8.2 Bias and Mean Square Error of Point estimators

We are interested in estimating a parameter, θ (or a fct of θ, $\tau(\theta)$) based on data from a r.s. $y_1 \cdots y_n$

e.g. $Y \sim N(\mu, \sigma)$ $\quad \theta_1 = \mu$, $\theta_2 = \sigma$

e.g. $Y \sim POI(\lambda)$ $\quad \theta = \lambda$

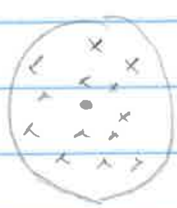
Estimate $P(Y=0) = e^{-\lambda}$ or $\tau(\lambda) = e^{-\lambda}$

The point estimate of θ using sample data is denoted $\hat{\theta}$ (or $\hat{\tau}(\theta)$)

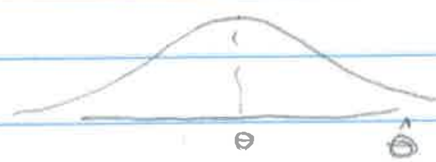
e.g. We might use $\hat{\theta} = \bar{x}$ to estimate $\theta_1 = \mu$

" " " $\hat{\theta} = s^2$ to " $\theta_2 = \sigma^2$

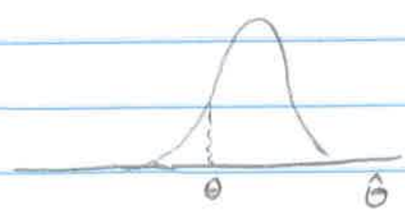
What are good and bad estimators? We evaluate estimators using their sampling dist. i.e. How do they perform over the longhaul? (marksmen analogy)



unbiased, high variance



biased, low variance



To evaluate how good an estimator is, we are certainly interested in $V(\hat{\theta})$ (lower is better). However, $V(\hat{\theta})$ is not the entire story if $\hat{\theta}$ is biased.

Def 8.2 Let $\hat{\theta}$ be a point estimator for a parameter θ. Then $\hat{\theta}$ is an unbiased estimator if $E(\hat{\theta}) = \theta$. If $E(\hat{\theta}) \neq \theta$ then $\hat{\theta}$ is said to be biased.

Def. 8.3 The bias of a pt. estimator is given by $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.

Ideally, we would like an unbiased estimator with low variance. Instead of using $B(\hat{\theta})$ and $V(\hat{\theta})$, we use $E[(\hat{\theta} - \theta)^2]$, the avg. of the square of the distance between the estimator and its target parameter.

Def 8.4 The mean square error of a pt. estimator $\hat{\theta}$ is

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

note:

$$\hat{\theta} - \theta = (\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta) = \hat{\theta} - E(\hat{\theta}) + B(\hat{\theta})$$

$$\Rightarrow MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E(\hat{\theta}) + B(\hat{\theta}))^2]$$

$$= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2 \cdot \underbrace{B(\hat{\theta})}_{\text{constant}} \cdot (\hat{\theta} - E(\hat{\theta})) + \underbrace{B(\hat{\theta})^2}_{\text{constant}}]$$

$$= var(\hat{\theta}) + 2 \cdot B(\hat{\theta}) \cdot E(\hat{\theta} - E(\hat{\theta})) + B(\hat{\theta})^2$$

$$= var(\hat{\theta}) + 2B(\hat{\theta})[\cancelto{0}{E(\hat{\theta}) - E(\hat{\theta})}] + B(\hat{\theta})^2$$

$$= var(\hat{\theta}) + B(\hat{\theta})^2$$

$$\Rightarrow \boxed{MSE(\hat{\theta}) = var(\hat{\theta}) + B(\hat{\theta})^2}$$

8.8 Y_1, Y_2, Y_3 is a r.s. from $Y \sim EXP(\theta)$

$\hat{\theta}_1 = Y_1$ $\quad \hat{\theta}_2 = \frac{Y_1 + Y_2}{2}$ $\quad \hat{\theta}_3 = \frac{Y_1 + 2Y_2}{3}$ $\quad \hat{\theta}_4 = \min(Y_1, Y_2, Y_3)$

$\hat{\theta}_5 = \bar{Y}$

a) which estimators are unbiased?

$E(\hat{\theta}_1) = E(Y_1) = \theta$ ✓

$E(\hat{\theta}_2) = \frac{1}{2}[E(Y_1) + E(Y_2)] = \frac{1}{2}[\theta + \theta] = \theta$ ✓

$E(\hat{\theta}_3) = \frac{1}{3}[E(Y_1) + 2E(Y_2)] = \frac{1}{3}[\theta + 2\theta] = \theta$ ✓

from before $g_{(1)}(y) = n[1 - F(y)]^{n-1} f(y)$

$$= 3\left[1 - (1 - e^{-y/\theta})\right]^2 \cdot \frac{1}{\theta} e^{-y/\theta}$$

$$= 3\left[e^{-2y/\theta}\right] \cdot \frac{1}{\theta} e^{-y/\theta}$$

$$= \frac{3}{\theta} e^{-3y/\theta} \quad \sim EXP(\theta' = \theta/3)$$

$\Rightarrow E(\hat{\theta}_4) = \theta/3$ X biased

$E(\hat{\theta}_5) = E(\bar{Y}) = \mu = \theta$ ✓

b) For unbiased estimators, which one has minimum variance?

