H_1 : \theta > \theta_0$  OR  $H_0 : \theta \geq \theta_0$  (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  $H_1 : \theta \neq \theta_0$ ) 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  $\alpha$  rejection region

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

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  $n \rightarrow \infty$ , under  $H_0$ ,

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

where  $\nu =$  the difference of numbers of free parameters between  $\Theta_0$  and  $\Theta$ .

- Implementation (asymptotic)
  1. Confirm the value of  $\alpha$ ;
  2. Figure out  $\hat{\theta}_{ML,0}$  and  $\hat{\theta}_{ML}$ ;
  3. Check  $\nu$ , the difference of numbers of free parameters between  $\Theta_0$  and  $\Theta$ ;
  4. The rejection region is  $\{(x_1, \dots, x_n) : \lambda(x_1, \dots, x_n) \leq \exp(-\chi^2_{\nu, 1-\alpha}/2)\} \Leftrightarrow$  reject  $H_0$  if  $\lambda(x_1, \dots, x_n) \leq \exp(-\chi^2_{\nu, 1-\alpha}/2)$ .
- Numerical illustration of the chi-square approximation of  $-2 \ln \lambda(X_1, \dots, X_n)$  under  $H_0$ : Collecting sample  $X_1, \dots, X_{1000} \stackrel{iid}{\sim} f_X(x | p = 1/4) = (1/4)^x (3/4)^{1-x}$ ,  $x = 0, 1$ , we test  $H_0 : p = 1/4$  vs.  $H_1 : p \neq 1/4$ .
  1. Generate  $B = 10,000$  bootstrap samples of  $-2 \ln \lambda(X_1, \dots, X_{1000})$ , where  $\lambda(X_1, \dots, X_{1000})$  is the likelihood ratio.
  2. Generate a figure to compare the simulated distribution of  $-2 \ln \lambda(X_1, \dots, X_{1000})$  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  $H_0 : \text{_____}$  vs.  $H_1 : \text{_____}$ , 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 \_\_\_\_\_.
- Adapting the demo report to this question: Testing hypotheses  $H_0 : \text{accident rate} = 15$  vs.  $H_1 : \text{accident rate} < 15$ , we carried out the asymptotic likelihood ratio test and obtained .39 as the value of test statistic. Since the critical point is .15, there wasn't a statistical evidence against  $H_0$  at the .05 (significance) level, i.e., we believed that the accident rate hasn't dropped.

```

xs = 10
n = length(xs)
alpha = .05
ell = function(lambda, xs){
  -lambda*n + sum(xs)*log(lambda)-sum(log(factorial(xs)))
}
lambda_ml_0 = 15
lambda_ml = optim(
  par = 10, lower = .00001, upper = 15,
  fn = ell, xs = xs,
  method="L-BFGS-B",
  control=list(fnscale=-1))$par
test_stat = exp( # likelihood ratio is exp(difference of log-likelihoods)
  ell(lambda_ml_0, xs) -
  ell(lambda_ml, xs)
)
test_stat <= exp(-qchisq(1-alpha, df=1)/2)

```

---

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

```
xs = 25
n = length(xs)
alpha = .05
ell = function(lambda, xs){
  -lambda*n + sum(xs)*log(lambda)-sum(log(factorial(xs)))
}
lambda_ml_0 = 42
lambda_ml = optim(
  par = 10, lower = .00001, upper = 42,
  fn = ell, xs = xs,
  method="L-BFGS-B",
  control=list(fnscale=-1))$par
test_stat = exp( # likelihood ratio is exp(difference of log-likelihoods)
  ell(lambda_ml_0, xs) -
  ell(lambda_ml, xs))
)
test_stat <= exp(-qchisq(1-alpha, df=1)/2)
```

### 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  $\rightarrow$  1,986,000 successes;
- Treatment (new biometric login flow): 2,000,000 login attempts  $\rightarrow$  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  $\alpha$  rejection region  $R$  consists of a test statistic (e.g.,  $\lambda(X_1, \dots, X_n)$  for LRT) and a critical point (e.g.,  $c_\alpha$  for LRT)
  - Would like to fix the critical point to be  $\alpha$  by defining a test statistic  $p(X_1, \dots, X_n)$  (i.e., *p*-value) such that

$$R = \{(x_1, \dots, x_n) : p(x_1, \dots, x_n) \leq \alpha\}$$

\* More convenient in communication because the critical point is  $\alpha$  by default

- NOT always well-defined but working well for LRT
  - For LRT, asymptotically,

$$p\text{-value} = 1 - F_{\chi^2(\nu)}(-2 \ln \lambda(x_1, \dots, x_n)).$$

\*  $F_{\chi^2(\nu)}(\cdot)$ : 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  $H_0 : \underline{\hspace{2cm}}$  vs.  $H_1 : \underline{\hspace{2cm}}$ , 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.  $H_1 : \theta \neq \theta_0$
- Test statistic:  $(\hat{\theta}_{ML} - \theta_0)/\sqrt{\widehat{\text{var}}(\hat{\theta}_{ML})}$ 
  - Asymptotically equivalent to LRT for hypotheses  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta \neq \theta_0$
  - Refer to the previous part for how to obtain  $\widehat{\text{var}}(\hat{\theta}_{ML})$  (via the observed Fisher information or delta method)
- Level  $\alpha$  Wald rejection region:  $\{(x_1, \dots, x_n) : |\hat{\theta}_{ML} - \theta_0|/\sqrt{\widehat{\text{var}}(\hat{\theta}_{ML})} \geq \Phi_{1-\alpha/2}^{-1}\}$ 
  - $\Phi_{1-\alpha/2}^{-1}$ : the  $(1 - \alpha/2)$  quantile of  $\mathcal{N}(0, 1)$
- $p$ -value =  $2\Phi\left(-|\hat{\theta}_{ML} - \theta_0|/\sqrt{\widehat{\text{var}}(\hat{\theta}_{ML})}\right)$ 
  - $\Phi(\cdot)$ : 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 been changed?
- Demo report: Testing hypotheses  $H_0 : \underline{\hspace{2cm}}$  vs.  $H_1 : \underline{\hspace{2cm}}$ , 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  .