Week 5 QCB 408/508 Spring 2020

Hypothesis Testing

Making probabilistic statements about the value of the parameter(s) of interest wrt categories (or sets) of interest.

Simple vs. Simple hypotheses:

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$

$H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$

θ_0: null value

θ_1: alternative value

Significance regions:

$\Gamma_c = \{ \underline{x} : L(\theta_1; \underline{x}) / L(\theta_0; \underline{x}) \geq c \}$ $c \geq 0$

$c' \leq c \Rightarrow \Gamma_c \subseteq \Gamma_{c'}$

Γ_c as a function of $c \geq 0$, this forms a nested set of significance regions.

$S(\underline{x}) = \dfrac{L(\theta_1; \underline{x})}{L(\theta_0; \underline{x})} \Rightarrow S(\underline{x}) \geq c$ captures Γ_c.

$S(\underline{x})$ is an example of a test-statistic.

	test is "significant"	test is not "significant"
H_0 is true	false positive	true negative
H_1 is true	true positive	false negative

false positive = Type I error

false negative = Type II error

false positive rate = Type I error rate

$$= \int_{\Gamma_c} f(\underline{x};\theta_0)\,d\underline{x} = \Pr(\underline{X} \in \Gamma_c ; \theta = \theta_0)$$

false negative rate = Type II error rate

$$= 1 - \int_{\Gamma_c} f(\underline{x};\theta_1)\,d\underline{x} = 1 - \Pr(\underline{X} \in \Gamma_c ; \theta = \theta_1)$$

Power = 1 − false negative rate = $\Pr(X \in \Gamma_c ; \theta = \theta_1)$

Type II

Neyman - Pearson Lemma

$\Gamma_c = \{ \underline{x} : L(\theta_1; \underline{x}) / L(\theta_0; \underline{x}) \geq c \}$

are the "most powerful" significance regions.

Given a FPR (false positive rate), the highest power possible is Γ_c.

Generalized Likelihood Ratio Test

Θ be the set of all possible θ

Θ_0 null values

Θ_1 alternative values

$\Theta_0 \cap \Theta_1 = \phi$

$\Theta_0 \cup \Theta_1 = \Theta$

$H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$

Test statistic: $\lambda(\underline{x}) = \dfrac{\max_{\theta \in \Theta} L(\theta; \underline{x})}{\max_{\theta \in \Theta_0} L(\theta; \underline{x})}$

$$= \frac{L(\hat{\theta}_{MLE}; \underline{x})}{L(\hat{\theta}_{0,MLE}; \underline{x})}$$

Null distribution: the probability distribution of the test-statistic when H_0 is true

Here, we have an approximate ($n \to \infty$) null distribution ($\theta \in \Theta_0$).

When H_0 is true, $2\log(\lambda(\underline{x})) \xrightarrow{D} \chi^2_r$.

$$r = \dim(\Theta) - \dim(\Theta_0).$$

$$\Gamma_c = \{ \underline{x} : \lambda(\underline{x}) \geq c \}$$

$$\text{Type I error rate} = \int_{\Gamma_c} f(\underline{x}; \theta_0)\, d\underline{x}$$

$$\underset{n \to \infty}{\approx} \int_{\Gamma_c} g(\underline{x})\, d\underline{x} \quad \text{where } g(\underline{x}) \text{ is } \chi^2_r \text{ pdf.}$$

p-value is the minimum Type I error rate obtainable when calling a test based on data $\underline{x}$ significant

$$\text{p-value}(\underline{x}) = \min_{\Gamma_c : \underline{x} \in \Gamma_c} \int_{\Gamma_c} f(\underline{x};\theta_0)\, d\underline{x}$$

Let $X_1^*, X_2^*, \ldots, X_n^* \sim F_{\theta_0}$

If $S(\underline{X})$ is s.t. the larger $S(\underline{x})$, the smaller the significance region, then:

$$\text{p-value}(\underline{x}) = \Pr\big(\underbrace{S(\underline{X}^*)}_{\text{random variable}} \geq \underbrace{S(\underline{x})}_{\text{observed}} ; \theta_0 \big)$$

There's usually a 1-to-1 correspondence between test-statistics and significance regions such that the larger the test-statistic, the smaller the significance region.