note: Y_i are ind., $Var(Y_i) = \theta^2$

$var(\hat{\theta}_1) = \theta^2$

$var(\hat{\theta}_2) = \frac{1}{4}(var(Y_1) + var(Y_2)) = \frac{2\theta^2}{4} = \frac{\theta^2}{2}$

$var(\hat{\theta}_3) = \frac{1}{9} var(Y_1) + \frac{4}{9} var(Y_2) = \frac{5}{9}\theta^2$

$var(\hat{\theta}_5) = var(\bar{Y}) = \frac{\sigma^2}{n} = \frac{\theta^2}{3}$ $\Rightarrow$ minimum variance

8.10 $Y \sim POI(\lambda)$

a) Try $\boxed{\hat{\lambda} = \bar{y}}$ $\Rightarrow E(\hat{\lambda}) = \mu = \lambda$ unbiased

b) $C = 3Y + Y^2 \Rightarrow E(C) = 3E(Y) + E(Y^2)$

$= 3\lambda + (\lambda + \lambda^2) = 4\lambda + \lambda^2$

c) Find $\hat{c}$ where $E(\hat{c}) = E(c)$

$E(\bar{Y}) = \mu = \lambda$

$E(\bar{Y}^2) = var(\bar{Y}) + E(\bar{Y})^2 = \frac{\sigma^2}{n} + \mu^2 = \frac{\lambda}{n} + \lambda^2$

Let $\hat{c} = a \cdot \bar{Y} + \bar{Y}^2 \Rightarrow E(\hat{c}) = a \cdot \lambda + \frac{\lambda}{n} + \lambda^2 \overset{SET}{=} 4\lambda + \lambda^2$

$\Rightarrow (a + \frac{1}{n}) = 4 \Rightarrow a = 4 - 1/n$

$\Rightarrow \boxed{\hat{c} = (4 - \frac{1}{n})\bar{y} + \bar{y}^2}$

8.14 $f(y) = \begin{cases} \dfrac{\alpha y^{\alpha-1}}{\theta^\alpha}, & 0 \le y \le \theta \\ 0, & \text{———} \end{cases}$ $\alpha > 0$ known, fixed value

θ unknown

$\hat{\theta} = \max(Y_1, \dots, Y_n)$

a) Recall $g_{(n)}(y) = n[F(y)]^{n-1} f(y)$

$$F(y) = \int_0^y \frac{\alpha t^{\alpha-1}}{\theta^\alpha}\, dt = \frac{t^\alpha}{\theta^\alpha}\Big|_0^y = \left(\frac{y}{\theta}\right)^\alpha, \quad 0 \le y < \theta$$

$$g_{(n)}(y) = n\left[\left(\frac{y}{\theta}\right)^\alpha\right]^{n-1} \cdot \frac{\alpha y^{\alpha-1}}{\theta^\alpha} = n\left(\frac{y}{\theta}\right)^{\alpha(n-1)} \cdot \frac{\alpha \cdot y^{\alpha-1}}{\theta^\alpha}$$

$$= \frac{n \cdot \alpha}{\theta^{\alpha n}} \cdot y^{\alpha n-1}, \quad 0 \le y < \theta$$

$$E[\hat{\theta}] = \int_0^\infty \frac{\alpha n}{\theta^{\alpha n}} \cdot y^{\alpha n}\, dy = \frac{\alpha n \cdot y^{\alpha n+1}}{(\alpha n+1) \cdot \theta^{\alpha n}}\Big|_0^\theta$$

$$= \boxed{\frac{\alpha n}{(\alpha n+1)} \cdot \theta} \ne \theta \Rightarrow \text{biased}$$

b) Let $\boxed{\hat{\theta}_2 = \dfrac{\alpha n+1}{\alpha n} \cdot \hat{\theta}} \Rightarrow E(\hat{\theta}_2) = \dfrac{\alpha n+1}{\alpha n} E(\hat{\theta}) =$

$$= \frac{\alpha n+1}{\alpha n} \cdot \frac{\alpha n}{\alpha n+1} \theta = \theta \checkmark \text{ unbiased}$$

c) $B(\hat{\theta}) = \frac{\alpha n}{\alpha n+1} \cdot \theta - \theta = \frac{\alpha n - \alpha n - 1}{\alpha n+1} \cdot \theta = \frac{-1}{\alpha n+1} \cdot \theta$

$E(\hat{\theta}^2) = \int_0^\theta \frac{\alpha n}{\theta^{\alpha n}} y^{\alpha n+1} \, dy = \frac{\alpha n y^{\alpha n+2}}{(\alpha n+2) \cdot \theta^{\alpha n}} \Big|_0^\theta = \frac{\alpha n}{\alpha n+2} \cdot \theta^2$

