STAT 231 Online Spring 2020

Linxuntian

July 2020

Contents

1 Week 7

1.1 Test of Hypothesis Part I

Definition

A **test statistic** or **discrepancy measure** is a function of the data $D = \mathrm{g}(Y)$ that is constructed to measure the degree of "agreement" between the data Y and the null hypothesis H_0.

Summary (Steps of a Statistical Test of Hypothesis)

1. Assume that the null hypothesis H_0 will be tested using data Y
2. Adopt a test statistic or discrepancy measure $\mathrm{D}(Y)$ for which large value of D are less consistent with H_0. Let $d = \mathrm{D}(y)$ be the corresponding observed value of D
3. Calculate

$$\begin{aligned} p\text{-value} &= \mathrm{P}(D \geqslant d; \text{ assuming } \mathrm{H}_0 \text{ is true}) \\ &= \mathrm{P}(D \geqslant d; \mathrm{H}_0) \end{aligned}$$

4. Draw a conclusion based on the p-value

1.2 Test of Hypothesis Part II

Summary (Hypothesis Test for Binomial Model)

1. Test $\mathrm{H}_0 : \theta = \theta_0$ using $Y \sim \mathrm{Bi}(n, \theta)$
2. Test statistic $\mathrm{D}(Y) = |Y - n\theta_0|$. Let $d = |y - n\theta_0|$ be the observed value of D
3. Calculate

$$\begin{aligned} p\text{-value} &= \mathrm{P}(D \geqslant d; \theta = \theta_0) \\ &= \mathrm{P}(|Y - n\theta_0| \geqslant d) \end{aligned}$$

   where $Y \sim \mathrm{Bi}(n, \theta)$

   If n is large, we can approximate

$$\begin{aligned} p\text{-value} &\approx \mathrm{P}\left(|Z| \geqslant \frac{d}{\sqrt{n\theta_0(1-\theta_0)}}\right) \quad \text{where } Z \sim \mathrm{N}(0,1) \\ &= 2\left[1 - \mathrm{P}\left(Z \leqslant \frac{d}{\sqrt{n\theta_0(1-\theta_0)}}\right)\right] \end{aligned}$$

4. Draw a conclusion

Summary (Hypothesis Test for Gaussian Model)

When the **variance** is unknown

1. 123

When the **mean** is unknown

2 Week 8

Summary (Likelihood Ratio Test of Hypothesis)

1. To test H_0: $\theta = \theta_0$
2. Likelihood ratio test statistic $\Lambda(\theta_0) = -2\log\left[\frac{L(\theta_0)}{L(\tilde{\theta})}\right]$,

   with the observed value $\lambda(\theta_0) = -2\log\left[\frac{L(\theta_0)}{L(\hat{\theta})}\right]$
3. Calculate

$$\begin{aligned} p\text{-value} &\approx P(W \leqslant \lambda(\theta_0)) \quad \text{where } W \sim \chi^2(1) \\ &= 2[1 - P(Z \leqslant \sqrt{\lambda(\theta_0)})] \quad \text{where } Z \sim G(0,1) \end{aligned}$$

4. Draw a conclusion

3 Week 9

3.1 Gaussian Response Models Part I

Confidence Interval for beta β

A $100p\%$ Confidence Interval for β is given by

$$\hat{\beta} \pm a \frac{s_e}{\sqrt{S_{xx}}}$$

where $\mathrm{P}(T \leqslant a) = \dfrac{(1+p)}{2}$ and $T \sim t(n-2)$

3.2 Gaussian Response Models Part II

Summary (Hypothesis testing of No Relationship for beta β)

1. Test $\mathrm{H}_0 : \beta = 0$
2. Test statistic $D = \dfrac{|\tilde{\beta} - 0|}{S_e/\sqrt{S_{xx}}}$, with observed value $d = \dfrac{|\hat{\beta} - 0|}{s_e/\sqrt{S_{xx}}}$
3. Calculate

$$\begin{aligned} p\text{-value} &= \mathrm{P}\left(|T| \geqslant \frac{|\hat{\beta} - 0|}{s_e/\sqrt{S_{xx}}}\right) \\ &= 2\left[1 - \mathrm{P}\left(T \leqslant \frac{|\hat{\beta} - 0|}{s_e/\sqrt{S_{xx}}}\right)\right] \end{aligned}$$

   where $T \sim t(n-2)$
4. Draw Conclusion

Confidence Interval for Mean $\mu(x) = \alpha + \beta x$

A $100p\%$ Confidence Interval for $\mu(x) = \alpha + \beta x$ is given by

$$\begin{aligned} &\hat{\mu}(x) \pm a s_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \\ &= \hat{\alpha} + \hat{\beta} x \pm a s_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \end{aligned}$$

where $\mathrm{P}(T \leqslant a) = \dfrac{(1+p)}{2}$ and $T \sim t(n-2)$

4 Week 10

4.1 Gaussian Response Models Part III

Prediction Interval For Future Response Y

A $100p\%$ Prediction Interval for a future response Y is given by

$$\begin{aligned} &\hat{\mu}(x) \pm a s_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \\ &= \hat{\alpha} + \hat{\beta} x \pm a s_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \end{aligned}$$

where $\mathrm{P}(T \leqslant a) = \dfrac{(1+p)}{2}$ and $T \sim t(n-2)$

Definition

For the simple linear regression model, let

$$\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i$$

and let

$$\hat{r}_i = y_i - \hat{\mu}_i \quad i = 1, 2, \ldots, n$$

The $\hat{r}_i$'s are called **residuals** since $\hat{r}_i$ represents what is "left" after the model has been "fitted" to the data