Example: HWE

observe $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$

$X_i \in \{0,1,2\}$ $\quad \theta = \{r_0, r_1\}$

$$\Pr(X_i = k) = r_k \qquad k = 0, 1, 2$$

Note: $r_2 = 1 - r_0 - r_1$

$$\dim(\Theta) = 2.$$

$$p = \frac{r_1}{2} + r_2$$

H_0: $X \sim \text{Binomial}(2, p)$

H_1: $X \sim \text{Multinomial}(1, (r_0, r_1, r_2))$

H_0: $\theta \in \overbrace{\{p^2, 2p(1-p)\}, \; p = \frac{r_1}{2} + r_2}^{\Theta_0}$

H_1: Not H_0

$$n_0 = \#\{X_i = 0\} \qquad n_0 + n_1 + n_2 = n$$
$$n_1 = \#\{X_i = 1\}$$
$$n_2 = \#\{X_i = 2\}$$

MLE $\hat{r}_k = \dfrac{n_k}{n} \qquad k = 0, 1, 2$

MLE with $\theta \in \Theta_0$: $\hat{p} = \dfrac{\frac{n_1}{2} + n_2}{n}$

$\hat{r}_0^0 = (1-\hat{p})^2$, $\hat{r}_1^0 = 2\hat{p}(1-\hat{p})$,

$\hat{r}_2^0 = \hat{p}^2$

$$L(\underline{r}; \underline{x}) \propto r_0^{n_0} r_1^{n_1} r_2^{n_2}$$

$$\lambda(\underline{x}) = \frac{\hat{r}_0^{n_0} \hat{r}_1^{n_1} \hat{r}_2^{n_2}}{(\hat{r}_0^0)^{n_0} (\hat{r}_1^0)^{n_1} (\hat{r}_2^0)^{n_2}}$$

$$= \frac{\hat{r}_0^{n_0} \hat{r}_1^{n_1} \hat{r}_2^{n_2}}{(1-\hat{p})^{2n_0} (2\hat{p}(1-\hat{p}))^{n_1} \hat{p}^{2n_2}}$$

Null distribution $2 \log \lambda(\underline{x}) \,\dot\sim\, \chi^2_\nu$

$$\nu = \dim(\Theta) - \dim(\Theta_0)$$

$$= 2 - 1 = 1$$

<u>Pivotal Statistic HT</u>

Suppose we have a pivotal statistic of the form $\frac{\hat{\theta} - \theta}{\hat{se}(\hat{\theta})} \,\dot\sim\, \text{Normal}(0,1)$.

Consider $H_0: \theta = \theta_0$ vs. $H_1: \theta \neq \theta_0$

known value

When $\theta = \theta_0$, $\dfrac{\hat{\theta} - \theta_0}{\hat{se}(\hat{\theta})} \overset{\cdot}{\sim} \text{Normal}(0,1)$

$$\frac{\hat{\theta} - \theta_0}{se(\theta_0)} \overset{\cdot}{\sim} \text{Normal}(0,1)$$

Example: $X_1, X_2, \ldots, X_n \sim \text{Poisson}(\lambda)$

$H_0: \lambda = 5 \qquad H_1: \lambda \neq 5$

In general $\dfrac{\hat{\lambda} - \lambda}{\sqrt{\hat{\lambda}/n}} \overset{\cdot}{\sim} \text{Normal}(0,1)$

$\hat{\lambda} = \bar{X}$.

When $\lambda = 5$, $\dfrac{\hat{\lambda} - 5}{\sqrt{\hat{\lambda}/n}} \sim \text{Normal}(0,1)$

$$\frac{\hat{\lambda} - 5}{\sqrt{5/n}} \overset{\cdot}{\sim} \text{Normal}(0,1)$$

Statistic : $|z| = \left| \dfrac{\hat{\theta} - \theta_0}{\hat{se}(\hat{\theta})} \right|$ $\Rightarrow$ larger is more evidence against H_0 in favor of H_1

observed statistic

Type I error rate:

Suppose c is our significance cut-off, $|z| \geq c \Rightarrow$ significant

Let $Z^* \sim \text{Normal}(0,1)$;

TIE rate is $\Pr(|Z^*| \geq c)$.

P-value :

$$\Pr(|Z^*| \geq |z|)$$

observed

rv ~ null dist'n

Implied significance regions:

$$\Gamma_c = \{ z : |z| \geq c \}$$

$c' \geq c \Rightarrow \Gamma_c \subseteq \Gamma_{c'}$.