$\Rightarrow var(\hat{\theta}) = \frac{\alpha n}{\alpha n+2} \theta^2 - \left(\frac{\alpha n}{\alpha n+1}\right)^2 \cdot \theta^2$

$\Rightarrow MSE(\hat{\theta}) = \overbrace{\frac{\alpha n}{\alpha n+2} \theta^2 - \frac{\alpha^2 n^2 \theta^2}{(\alpha n+1)^2}}^{var(\hat{\theta})} + \overbrace{\frac{\theta^2}{(\alpha n+1)^2}}^{B(\hat{\theta})^2}$

$= \boxed{\frac{\alpha n \theta^2}{\alpha n+2} + \frac{(1-\alpha^2 n^2)\theta^2}{(\alpha n+1)^2}}$

8.16

From Ch.4 ... $Y \sim GAM(\alpha, \beta)$ $\quad E(Y^a) = \frac{\beta^a \Gamma(\alpha+a)}{\Gamma(\alpha)}$

If $y_1 \cdots y_n$ B a r.s. from $Y \sim N(\mu, \sigma^2)$

$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$ or $GAM(\alpha = \frac{n-1}{2}, \beta = 2)$

$W^{1/2} = \left[\frac{(n-1)S^2}{\sigma^2}\right]^{1/2} = \frac{\sqrt{n-1}\, S}{\sigma}$

$\Rightarrow E(W^{1/2}) = \frac{2^{1/2} \cdot \Gamma(\frac{n-1}{2} + \frac{1}{2})}{\Gamma(\frac{n-1}{2})}$

⑦

$$S = \frac{\sigma}{\sqrt{n-1}}\left(W^{1/2}\right) \Rightarrow E(S) = \frac{\sigma}{\sqrt{n-1}} \cdot \frac{2^{1/2} \cdot \Gamma\left(\frac{n-1}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$$ (i.e. S is biased)

let $\boxed{\hat{\sigma} = \frac{\sqrt{n-1}\,\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\cdot\Gamma\left(\frac{n-1}{2}+\frac{1}{2}\right)} \cdot S}$ $\hat{\sigma}$ is unbiased for σ

8.3 Common Unbiased Point Estimators

1. Parameter: μ

pt. estimate: $\bar{y}$; $y_1 \cdots y_n$ is a r.s. from pop. w/ μ, σ

$$E[\bar{y}] = E\left[\frac{y_1 + \cdots + y_n}{n}\right] = \frac{1}{n}E(y_1) + \cdots + \frac{1}{n}E(y_n)$$

$$= \frac{1}{n}\cdot\mu + \cdots + \frac{1}{n}\cdot\mu = \frac{n\mu}{n} = \boxed{\mu}$$

$$var(\bar{y}) = var\left(\frac{y_1 + \cdots + y_n}{n}\right)$$

$$= \frac{1}{n^2}\left[var(y_1) + \cdots + var(y_n)\right] = \frac{n\sigma^2}{n^2} = \boxed{\frac{\sigma^2}{n}}$$

Simulation

2. Parameter: p

pt. estimate: $\hat{p}$

$y_1 \cdots y_n$ a r.s from $Bin(1, p)$ $\Rightarrow E(y_i) = p$, $var(Y_i) = p(1-p)$

$$\Rightarrow E(\hat{p}) = E\left[\frac{y_1 + \cdots + y_n}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \frac{np}{n} = \boxed{p}$$

$$\Rightarrow var(\hat{p}) = var\left[\frac{y_1 + \cdots + y_n}{n}\right] = \frac{1}{n^2}\sum_{i=1}^{n} var(y_i) = \frac{np(1-p)}{n^2} = \boxed{\frac{p(1-p)}{n}}$$

3. parameter: $\mu_1 - \mu_2$

pt. estimate: $\bar{y}_1 - \bar{y}_2$

$y_{11} \cdots y_{1n_1}$ is a r.s. from Pop 1 w/ μ_1, σ_1

$y_{21} \cdots y_{2n_2}$ " " " " 2 w/ μ_2, σ_2

from 1 ...

$E(\bar{y}_1) = \mu_1$, $var(\bar{y}_1) = \frac{\sigma_1^2}{n_1}$

$E(\bar{y}_2) = \mu_2$, $var(\bar{y}_2) = \sigma_2^2/n_2$

$$E[\bar{y}_1 - \bar{y}_2] = E(\bar{y}_1) - E(\bar{y}_2) = \boxed{\mu_1 - \mu_2}$$

$var(\bar{y}_1 - \bar{y}_2) = var(\bar{y}_1) + var(\bar{y}_2)$ note: $\bar{y}_1, \bar{y}_2$ are ind.

$$= \boxed{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

4 parameter: $p_1 - p_2$

pt. estimate: $\hat{p}_1 - \hat{p}_2$

$y_{11} \cdots y_{1n_1}$ is a r.s. from $Bin(1, p_1)$

$y_{21} \cdots y_{2n_2}$ " " " $Bin(1, p_2)$

from 2 ...

