Chapter 5

Two-Samples

Ⓐ Bernoulli

Suppose we have two sets of coin flips.

$$Y_1, \dots, Y_{n_y} \overset{iid}{\sim} Bern(\theta_y) \quad X_1, \dots, X_{n_x} \overset{iid}{\sim} Bern(\theta_x)$$

& we want to know if $\theta_y = \theta_x$?

$H_0: \theta_y = \theta_x$

$H_1: \theta_y \neq \theta_x$

If H_0 is true $\theta \equiv \theta_y = \theta_x$ then MLE

$$\hat{\theta}_{pool} = \left(\sum_{i=1}^{n_y} y_i + \sum_{j=1}^{n} x_j \right) / (n_y + n_x)$$

$$= \frac{S_y + S_x}{n_y + n_x}$$

A natural stat for testing

$$\hat{\theta}_\Delta = \hat{\theta}_y - \hat{\theta}_x$$

if H_0 then $\theta_\Delta = \emptyset$.

$$\frac{s_y}{n_y} \quad \frac{s_x}{n_x}$$

In Chap 4 CLT/AMLE

$$\hat{\theta}_y \sim N(\theta_y, \theta_y(1-\theta_y)/n_y) \quad \text{as } n_y \to \infty$$

$$\hat{\theta}_x \sim N(\theta_x, \theta_x(1-\theta_x)/n_x) \quad n_x \to \infty$$

Using that sums of Gaussian are Gaussian

$$\hat{\theta}_\Delta = \hat{\theta}_y - \hat{\theta}_x \sim N\left(\theta_y - \theta_x, \frac{\theta_y(1-\theta_y)}{n_y} + \frac{\theta_x(1-\theta_x)}{n_x}\right)$$

as $n_y, n_x \to \infty$

w/ $\theta \equiv \theta_y = \theta_x$

Under H_0 $\theta_y = \theta_x$

so $$\hat{\theta}_\Delta \sim N\left(0, \theta(1-\theta) \times \left(\frac{1}{n_y} + \frac{1}{n_x}\right)\right)$$

WLLN to replace θ w/ an estimate

w/ $\hat{\theta}_{pool}$

$$\frac{\hat{\theta}_y - \hat{\theta}_x}{\sqrt{\hat{\theta}_{pool}(1-\hat{\theta}_{pool}) \times \left(1/n_y + 1/n_x\right)}} \sim N(0,1)$$

as $n_x, n_y \to \infty$

Chapter 5
Continued

---

⑤ Two-sample Gaussian

$Y_1, \ldots, Y_{n_y} \overset{iid}{\sim} N(\mu_y, \sigma_y^2)$

$X_1, \ldots, X_{n_x} \overset{iid}{\sim} N(\mu_x, \sigma_x^2)$

$H_0: \mu_y \neq \mu_x$

$H_1: \mu_y \neq \mu_x$

~~Sometimes $\sigma_y^2 = \sigma_x^2$ (pooled analysis)~~

A sensible test stat $\hat{\mu}_y - \hat{\mu}_x \equiv \hat{\mu}_\Delta$

$\bar{Y} - \bar{X}$

Welch (1938) showed

$$T \equiv \frac{\bar{Y} - \bar{X}}{\sqrt{s_y^2/n_y + s_x^2/n_x}} \sim t_v$$

where $v \approx \min(n_y, n_x) - 1$

↓

book

is a good approx

An important variation involves "paired" data.

$(x_1, y_1), \ldots, (x_n, y_n)$

We can reduce this 2-sample testing problem in a <u>one-sample</u> one

let $D_i = Y_i - X_i \overset{iid}{\sim} N(\mu_d, \sigma_d^2)$

$i = 1, \ldots, n$

$H_0: \mu_y = \mu_x \equiv \mu_d = \emptyset$

$H_1: \mu_d \neq \emptyset$

$\rightarrow$ 1-sample t-test on $D_1, \ldots, D_n$.