Lecture 8: Sampling Distributions

STATS 101: Foundations of Statistics

Linh Tran

linh@thetahat.ai

January 22, 2019

Announcements

- No more assignments
  - Questions will still be posted (and reviewed)
- A *Colab* script is available for today's class.

Outline

Sampling distributions

- Review
- Efficiency
- χ^2 and t distribution's
- Confidence intervals

Recall

Given $X_n, \ldots, X_n \overset{iid}{P}_0$, we can form an estimator

$$\hat{\theta}_n = \omega(X_1, \ldots, X_n) \tag{1}$$

of some underlying parameter on P_0.

The parameter estimates $\hat{\theta}$ are random and therefore have a *sampling distribution*

Colab link

Evaluating estimators

How to evaluate estimators:

- Mean squared error, i.e.

$$MSE(\hat{\theta}_n) = \mathbb{E}_0[\left(\theta_0 - \hat{\theta}\right)^2] \tag{2}$$

- The estimator's bias, i.e.

$$Bias(\hat{\theta}_n) = \mathbb{E}_0[\theta_0 - \hat{\theta}] \tag{3}$$

- The estimator's variance, i.e.

$$var(\hat{\theta}_n) = \mathbb{E}_0\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right] \tag{4}$$

Colab link

Efficiency

We are typically interested in estimators having low MSE. This can be defined in as efficiency, i.e.

Efficiency

An estimator $\hat{\theta}_n$ is *efficient* relative to W if

$$MSE(\hat{\theta}_n) \leq MSE(w) \forall \theta \in \Theta, \forall w \in W \tag{5}$$

Efficiency

We are typically interested in estimators having low MSE. This can be defined in as efficiency, i.e.

Efficiency

An estimator $\hat{\theta}_n$ is *efficient* relative to W if

$$MSE(\hat{\theta}_n) \leq MSE(w) \, \forall \theta \in \Theta, \forall w \in W \tag{5}$$

Problem: Unless we restrict W in some way, we can often find many estimators with equal MSE by trading off bias for variance.

Efficiency

Question: Is there a lower bound that we can aim for in terms of variance?

Efficiency

Question: Is there a lower bound that we can aim for in terms of variance?

Answer: Yes!

The Cramer-Rao Lower Bound (CRLB)

Under some regularity conditions (i.e. finite variance and differentiation/integration interchangability), we have that

$$var(\hat{\theta}_n) \geq \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}\left[\hat{\theta}_n\right]\right)^2}{n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]} \tag{6}$$

Efficiency

n.b. The CRLB has three possible cases:

- The CRLB is applicable and attainable, e.g.
  - Estimating p when $X_i \sim Ber(p)$
- The CRLB is applicable, but not attainable, e.g.
  - Estimating $\hat{\sigma}^2 = s^2$ when $X_i \sim N(\mu, \sigma^2)$.

$$var(\hat{\sigma}^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n} \tag{7}$$

    the latter of which is the CRLB
- The CRLB is not applicable, e.g.
  - Estimating θ when $X_i \sim U(0, \theta) : \theta < \infty$

$$var(\hat{\theta}) = \frac{1}{n(n+2)\theta^2} \tag{8}$$

    while the CRLB is $\frac{\theta^2}{n}$

Efficiency

Some notes:

- If $\hat{\theta}_n$ is unbiased then we have that $\mathbb{E}\left[\hat{\theta}_n\right] = \theta$ and, consequently,

$$var(\hat{\theta}_n) \geq \frac{1}{n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log\ f(X|\theta)\right)^2\right]} \tag{9}$$

- We commonly refer to $I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log\ f(X|\theta)\right)^2\right]$ as the Fisher information (of a single observation)
- We commonly refer to $I_n(\theta) = n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log\ f(X|\theta)\right)^2\right]$ as the Fisher information (of a random sample)
  - Our lower bound is therefore

$$var(\hat{\theta}_n) \geq \frac{1}{nI(\theta)} = \frac{1}{I_n(\theta)} \tag{10}$$