$E(\hat{p}_1) = p_1$ $var(\hat{p}_1) = \frac{p_1(1-p_1)}{n_1}$

$E(\hat{p}_2) = p_2$ $var(\hat{p}_2) = \frac{p_2(1-p_2)}{n_2}$

$E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = \boxed{p_1 - p_2}$

$V(\hat{p}_1 - \hat{p}_2) = var(\hat{p}_1) + var(\hat{p}_2) = \boxed{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

note: $\hat{p}_1, \hat{p}_2$ are ind.

note: These results are valid regardless of sample size and shape of the dist. we are sampling from. However, the CLT says the shape will be approx. normal for $n \geq 30$.

Example 8.1

$\sigma^2 = \frac{\sum_{i=1}^{N} (y_i - \mu)^2}{N}$ why don't we use $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}$?

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \left[y_i^2 - 2y_i \cdot \bar{y} + \bar{y}^2 \right] = \sum y_i^2 - 2\bar{y} \cdot \sum y_i + n\bar{y}^2$$

$= \sum y_i^2 - 2n\bar{y}^2 + n\bar{y}^2 = \sum y_i^2 - n\bar{y}^2$

$E(y_i^2) = \sigma^2 + \mu^2 \qquad E(\bar{y}^2) = \frac{\sigma^2}{n} + \mu^2$

$$\Rightarrow E\left[\sum y_i^2 - n\bar{y}^2\right] = \sum E(y_i^2) - nE(\bar{y}^2)$$
$$= n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= n\cdot\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2$$
$$= (n-1)\sigma^2$$

$$\Rightarrow E[\hat{\sigma}^2] = E\left[\frac{\sum(y_i - \bar{y})^2}{n}\right] = \frac{1}{n}\left[(n-1)\sigma^2\right] = \boxed{\frac{(n-1)\sigma^2}{n}}$$ simulation

$\Rightarrow \hat{\sigma}^2$ is biased i.e. it slightly underestimates σ on avg

note: bias is small for large n

let $s^2 = \frac{n\cdot\hat{\sigma}^2}{n-1} = \frac{\sum(y_i - \bar{y})^2}{n-1}$

$$\Rightarrow E(s^2) = \frac{n}{n-1}E(\hat{\sigma}^2) = \frac{n}{n-1}\left(\frac{(n-1)\sigma^2}{n}\right) = \sigma^2$$

Now we know why we divide by $n-1$ in 2050!

Another approach...

Recall $y_1, \dots, y_n$ a r.s. from $Y \sim N(\mu, \sigma^2)$

$$\Rightarrow \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

$$E\left[\frac{(n-1)s^2}{\sigma^2}\right] = n-1 = \frac{(n-1)}{\sigma^2}\cdot E(s^2)$$

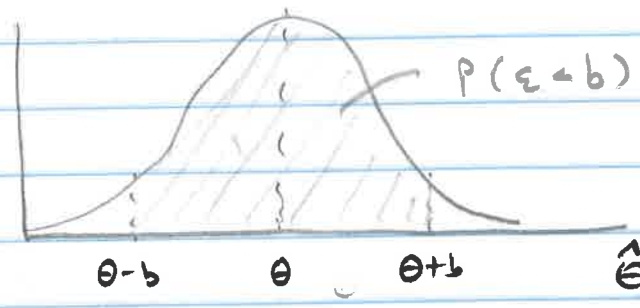
$$\Rightarrow E(s^2) = \frac{\sigma^2}{n-1}(n-1) = \sigma^2$$

note: this depends on sampling from a normal dist.

8.4 Goodness of a Point Estimator

Def 8.5 The error of estimation ε is the distance between an estimator and its target parameter. That is, $\varepsilon = |\hat{\theta} - \theta|$

Fig 8.4 Sampling dist. of $\hat{\theta}$ (unbiased)



$$P(\varepsilon < b) = \int_{\theta-b}^{\theta+b} f(\hat{\theta})\,d\hat{\theta}$$

note: we must know $f(\hat{\theta})$ to compute $P(\varepsilon < b)$ exactly. otherwise we estimate it (lower bound) using Tchebysheff's inequality.

a commonly used range is $b = 2 \cdot SE(\hat{\theta})$. For many dists. $P(\varepsilon < 2 \cdot SE(\hat{\theta})) \approx .95$

Table 8.2

Dist.	$P(\mu - 2\sigma < Y < 2\sigma)$
normal	.9544
Exponential	.9502
uniform	1

Exponential $Y \sim Exp(\beta)$

$E(Y) = \beta \quad var(Y) = \beta^2$

$\mu \pm 2\sigma = \beta \pm 2 \cdot \beta = (-\beta, 3\beta)$

$$\int_{-\beta}^{3\beta} \frac{1}{\beta} e^{-y/\beta} dy = \int_{0}^{3\beta} \frac{1}{\beta} e^{-y/\beta} dy = -e^{-y/\beta} \Big|_0^{3\beta}$$

$$= 1 - e^{-3\beta/\beta} = 1 - e^{-3} = .9502$$

e.g. 8.27 $n = 985$ "likely voters"

$x = 592$ will vote Republican

a) p is estimated to be $.601$ (i.e. $\hat{p} = \frac{592}{985} = .601$)

$$b = 2 \cdot \hat{SE}(\hat{p}) = 2 \cdot \sqrt{\frac{(.601)(.399)}{985}} = .0312$$

b) Do you think the Republican candidate will be elected?

