Interval Estimation 2

Day 10 (2/6/20)

Example 4.2.3 (Large sample confidence interval for p). 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. For example, consider the Bernoulli sample: S, F, F, S, F

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

"The sample proportion of successes ($\hat{p}$) is a sample mean ($\overline{X}$) of 1s and 0s", 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

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

$$.20 \pm 1.96(.06)$$
$$(.08, 32)$$

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 $\text{SE}(\hat{\Delta}) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$. Using the pivot method

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

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 of the following 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

An exact confidence interval for $\mu_1 - \mu_2$

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

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} = \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$$

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

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\frac{(n_1+n_2-2)S_p^2}{\sigma^2}/(n_1+n_2-2)}} \stackrel{\mathcal{D}}{=} \frac{N(0,1)}{\sqrt{\chi^2_{n_1+n_2-2}/(n_1+n_2-2)}}$$

Student named the right side a t distribution (with n_1+n_2-2 degrees of freedom).

$$.95 = P\left[-t_{.025,n_1+n_2-2} \le \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \le t_{.025,n_1+n_2-2}\right]$$

$$\vdots$$

$$= P\left[(\overline{X}-\overline{Y})-t_{.025,n_1+n_2-2}\,S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1-\mu_2 \le (\overline{X}-\overline{Y})+t_{.025,n_1+n_2-2}\,S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$$

$$\equiv P[L \le \mu_1-\mu_2 \le U]$$

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

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

Example 4.2.4 The baseball data contains heights of $n_1 = 33$ hitters and $n_2 = 26$ pitchers. The difference between pitcher and hitter heights are estimated as

$$\hat{\Delta} = \overline{X}-\overline{Y} = 75.19 - 72.67 = 2.53 \text{ inches}$$

```
> load(url('http://www.stat.wmich.edu/~mckean/hmchomepage/Data/bb.rda'))
> head(bb)
  hand height weight hitind hitpitind average
1    1     74    218      1         0   3.330
2    0     75    185      1         1   0.286
```

```
3    1     77    219      2          0   3.040
4    0     73    185      1          1   0.271
5    0     69    160      3          1   0.242
6    0     73    222      1          0   3.920

> x<-bb$height[bb$hitpitind==0]
> x
 [1] 74 77 73 78 76 78 76 73 75 76 76 76 75 79 75 78 73 76 75 73 74 73 71 73
     76 76
> y<-bb$height[bb$hitpitind==1]
> y
 [1] 75 73 69 77 72 73 74 72 75 72 68 73 69 76 77 74 73 72 70 75 75 74 71 73
     73 73 72 71 71 74 71 71 70
> cbind(mean(x), mean(y), mean(x)-mean(y))
         [,1]     [,2]     [,3]
[1,] 75.19231 72.66667 2.525641
> s1<-sd(x)
> s2<-sd(y)
> sp<-sqrt(( (26-1)*s1^2+(33-1)*s2^2 )/(26+33-2))
> cbind(s1, s2, sp)
           s1       s2       sp
[1,] 1.959984 2.217356 2.108345
> SE <-sp*sqrt(1/26 + 1/33)
> SE
[1] 0.5528713
> qt(.975,26+33-2)
[1] 2.002465
> L<-2.525641-2.002465*.5528713
> U<-2.525641+2.002465*.5528713
> cbind(L,U)
            L        U
[1,] 1.418536 3.632746
>
> # Using t.test
> t.test(x,y,var.equal=T,conf.level=.95)

Two Sample t-test
data:  x and y
t = 4.5682, df = 57, p-value = 2.682e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.418535 3.632747
sample estimates:
mean of x mean of y
 75.19231  72.66667
```