STAT 111

Recitation 6

Mo Huang

Email: mohuang@wharton.upenn.edu

Office Hours: Wednesdays 3:00 - 4:00 pm, JMHH F96

Slides: `github.com/mohuangx/STAT111-Spring2019`

March 15, 2019

Estimation of a binomial parameter θ

- Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Bin(1, \theta)$, and you want to estimate θ.
- The **estimate** for θ is $\mathrm{p} = \mathrm{x}/\mathrm{n}$.
- The **estimator** is P.
- 95% confidence interval is $p \pm \sqrt{1/n}$.

Estimation of a mean μ

- Suppose we have i.i.d. random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2. We do not assume any distribution!
- We observe $x_1, \ldots, x_n$ and we want to estimate μ. How do we do this?
- Recall by the Central Limit Theorem, for large n, $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
- Thus, $\bar{x}$ is an unbiased **estimate** for μ.

Estimation of a mean μ

- How precise an estimate is $\bar{x}$?

- This is where the variance comes in:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

- We can use the two-standard deviation rule:

$$\text{Prob}\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

- Then our 95% confidence interval for our data is:

$$\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}.$$

Note: $(1-\alpha)\%$ confidence interval is $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ where z^* is the value in the z-chart where the probability is $1 - \frac{\alpha}{2}$.

Estimation of a mean μ

- What if we don't know the true variance σ^2?
- We need to estimate the variance from the data.
- The unbiased estimate of σ^2 is s^2:

$$s^2 = \frac{x_1^2 + x_2^2 + \cdots + x_n^2 - n(\bar{x}^2)}{n-1}.$$

- Then, our approximate 95% confidence interval is:

$$\bar{x} - 2\frac{s}{\sqrt{n}} \quad \text{to} \quad \bar{x} + 2\frac{s}{\sqrt{n}}.$$

Note: 95% confidence intervals are always of the form:

$$Estimator - 2 * SD(Estimator) \quad \text{to} \quad Estimator + 2 * SD(Estimator)$$

Example

- Suppose we observe 15 iid data points:

  $104, 127, 153, 164, 115, 143, 193, 151, 129, 139, 122, 144, 108, 148, 132.$

Q: Find the sample average of the data.

A: $\bar{x} = (104 + 127 + \cdots + 132)/15 = 138.13$

Q: Find the sample standard deviation of the data.

A:

$$s^2 = \frac{104^2 + 127^2 + \cdots + 132^2 - 15 * 138.13^2}{15 - 1} \Rightarrow s = 22.89$$

Q: Find the 95% confidence interval for the mean.

A:

$$138.13 - 2 * \frac{22.89}{\sqrt{15}} \quad \text{to} \quad 138.13 + 2 * \frac{22.89}{\sqrt{15}} = [126.31, 149.95]$$

Estimating the difference between proportions $\theta_1 - \theta_2$

- Let Population 1 have proportion P_1 and sample size n, and Population 2, P_2 and m respectively. Then for large n, m, we have

$$P_1 \sim N\left(\theta_1, \frac{\theta_1(1-\theta_1)}{n}\right), \quad P_2 \sim N\left(\theta_2, \frac{\theta_2(1-\theta_2)}{m}\right) \quad (CLT)$$

- Let $D = P_1 - P_2$. Then D is normal with

$$D \sim N\left(\theta_1 - \theta_2, \frac{\theta_1(1-\theta_1)}{n} + \frac{\theta_2(1-\theta_2)}{m}\right).$$

- The **estimate** for $\theta_1 - \theta_2$ is $p_1 - p_2$.
- The **95% confidence interval** is then:

$$p_1 - p_2 \pm \sqrt{\frac{1}{n} + \frac{1}{m}}$$

Example

- Suppose we have two drugs to cure a headache, 1 and 2. Let θ_1 be the probability that Drug 1 cures a headache and θ_2 is the probability that Drug 2 cures a headache.
- In a group of 250 people, Drug 1 cures 189 people. In a different group of 300 people, Drug 2 cures 256 people.
- Find the 95% confidence interval for the difference $\theta_1 - \theta_2$.

A:

$$p_1 - p_2 = \frac{189}{250} - \frac{256}{300} = -0.0973, \quad \sqrt{\frac{1}{250} + \frac{1}{300}} = 0.0856$$

$\Rightarrow$ 95% confidence interval:

$$-0.0973 \pm 0.0856 \Rightarrow [-0.1829, -0.0117].$$

Estimating the difference between means $\mu_1 - \mu_2$

- Suppose we have two independent populations:
  - $X_{11}, X_{12}, \ldots, X_{1n}$: i.i.d. random variables with mean μ_1 and variance σ_1^2 (both unknown),
  - $X_{21}, X_{22}, \ldots, X_{2m}$: i.i.d. random variables with mean μ_2 and variance σ_2^2 (both unknown).
- By the CLT,

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

- The **estimate** for $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$.
- The **95% confidence interval** is then

$$\bar{x}_1 - \bar{x}_2 \pm 2\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}$$

where $s_1^2 = \dfrac{x_{11}^2 + \cdots + x_{1n}^2 - n(\bar{x}_1^2)}{n-1}, \quad s_2^2 = \dfrac{x_{21}^2 + \cdots + x_{2m}^2 - m(\bar{x}_2^2)}{m-1}$

Example

- Suppose we want to estimate the difference in yield of wheat from two different, independent fields, A and B.

- Data:

Field A: $n = 12, \bar{x}_1 = 121.4, s_1^2 = 10.1$

Field B: $m = 15, \bar{x}_2 = 113.8, s_2^2 = 12.1$

- Find the 95% confidence interval for the difference in mean yield $(\mu_1 - \mu_2)$.

A:

$$121.4 - 113.8 \pm 2\sqrt{\frac{10.1}{12} + \frac{12.1}{15}} \Rightarrow [5.032, 10.168]$$

Estimation Summary

- Binomial parameter θ:

  Estimate: p

  95% confidence interval: $p \pm \sqrt{1/n}$

- Mean μ:

  Estimate: $\bar{x}$

  95% confidence interval: $\bar{x} \pm 2\frac{s}{\sqrt{n}}$

- Difference between proportions $\theta_1 - \theta_2$:

  Estimate: $p_1 - p_2$

  95% confidence interval: $p_1 - p_2 \pm \sqrt{\frac{1}{n} + \frac{1}{m}}$

- Difference between means $\mu_1 - \mu_2$:

  Estimate: $\bar{x}_1 - \bar{x}_2$

  95% confidence interval: $\bar{x}_1 - \bar{x}_2 \pm 2\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}$

Note:

$$s = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2 - n(\bar{x}^2)}{n-1}}.$$