Definition

Define the **standardized residuals** as

$$\hat{r}_i = \frac{\hat{r}_i}{s_e} = \frac{y_i - \hat{\mu}_i}{s_e} \quad i = 1, 2, \ldots, n$$

4.2 Comparing Means of Two Populations

Definition

Pooled estimate of variance is defined as

$$\begin{aligned} s_p^2 &= \frac{1}{n_1 + n_2 - 2} \left[\sum_{i=1}^{n_1} (y_{1i} - \overline{y_1})^2 + \sum_{i=1}^{n_2} (y_{2i} - \overline{y_2})^2 \right] \\ &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \end{aligned}$$

C.I. for Difference in Means $\mu_1 - \mu_2$

A $100p\%$ C.I. for $\mu_1 - \mu_2$ is given by

$$\overline{y_1} - \overline{y_2} \pm a s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $\mathrm{P}(T \leqslant a) = \dfrac{(1+p)}{2}$ and $T \sim t(n_1 + n_2 - 2)$

Summary (Hypothesis testing of Difference in Means $\mu_1 - \mu_2$)

1. Test $\mathrm{H}_0 : \mu_1 - \mu_2 = 0$
2. Test statistic $D = \dfrac{|\overline{Y_1} - \overline{Y_2} - 0|}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ and observed value $d = \dfrac{|\overline{y_1} - \overline{y_2} - 0|}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
3. Calculate

$$\begin{aligned} p\text{-value} &= \mathrm{P}(|T| \geqslant d) \\ &= 2[1 - \mathrm{P}(T \leqslant d)] \end{aligned}$$

   where $T \sim t(n_1 + n_2 - 2)$
4. Draw a conclusion

C.I. for σ

A $100p\%$ Confidence Interval for σ is given by

5 Week 11

5.1 Gaussian Response Models Part IV

Summary (Hypothesis testing of Difference in Means $\mu = \mu_1 - \mu_2$ within-pair)

1. Test H_0: $\mu = 0$ using $Y_i = Y_{1i} - Y_{2i} \sim \mathrm{G}(\mu_1 - \mu_2, \sigma)$, $i = 1, 2, \ldots, n$, independently
2. Test statistic $D = \dfrac{|\overline{Y} - 0|}{S/\sqrt{n}}$, with observed value $d = \dfrac{|\overline{y} - 0|}{s/\sqrt{n}}$
3. Calculate
$$p\text{-value} = 2[1 - \mathrm{P}(T \leqslant d)]$$
where $T \sim t(n-1)$
4. Draw a conclusion

5.2 Multinomial Models and Goodness of Fit

Summary (Hypothesis testing of θ)

1. Test H_0: $\theta = \theta_0 = \left(\dfrac{1}{k}, \dfrac{1}{k}, \ldots, \dfrac{1}{k}\right)$ using $\mathrm{f}(y_1, y_2, \ldots, y_k; \theta_1, \theta_2, \ldots, \theta_k)$
2. Likelihood ratio test statistic $\Lambda = 2\displaystyle\sum_{j=1}^{k} Y_i \log\left(\frac{Y_j}{E_j}\right)$

   and observed value $\lambda = 2\displaystyle\sum_{j=1}^{k} y_i \log\left(\frac{y_j}{e_j}\right)$
3. Calculate
$$p\text{-value} = \mathrm{P}(\Lambda \geqslant \lambda; \mathrm{H}_0) \approx \mathrm{P}(W \geqslant \lambda)$$
where $W \sim \chi^2(k - 1 - p)$
4. Draw a conclusion

6 Week 12

6.1 Two-Way Tables and Tests of Hypothesis

Definition

An example of **Two-way table**...

Let θ_{ij} be the probability a randomly selected individual is combined type (A_i, B_j) and note that $\sum_{i=1}^{a}\sum_{j=1}^{b}\theta_{ij} = 1$. The $a \times b$ frequencies $(Y_{11}, Y_{12}, \ldots, Y_{ab})$ follow a Multinomial distribution with $k = ab$ classes.

Summary (Hypothesis testing of Dependence of A and B class)

1. Test H_0: $\theta_{ij} = \alpha_i \beta_j$

   Note that the expected frequencies under the the hypothesis H_0 are given by $e_{ij} = \dfrac{r_i c_j}{n}$

2. Likelihood ratio statistic $\Lambda = 2\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij}\log\left(\dfrac{Y_{ij}}{E_{ij}}\right)$,

   with observed value $\lambda = 2\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}\log\left(\dfrac{y_{ij}}{e_{ij}}\right)$

3. Calculate

   $$p\text{-value} \approx \mathrm{P}(W \geqslant \lambda)$$

   where $W \sim \chi^2((a-1)\times(b-1))$

   Note:

   (a) If $((a-1)\times(b-1)) = 1$, then

   $$p\text{-value} = 2[1 - \mathrm{P}(Z \leqslant \sqrt{\lambda})], \quad \text{where } Z \sim \mathrm{G}(0,1)$$

   (b) If $((a-1)\times(b-1)) = 2$, then

   $$p\text{-value} = \mathrm{P}(W \geqslant \lambda) = \mathrm{e}^{-\lambda/2}, \quad \text{where } W \sim \chi^2(2) = \text{Exponential}(2)$$

4. Draw a conclusion

6.2 Cause and Effect

Definition

Let be y a response variate and let be x an explanatory variate associated with units in a population or process. Then, if all other factors that affect are held constant, let us change x (or observe different values of x) and see if y changes. If it does, we say that x has a **causal effect** on y.

Definition

x has a **causal effect** on Y if, when all other factors that affect are held constant, a change in x (or observing different values of x) induces a change in a property of the distribution of Y.