2020-02-05

Homework Review

Some common numerical estimands

- $Y_1, \ldots, Y_n$ iid from dist. with finite variance σ^2 & mean μ

Then for the estimator $\bar{Y}$ of μ:

$$\mathbb{E}\,\bar{Y} = \mathbb{E}\left[\frac{1}{n}\sum Y_i\right] = \frac{1}{n}\sum \mathbb{E}\,Y_i = \frac{1}{n}\sum \mu = \frac{1}{n}\cdot n\cdot\mu = \mu$$

$$\mathbb{V}\,\bar{Y} = \mathbb{V}\left[\frac{1}{n}\sum Y_i\right] = \frac{1}{n^2}\left(\sum \mathbb{V}\,Y_i + \text{(covariances)}\right) = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{\sigma^2}{n}$$

(covariances: Y_i independent! $\Rightarrow$ $\text{cov} = 0$)

Central Limit Theorem:

$\bar{Y} \sim N(\mu, \sigma^2)$ approx.

- $Y \sim \text{Binomial}(n, p)$

$\hat{p} = Y/n$ estimates p.

$$\mathbb{E}\,\hat{p} = \mathbb{E}\left[\frac{Y}{n}\right] = \frac{1}{n}\,\mathbb{E}\,Y = \frac{1}{n}\cdot np = p$$

($\mathbb{E}\,Y = np$ because Y is binomial)

$$\mathbb{V}\,\hat{p} = \mathbb{V}\left[\frac{Y}{n}\right] = \frac{1}{n^2}\,\mathbb{V}\,Y = \frac{1}{n^2}\cdot npq = pq/n.$$

Sidebar:

$Y \sim \text{Bernoulli}(p)$ then

$\mathbb{E}\,Y = 1\cdot p + 0\cdot(1-p) = p.$

$\mathbb{V}\,Y = \mathbb{E}[(Y-p)^2] = p\cdot(1-p)^2 + (1-p)\cdot p^2$

$= pq(p+q) = pq.$

So if $Y \sim \text{Binomial}(n,p)$, then for $Y_1, \ldots, Y_n$ iid Bernoulli, $Y = \sum Y_i$.

$\mathbb{E}\,Y = np$ follow immediately

$\mathbb{V}\,Y = npq$

- $X_1, \ldots, X_{n_1}$ iid mean μ_1 variance σ_1^2 ⎫ independent
  $Y_1, \ldots, Y_{n_2}$ iid mean μ_2 variance σ_2^2 ⎭ of each other

$\bar{X} - \bar{Y}$ estimates $\mu_1 - \mu_2$.

$\mathbb{E}[\bar{X} - \bar{Y}] = \mathbb{E}\bar{X} - \mathbb{E}\bar{Y} = \mu_1 - \mu_2$.

$\mathbb{V}[\bar{X} - \bar{Y}] = \mathbb{V}\bar{X} + \mathbb{V}\bar{Y} + \text{covariances}$ ← independence ⇒ 0 $= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

- We could note that both the $\hat{p}$ and the $\hat{p}_1 - \hat{p}_2$ cases reduce to the means if we just express each binomial as a sum of Bernoulli r.v.s.

The standard deviation of the estimate is of critical importance in reporting any inferences.

This is called the <u>standard error</u>

CLT ⇒ all four have approx. normal sampl. distr. often for as small samples as $n=5$; definitely when $n=30$

For binomial: $n > 9 \cdot \frac{\max(p,q)}{\min(p,q)}$

Since the estimator $\hat{\vartheta}$ is a random variable so is the <u>estimation error</u> $\varepsilon = |\hat{\vartheta} - \vartheta|$

For approximately normal estimators, we can say a lot about this error!
Chebyshev's theorem lets us say quite a bit regardless of sampling distribution.

Pick points $\vartheta - b$ and $\vartheta + b$ in the tails of the sampling distribution of $\hat{\vartheta}$. We can quantify

$$P(\varepsilon < b)$$

Easy way to work with this is to express b in terms of the standard error $\sigma_{\hat{\vartheta}}$.

To ensure $P(\varepsilon < k\sigma_{\hat{\vartheta}}) > 0.90$, say, we just need to solve for k in the equality

$$\int_{\vartheta - k\sigma_{\hat{\vartheta}}}^{\vartheta + k\sigma_{\hat{\vartheta}}} p(\hat{\vartheta})\, d\hat{\vartheta} = 0.90$$

for p the density function for $\hat{\vartheta}$.

This is relatively easy when we deal with approx. normal distributions. There for $k = 2$:

$$\int_{\vartheta - 2\sigma_{\hat{\vartheta}}}^{\vartheta + 2\sigma_{\hat{\vartheta}}} p(\hat{\vartheta})\, d\hat{\vartheta} \approx 0.96$$

In general, Chebyshev's theorem gives us a (very!) loose bound:

Theorem $y_1, \dots, y_n$ iid. Let s be sample std. dev..

A fraction of at least $1 - \frac{1}{k^2}$ is contained in $(\bar{y} - k \cdot s, \bar{y} + k \cdot s)$.

Proof $(n-1)s^2 = \sum |y_i - \bar{y}|^2 \leq \sum \min(|y_i - \bar{y}|, k \cdot s)^2$.

Divide by $k^2 s^2$:

$$\frac{n-1}{k^2} \leq \sum \min\left(\frac{|y_i - \bar{y}|}{ks}, 1\right)^2$$

If more than $\frac{n-1}{k^2}$ entries are out of range, inequality breaks. So at most $\frac{n-1}{k^2}$ out of n entries are out of range. Hence

$n - \frac{n-1}{k^2}$ entries are in range.

$$n - \frac{n-1}{k^2} \geq n - \frac{n}{k^2} = n\left(1 - \frac{1}{k^2}\right). \quad \blacksquare$$

Corollary $\mathbb{P}\left(\varepsilon \leq k \sigma_{\hat{y}}\right) \geq 1 - \frac{1}{k^2}$.

So for $k=2$, at least 0.75 of probability mass is in the range for any distribution

Example 8.2.

8.3.

8.34, 8.36, 8.37,

Exercises 8.23 Bring up and use RStudio

8.26

8.29

8.32

8.33