$.601 \pm .0312 \rightarrow (.5698, .6322)$

Republican should be elected. It's highly unlikely $\hat{p} = .601$ could be more than $.10$ away from p.

c) Trump ran it.

$\boxed{\text{Simulation}}$

e.g. 8.23 EPA and Univ. of Florida

Covariates: calcium in drinking water, smoking activity

response: kidney-stone disease

Data is collected on individuals w/ recurring kidney stone problems.

<u>note</u>: retrospective observational study i.e. collect data based on response

	Carolinas	Rockies
n	467	191
avg. age	45.1	46.4
SD(age)	10.2	9.8
avg. calcium	11.3	40.1
SD(calcium)	16.6	28.4
proportion smoking	.78	.61

a) $\hat{\mu}_c = 11.3 \quad b = 2 \cdot \frac{16.6}{\sqrt{467}} = 1.54$

b) $\hat{\mu}_R - \hat{\mu}_c = (40.1 - 11.3) = 28.8$

$$b = 2\sqrt{\frac{16.6^2}{467} + \frac{28.4^2}{191}} = 4.39$$

→ significant difference in calcium
"at least" 24.41 higher in Rockies

c) $\hat{p}_c - \hat{p}_R = .78 - .61 = .17$

$$b = 2\sqrt{\frac{.78(.22)}{467} + \frac{(.61)(.39)}{191}} = .08$$

→ significant difference in smoking
"at least" .09 higher in Carolinas

<u>note</u>: Retrospective studies might help in direction for future studies. We can't draw any conclusions from retrospective studies themselves.

e.g. what % of cancer patients have 2 arms? 99%?

8.5 Confidence Intervals

An interval estimator will specify an interval in which we think it is highly likely that the parameter θ is located.

Ideally, the likelihood our interval contains θ is high (e.g. 95%, 98%, etc.) and the interval is narrow.

$P(\hat{\theta}_L \le \theta \le \hat{\theta}_u) = 1-\alpha$ → confidence coefficient

fcts of sample data

Since the interval $(\hat{\theta}_L, \hat{\theta}_u)$ depends on sample data, it is random and will vary from one sample to the next. In truth, the confidence coefficient $1-\alpha$, is the proportion of time $\theta \in (\hat{\theta}_L, \hat{\theta}_u)$ in repeated sampling.

One-sided confidence intervals:

- $P(\hat{\theta}_L \le \theta) = 1-\alpha \Rightarrow$ We are $(1-\alpha)\times 100\%$ confident $\theta \ge \hat{\theta}_L$

  i.e. one-sided lower CI for θ, $[\hat{\theta}_L, \infty)$

- $P(\theta \le \hat{\theta}_u) = 1-\alpha \Rightarrow$ We are $(1-\alpha)\times 100\%$ confident $\theta \le \hat{\theta}_u$

  i.e. one-sided upper CI for θ, $(-\infty, \hat{\theta}_u]$.

How do we determine these intervals? We need a pivotal quantity

(1) fct of sample data and θ

note: θ is the only unknown quantity

② The prob. dist. does <u>not</u> depend on θ

e.g. 8.39

$Y \sim GAM(\alpha = 2, \beta)$ Find a 90% CI for β

Pivotal Quantity?

recall, $\frac{2Y}{\beta} \sim \chi^2_{(4)}$ (see 6.46)

$\frac{2Y}{\beta}$ is ① a fct of sample data and β

② the prob. dist. does not depend on β

$$P\left[\chi^2_{.95} \le \frac{2Y}{\beta} \le \chi^2_{.05}\right] = .90$$

$$P\left[\frac{2Y}{\chi^2_{.05}} \le \beta \le \frac{2Y}{\chi^2_{.95}}\right] = .90$$

i.e. 90% CI for β: $\boxed{\left(\frac{2Y}{\underset{9.49}{\chi^2_{.05}}}, \frac{2Y}{\underset{.71}{\chi^2_{.95}}}\right)}$ or $\boxed{(.211Y, 2.817Y)}$

e.g. 8.40

$Y \sim N(\mu, \sigma^2 = 1)$

a) Find a 95% CI for μ:

Pivotal Qty?

$$\underbrace{Y - \mu}_{\text{fct of } Y, \mu} \sim \underbrace{N(0,1)}_{\text{does not depend on } \mu}$$

$$P\left[-z_{.025} \le Y - \mu \le z_{.025}\right] = .95$$

$$P\left[-1.96 \le Y - \mu \le 1.96\right] = .95$$

$= P[Y - 1.96 \le \mu \le Y + 1.96] = .95$

i.e. 95% CI for μ: $\boxed{(Y-1.96,\ Y+1.96)}$

b) Find a 95% upper CI for μ:

$P[-z_{.05} \le Y - \mu] = .95$

$P[-1.645 \le Y - \mu] = .95$

$P[\mu \le Y + 1.645] = .95$ i.e. 95% upper CI for μ: $\boxed{(-\infty,\ Y+1.645)}$

e.g. 8.43 $Y_1 \cdots Y_n$ a r.s. from $Y \sim \text{unif}(0, \theta)$ $\Rightarrow F(y) = \frac{y}{\theta}$

let $Y_{(n)} = \max(Y_1, \dots, Y_n)$ and $U = \frac{Y_{(n)}}{\theta}$

a) $F_{Y_{(n)}}(y) = P[Y_{(n)} \le y] = \left(\frac{y}{\theta}\right)^n$

$F_U(u) = P[U \le u] = P\left[\frac{Y_{(n)}}{\theta} \le u\right] = P[Y_{(n)} \le u\theta]$

$= \left(\frac{u\theta}{\theta}\right)^n = u^n$

$$\Rightarrow \boxed{F_u(u) = \begin{cases} 0, & u < 0 \\ u^n, & 0 \le u < 1 \\ 1, & 1 \le u \end{cases}}$$

b) Find 95th percentile of U:

$u^n = .95 \Rightarrow u = (.95)^{1/n}$ or $\phi_{.95} = (.95)^{1/n}$

$P[U \le (.95)^{1/n}] = P\left[\frac{Y_{(n)}}{\theta} \le (.95)^{1/n}\right] = P\left[\frac{Y_{(n)}}{(.95)^{1/n}} \le \theta\right]$

$\Rightarrow$ 95% one-sided lower CI for θ: $\boxed{\left[\frac{Y_{(n)}}{(.95)^{1/n}}, \infty\right)}$

(8.48) $Y_1 \cdots Y_n$ a r.s. from $Y \sim GAM(\alpha=2, \beta)$

a) Let $W = \frac{2 \sum y_i}{\beta}$

$$m_W(t) = E\left[e^{tW}\right] = E\left[e^{t \cdot \frac{2(\sum y_i)}{\beta}}\right] = E\left[e^{\frac{2t}{\beta}(\sum y_i)}\right]$$

$$= \prod_{i=1}^{n} E\left[e^{\frac{2t}{\beta} \cdot y_i}\right] = \prod_{i=1}^{n} m_Y\left(\frac{2t}{\beta}\right) = \prod_{i=1}^{n} \frac{1}{\left(1 - \frac{2t}{\beta} \cdot \beta\right)^2}$$

$$= \frac{1}{(1-2t)^{2n}} \Rightarrow W \sim \chi^2_{(4n)}$$

b) $P\left[\chi^2_{.975} \le W \le \chi^2_{.025}\right] = .95$

$$P\left[\chi^2_{.975} \le \frac{2(\sum y_i)}{\beta} \le \chi^2_{.025}\right] = .95$$

$$P\left(\frac{2(\sum y_i)}{\chi^2_{.025}} \le \beta \le \frac{2(\sum y_i)}{\chi^2_{.975}}\right] = .95$$

$\Rightarrow$ 95% CI for β: $\left(\frac{2(\sum y_i)}{\chi^2_{.025}}, \frac{2(\sum y_i)}{\chi^2_{.975}}\right)$

c) $n=5 \quad \bar{y} = 5.39$

$$\left(\frac{2(5)(5.39)}{20.4831}, \frac{2(5)(5.39)}{3.24697}\right) = \boxed{(2.63, 16.60)}$$

8.6 Large Sample CIs

The CLT has shown the sums of RVs have an approximately normal dist as the sample size gets large.

In Sec 8.3 we found unbiased estimators for μ, p, $\mu_1 - \mu_2$, and $p_1 - p_2$. Standard errors for these pt. estimators are found in Table 8.1 (p. 397)

For large samples we can now use the CLT result to form a pivotal quantity for μ, p, $\mu_1 - \mu_2$, and $p_1 - p_2$.

$$z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

Prob. stmt:

$$P\left[-z_{\alpha/2} \le \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \le z_{\alpha/2}\right] = 1 - \alpha$$

manipulate algebraically to isolate the parameter

(note: This will be a confidence stmt)

$$C\left[\hat{\theta} - z_{\alpha/2} \cdot \sigma_{\hat{\theta}} \le \theta \le \hat{\theta} + z_{\alpha/2} \cdot \sigma_{\hat{\theta}}\right] = 1 - \alpha$$

$\Rightarrow$ $100(1-\alpha)\%$ CI for θ: $(\hat{\theta}_L, \hat{\theta}_U)$

$$\boxed{\hat{\theta}_L = \hat{\theta} - z_{\alpha/2} \cdot \sigma_{\hat{\theta}}, \quad \hat{\theta}_U = \hat{\theta} + z_{\alpha/2} \cdot \sigma_{\hat{\theta}}}$$