The normal distribution

Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu_0, \sigma_0^2)$. The MLE for μ_0, σ_0^2 are

$$\begin{aligned} \hat{\mu}_n &= \bar{X}_n & (11) \\ \hat{\sigma}_n^2 &= (1/n) \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 & (12) \end{aligned}$$

The normal distribution

Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu_0, \sigma_0^2)$. The MLE for μ_0, σ_0^2 are

$$\hat{\mu}_n = \bar{X}_n \tag{11}$$

$$\hat{\sigma}_n^2 = (1/n) \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \tag{12}$$

Remark: $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ are independent of each other!
Colab link

The normal distribution

The sampling distributions for $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ are

$$\begin{aligned} \hat{\mu}_n &\sim N(\mu_0, \sigma^2/n) \\ \hat{\sigma}_n^2 &\sim \chi^2(n-1) \end{aligned}$$

The normal distribution

The sampling distributions for $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ are

$$\begin{aligned} \hat{\mu}_n &\sim N(\mu_0, \sigma^2/n) \\ \hat{\sigma}_n^2 &\sim \chi^2(n-1) \end{aligned}$$

Some notes:

- The χ^2 distribution with n degrees of freedom is the gamma distribution with $\alpha = n/2$ and $\beta = 1/2$.
- If μ_0 is known, we instead have $\hat{\sigma}_n^2 \sim \chi^2(n)$.
- The χ^2 distribution is commonly thought of as the standard normal distribution squared (i.e. if $X \sim N(0,1)$, then $Y = X^2 \sim \chi^2(1)$)

The t-distribution

Widely used as test statistics. Let $Z \sim N(0,1)$ and $Y \sim \chi^2(n)$. Then

$$X = \frac{Z}{(Y/n)^{1/2}} \sim t(n) \tag{13}$$

follows a *t-distribution* with n degrees of freedom.

The t-distribution

Widely used as test statistics. Let $Z \sim N(0,1)$ and $Y \sim \chi^2(n)$. Then

$$X = \frac{Z}{(Y/n)^{1/2}} \sim t(n) \tag{13}$$

follows a *t-distribution* with n degrees of freedom.

Example: Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_0, \sigma_0^2)$. If

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{14}$$

$$\sigma_n^{'} = \left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right)^{1/2} \tag{15}$$

then $n^{1/2}(\bar{X}_n - \mu)/\sigma_n^{'} \sim t(n-1)$.
Colab link

Confidence intervals

Rather than coming up with a single estimate $\hat{\theta}$, we could instead come up with a range that we think contains θ (with high probability). e.g.

$$95\%\ CI = (-c_l, c_u) : P(-c_l \leq \theta \leq c_u) = 0.95 \tag{16}$$

Question: Why do this?

Confidence intervals

Rather than coming up with a single estimate $\hat{\theta}$, we could instead come up with a range that we think contains θ (with high probability). e.g.

$$95\%\ CI = (-c_l, c_u) : P(-c_l \leq \theta \leq c_u) = 0.95 \tag{16}$$

Question: Why do this?

Note: Most confidence intervals assume a (usually symmetric) distribution (e.g. normal) and apply a pivot against the estimate, e.g.

$$95\%\ CI = \hat{\theta} \pm z_{\alpha/2} * se(\hat{\theta}) \tag{17}$$

Confidence intervals

Example: $X_1, ..., X_n \sim N(\mu, \sigma^2 = 1)$.

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{18}$$

$$se(\hat{\mu}_n) = \sqrt{\sigma^2/n} = 1/\sqrt{n} \tag{19}$$

Our confidence interval is therefore

$$95\%\ CI = \hat{\theta} \pm z_{\alpha/2} * se(\hat{\theta}) \tag{20}$$

$$= \hat{\theta} \pm \frac{z_{\alpha/2}}{\sqrt{n}} \tag{21}$$

Colab link

References

- DeGroot & Schervish Chapters 8.1-8.5, 8.8