Section 4.2 (con't.)

01/17/2019

1 Large sample confidence interval for p

Example 4.2.3

Suppose $n = 40$ graduating students are asked if they plan to go to graduate school. If 8 out of 40 said *yes*, then $\hat{p} = 8/40 = .20$, or 20%.

Question: What is the expected size of the error of estimation? (SE=?, 95% CI=?)

Math trick: If we can represent $\hat{p}$ as a sample mean, then all results already known about the sample mean apply.

Ex. Consider the binary sample: S, F, F, S, F

S	$\to$	1
F	$\to$	0
F	$\to$	0
S	$\to$	1
F	$\to$	0
$\hat{p} = 2/5 = .40$		$\overline{X} = 2/5 = .40$

"The sample proportion $\hat{p}$ is a sample mean of 0s and 1s"

i.e. $\hat{p} = \frac{\sum X_i}{n}$ or $\sum X_i = n\hat{p}$.

Furthermore, the sample variance is

$$\begin{aligned}
S^2 &= \frac{\sum (X_i - \overline{X})^2}{n-1} = \frac{\sum \left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right)}{n-1} = \frac{\sum X_i^2 - 2\overline{X}\sum X_i + n\overline{X}^2}{n-1} \\
&= \frac{\sum X_i - 2n\overline{X}^2 + n\overline{X}^2}{n-1} = \frac{\sum X_i - n\overline{X}^2}{n-1} \\
&= \frac{n\hat{p} - n\hat{p}^2}{n-1} = \frac{n}{n-1}\hat{p}(1-\hat{p}) \\
&\doteq \hat{p}(1-\hat{p})
\end{aligned}$$

From Example 4.2.2, a large sample confidence interval for μ is

$$\left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

so equivalently, a large sample confidence interval for the population proportion p is

$$\left(\hat{p} - z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right)$$

In particular, a 95% confidence interval for p is

$$\left(\hat{p} - 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right)$$

The term $\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$ is called the *standard error* of $\hat{p}$.

Example 4.2.3 (con't): Recall that $\hat{p} = 8/40 = .20$. A 95% confidence interval for p is

$$.20 \pm 1.96\sqrt{\frac{(.20)(.80)}{40}}$$

$$.20 \pm 1.96(.06)$$

$$(.08, 32)$$

4.2.1 Confidence Intervals for Difference in Means

Let $X_1, \ldots, X_{n_1}$ be a random sample from $f_1(\cdot)$ with mean μ_1 and variance σ_1^2 and $Y_1, \ldots, Y_{n_2}$ be a random sample from $f_2(\cdot)$ with mean μ_2 and variance σ_2^2. In addition, assume that the X sample and Y sample are independent.

Let the difference between means $\Delta = \mu_1 - \mu_2$ be estimated by

$$\hat{\Delta} = \overline{X} - \overline{Y}$$

It can be shown that

$$\begin{aligned}\text{Var}(\hat{\Delta}) &= \text{Var}(\overline{X} - \overline{Y}) = \text{Var}(\overline{X}) + \text{Var}(\overline{Y}) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \doteq \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\end{aligned}$$

so that

$$\text{SE of } \hat{\Delta} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Using the pivot method

$$\begin{aligned}.95 &\doteq P\left[-1.96 \leq \frac{\hat{\Delta} - \Delta}{\text{SE}} \leq 1.96\right] \\ &\vdots \\ &\vdots \\ &= P\left[\hat{\Delta} - 1.96(\text{SE}) \leq \Delta \leq \hat{\Delta} + 1.96(\text{SE})\right] \\ &\equiv P[L \leq \Delta \leq U]\end{aligned}$$

so that $\hat{\Delta} \pm 1.96(\text{SE})$ is an approximate 95% confidence interval.

In general, a $(1-\alpha)100\%$ confidence interval for $\Delta = \mu_1 - \mu_2$ is given by

$$\left((\overline{X} - \overline{Y}) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\overline{X} - \overline{Y}) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$$

Comment: The confidence interval works reasonably well when either one of the following assumptions hold

1. Both distributions $f_1(\cdot)$ and $f_2(\cdot)$ are normal
2. Both sample sizes n_1 and n_2 are reasonably large so that $\overline{X}$ and $\overline{Y}$ are approximately normal by CLT effect

<u>An exact confidence interval for $\mu_1 - \mu_2$</u>

Suppose that the following assumptions hold

1. $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$
2. $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$
3. The X sample and Y sample are independent

Then

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has $N(0,1)$ distribution.

Let

$$S_p^2 = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{j=1}^{n_2}(Y_j - \overline{Y})^2}{n_1 + n_2 - 2} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$$

be a *pooled estimator* of the common variance σ^2. It can be shown that

$$\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

Then

$$\begin{aligned}
\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} &= \frac{\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{(n_1+n_2-2)S_p^2}{\sigma^2}/(n_1+n_2-2)}} \\
&\overset{\mathcal{D}}{=} \frac{N(0,1)}{\sqrt{\chi^2_{n_1+n_2-2}/(n_1+n_2-2)}} \\
&\sim t_{n_1+n_2-2} \text{ df}
\end{aligned}$$

Using the pivot method

$$
\begin{aligned}
.95 &= P\left[-t_{.025,n_1+n_2-2} \leq \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq t_{.025,n_1+n_2-2}\right] \\
&\vdots \\
&= P\left[(\overline{X}-\overline{Y}) - t_{.025,n_1+n_2-2}\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \mu_1-\mu_2 \leq (\overline{X}-\overline{Y}) + t_{.025,n_1+n_2-2}\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right] \\
&\equiv P[L \leq \mu_1-\mu_2 \leq U]
\end{aligned}
$$

so that $(\overline{X}-\overline{Y} \pm t_{.025,n_1+n_2-2}\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ is an exact 95% confidence interval for $\mu_1-\mu_2$.

In general, a $(1-\alpha)100\%$ exact confidence interval for $\mu_1-\mu_2$ is given by

$$\overline{X}-\overline{Y} \pm t_{\alpha/2,n_1+n_2-2}\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

<u>Example in R</u>

```
> trt <- c(91,101,110,103,93,99,104)
> ctl <- c(87,99,77,88,91)
> mean(trt)
[1] 100.1429
> mean(ctl)
[1] 88.4
> sd(trt)
[1] 6.542899
> sd(ctl)
[1] 7.924645
> n1<-length(trt)
> n1
[1] 7
> n2<-length(ctl)
> n2
[1] 5
> df<-n1+n2-2
> df
[1] 10
> qt(.975,df)
[1] 2.228139
> sp<-sqrt(((n1-1)*sd(trt)^2+(n2-1)*sd(ctl)^2)/(n1+n2-2))
> sp
[1] 7.127813
> SE<-sp*sqrt(1/n1+1/n2)
> SE
```

```
[1] 4.17362
> lcl<-mean(trt)-mean(ctl)-qt(.975,df)*SE
> lcl
[1] 2.443453
> ucl<-mean(trt)-mean(ctl)+qt(.975,df)*SE
> ucl
[1] 21.04226

> t.test(trt,ctl,var.equal=TRUE,conf.level=.95)

Two Sample t-test

data:  trt and ctl
t = 2.8136, df = 10, p-value = 0.01836
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  2.443453 21.042262
sample estimates:
mean of x mean of y
 100.1429   88.4000
```