8.60 Estimate μ = "normal" temp. for healthy humans

$n = 130 \quad \bar{y} = 98.25 \quad s = .73$

a) Find a 99% CI for μ:

Ideally we would use $\bar{y} \pm z_{\alpha/2} \cdot \underbrace{\sigma/\sqrt{n}}_{\sigma_{\bar{y}}}$

However, in this context σ is unknown so we substitute s

$98.25 \pm 2.576 \dfrac{(.73)}{\sqrt{130}} \rightarrow 98.25 \pm .165 \rightarrow \boxed{(98.085, 98.415)}$

b). 99% confident the avg. temp is below 98.6

note: This is avg healthy human temp. This does not answer the question of the range of temps for healthy humans

Chebychev's rule: at least 75% of healthy humans have temp. in range:

$98.25 \pm 2(.73) = 98.25 \pm 1.46 \rightarrow \boxed{(96.79, 99.71)}$

Compare large sample CIs for μ:

$Y \sim POI(\lambda = 3) \quad n = 50$

$\Rightarrow \mu = 3 \quad \sigma = \sqrt{3}$

- $\bar{y} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$ vs $\bar{y} \pm z_{\alpha/2} \cdot s/\sqrt{n}$

Simulation   note: both are approximate

8.65 Compare defective rates for two assembly lines

Line A	Line B
$n_1 = 100$	$n_2 = 100$
$Y_1 = 18$	$Y_2 = 12$
$\hat{p}_1 = \frac{18}{100} = .18$	$\hat{p}_2 = \frac{12}{100} = .12$

a) Find a 98% CI for $p_1 - p_2$

$$\hat{\theta} \pm z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

note: Since p_1, p_2 are unknown, substitute $\hat{p}_1$ and $\hat{p}_2$

$$(.18 - .12) \pm 2.326 \sqrt{\frac{(.18)(.82)}{100} + \frac{(.12)(.88)}{100}}$$

$$.06 \pm .12 \rightarrow \boxed{(-.06, .18)}$$

b) No statistically significant evidence that Line A has a higher defective rate than Line B, i.e. .18 vs. .12 could be reasonably explained as chance variation.

Compare CI coverage for p: $(p = .5,\ p = .1)$

- $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$ vs. $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\boxed{\text{Simulation}}$

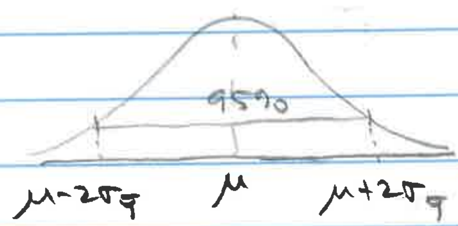
8.7 How large should n be?

A common question for statisticians, especially in design of experiment

How accurate does the researcher want to be?

1. within ...? (E)
2. with confidence level?

Spse we want to estimate μ using $\bar{Y}$. From CLT ($n \geq 30$)



i.e. 95% of the time, $\bar{Y}$ is within $2\sigma_{\bar{Y}}$ of μ

If the researcher wants to be within E of μ with 95% confidence then

$$2 \cdot \sigma_{\bar{Y}} \overset{\text{SET}}{=} E \rightarrow 2 \cdot \frac{\sigma}{\sqrt{n}} = E \rightarrow \sqrt{n} = \frac{2\sigma}{E}$$

$$\Rightarrow \boxed{n = \left(\frac{2\sigma}{E}\right)^2}$$

Problem: In this context (estimating μ) we wouldn't know σ, so it must be estimated before we take the sample.

1. use s from a prior sample
2. If we have an approximation of the range, use $\sigma \approx \frac{\text{range}}{4}$

Recall the empirical rule $\mu \pm 2\sigma \rightarrow$ approx 95%

note: We might think we are being more conservative and use $\mu \pm 3\sigma \rightarrow$ at least 89% (Chebychov's)

$\Rightarrow \sigma \approx \frac{\text{range}}{6}$

However, this will produce a smaller value for n, and we might not have the level of precision we wanted. Better to overestimate σ slightly ($\sigma \approx$ range/4) and then our actual level of confidence will be slightly higher than we specify.

In general to estimate μ ...

$\bar{y}$ will be within $z_{\alpha/2} \cdot \sigma_{\bar{y}}$ $100(1-\alpha)\%$ of the time

$$\boxed{n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2}$$

How about election polls? Estimating p

from sampling theory we know...

$$\hat{p} \text{ is approx } N\left(\mu_{\hat{p}} = p, \ \sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

95% of the time $\hat{p}$ is within $2\sigma_{\hat{p}}$ of p

$$E = 2\sigma_{\hat{p}} \rightarrow E = 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \rightarrow E = 2\sqrt{\frac{\overbrace{(.5)(.5)}^{\text{conservative estimate}}}{n}}$$

$$\rightarrow \boxed{E = \frac{1}{\sqrt{n}}}$$

$\frac{1}{\sqrt{n}}$ is very close to the margin-of-error reported in election polls.