Requirements in general:

1. A statistic that quantifies evidence against H_0 in favor H_1
2. The (approximate) distribution of the statistic when H_0 is true.

Requirement 1 is assessed in terms of power.

Distribution of a p-value

If we have the true null distribution identified, then the p-value is Uniform$(0,1)$ when H_0 is true.

When H_0 is true, $Pr_{H_0}(P \leq \alpha) = \alpha$. TIE rate

P ~ p-value

Conservative: $Pr_{H_0}(P \le \alpha) \le \alpha$.

Suppose H_1 is true, we want:

$$Pr_{H_1}(P \le \alpha) \gg \alpha.$$

power

One-sided test

$H_0: \theta \le \theta_0$ vs. $H_1: \theta > \theta_0$

$$z = \frac{\hat{\theta} - \theta_0}{\hat{se}(\theta)}$$

$$\Gamma_c = \{ z : z > c \}$$

Least favorable is when $\theta = \theta_0$.

When $\theta = \theta_0 \Rightarrow$ null dist'n approximately Normal $(0,1)$.

Let $Z^* \sim$ Normal $(0,1)$

TIE rate: $Pr(Z^* \ge c)$

P-value: $Pr(Z^* \ge z)$

observed

Two Sample Pivotal Statistic

If X and Y are independent rv's, then

$$E[X-Y] = E[X] - E[Y]$$

$$Var(X-Y) = Var(X) + Var(Y)$$

Suppose $X_1, X_2, \dots, X_n \overset{iid}{\sim} F_\theta$ and $Y_1, Y_2, \dots, Y_m \overset{iid}{\sim} F_\gamma$, and $\underline{X}$ and $\underline{Y}$ are independent.

MLEs $\hat{\theta}_n$ and $\hat{\gamma}_m$, $\hat{se}(\hat{\theta}_n)$ and $\hat{se}(\hat{\gamma}_m)$.

If W and V are Normal random variables, then $W+V \sim$ Normal.

Goal is to do inference on $\theta - \gamma$.

E.g., $H_0: \underbrace{\theta - \gamma = 0}_{\theta = \gamma}$ vs. $H_1: \underbrace{\theta - \gamma \neq 0}_{\theta \neq \gamma}$

$$\frac{\hat{\theta}_n - \hat{\gamma}_m - (\theta - \gamma)}{\sqrt{\hat{se}(\hat{\theta}_n)^2 + \hat{se}(\hat{\gamma}_m)^2}} \overset{\cdot}{\sim} \text{Normal}(0,1).$$

Pivotal statistics exist in many settings outside of MLEs, and the above ideas continue to hold.

Bayesian Inference

A prior distribution is assumed for the parameter(s) of interest.

Example: $P \sim \text{Uniform}(0,1)$

Data generating istribution $X \mid P=p \sim \text{Bin}(n,p)$.

pdf $\to$ $$f(p \mid X=x) = \frac{\Pr(X=x \mid P=p)\, f(p)}{\Pr(X=x)}$$

$f(p) = 1$ (Uniform(0,1)) is is prior dist'n.

$f(p \mid X=x)$ is the posterior dist'n.

$\Pr(X=x) = \int \Pr(X=x \mid P=p^*)\, f(p^*)\, dp^*$.

General set up:

$(X_1, X_2, \ldots, X_n) \mid \theta \overset{iid}{\sim} F_\theta$ and θ is drawn from a prior distribution $\theta \sim F_\tau$.

Essential task: Calculate the posterior distribution of $\theta | \underline{X} = \underline{x}$.

$$f(\theta | \underline{X}) = \frac{f(\underline{X} | \theta) f(\theta)}{f(\underline{X})}$$

$$= \frac{f(\underline{X} | \theta) f(\theta)}{\int f(\underline{X} | \theta^*) f(\theta^*) \, d\theta^*}$$

If there is indeed a "true" θ, then if certain regularity conditions are met then $f(\theta | \underline{X})$ concentrates around the true θ as $n \to \infty$.

Point Estimation:

For example, we might estimate θ with $E_{\theta | \underline{x}}[\theta | \underline{X} = \underline{x}] = \int \theta \, f(\theta | \underline{X} = \underline{x}) \, d\theta$.