In general use ...

$$z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} = E \rightarrow \sqrt{n} = \frac{z_{\alpha/2}}{E}\sqrt{p(1-p)}$$

$$\rightarrow \boxed{n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \cdot p(1-p)}$$

again we have a problem, we don't know p. It is what we are trying to estimate!

① $p(1-p) \leq .25$ so use $\boxed{n = \left(\frac{z_{\alpha/2}}{2E}\right)^2}$

② use a prior estimate for p

e.g. Estimate proportion of alcoholics in Utah County. Talk to doctors or social workers for a reasonable prior estimate. It certainly isn't as high as .5 so we can save some expense in sampling using $p < .5$

8.78 p_1 = proportion of defectives from assembly line 1
p_2 = " " " " " 2

Estimate $p_1 - p_2$ within $\overbrace{E = .02}^{\text{changed}}$ w/ 95% confidence

$$2 \cdot \sigma_{\hat{p}_1 - \hat{p}_2} = .2 \rightarrow \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = .02$$

assume equal sample sizes $(n_1 = n_2)$

$$\rightarrow \frac{1}{\sqrt{n}} \cdot \sqrt{p_1(1-p_1)+p_2(1-p_2)} = .02$$

$$\rightarrow n = 25\left[p_1(1-p_1)+p_2(1-p_2)\right]$$

Use estimates of p_1 and p_2 from 8.65
(.5 for defectives is too high)

$$\hat{p}_1 = \frac{18}{100} = .18 \qquad \hat{p}_2 = \frac{12}{100} = .12$$

$$n = 2500\left[\underbrace{.18(.82)}_{.1476} + \underbrace{.12(.88)}_{.1056}\right] = 633$$

Take samples of size 633 from each line.

Example 8.10

μ_1 = avg. assembly time for workers using training method I
μ_2 = " " " " " " " II

range $\approx$ 8 mins. Estimate $\mu_1 - \mu_2$ within 1 min. w/ 95% confidence and using equal sample sizes

$\Rightarrow \sigma \approx 8/2 = 4$

$$2 \cdot \sigma_{\bar{Y}_1 - \bar{Y}_2} = 1 \rightarrow 2\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1$$

$$2\sqrt{\frac{4}{n} + \frac{4}{n}} = 1$$

$$\rightarrow 2\sqrt{8/n} = 1 \rightarrow n = (2\sqrt{8})^2 = 32 \text{ workers}$$

(from each assembly line training method)

8.8 Small sample CIs for μ and $\mu_1 - \mu_2$

When n is too small to use CLT ($n<30$) we have "small sample" procedures.

The main assumption here is that we are sampling from a normal dist or "the departure from normality is not excessive"

If we are estimating μ, in practice we won't know σ either. We have seen ...

$$T = \frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

$$P\left[-t_{\alpha/2} \le \frac{\bar{Y}-\mu}{S/\sqrt{n}} \le t_{\alpha/2}\right] = 1-\alpha$$ note: This is a prob stmt.

manipulating the inequalities ...

$$\frac{\bar{Y}-\mu}{S/\sqrt{n}} \le t_{\alpha/2} \Rightarrow \bar{Y}-\mu \le t_{\alpha/2}\cdot S/\sqrt{n} \Rightarrow \bar{Y}-t_{\alpha/2}\cdot S/\sqrt{n} \le \mu$$

$$-t_{\alpha/2} \le \frac{\bar{Y}-\mu}{S/\sqrt{n}} \Rightarrow -t_{\alpha/2}\cdot \frac{S}{\sqrt{n}} \le \bar{Y}-\mu \Rightarrow \mu \le \bar{Y}+t_{\alpha/2}\cdot S/\sqrt{n}$$

$$\Rightarrow C\left\{\bar{Y}-t_{\alpha/2}\cdot \frac{S}{\sqrt{n}} \le \mu \le \bar{Y}+t_{\alpha/2}\cdot S/\sqrt{n}\right] = 1-\alpha$$

note: This is now a confidence stmt. because we have isolated μ which is a fixed unknown parameter.

$100(1-\alpha)\%$ CI for μ: $\boxed{\bar{Y} \pm t_{\alpha/2} \cdot s/\sqrt{n}}$

Example 8.11

μ = avg. muzzle velocity $\boxed{\text{check NPP / Histogram}}$

$n=8$ $\bar{Y}=2959$ $t_{.025} = 2.365$ $s = 39.1$
(d.f. = 7)

$2959 \pm 2.365\,(39.1)/\sqrt{8} \rightarrow 2959 \pm 32.7$ $\boxed{(2926.3,\ 2991.7)}$

Early 1900s:

Gossett using small samples at Guinness Brewing Co.

$\boxed{\text{Simulation}}$ CI: $\bar{Y} \pm z_{\alpha/2} \cdot s/\sqrt{n}$ vs. $\bar{Y} \pm t_{\alpha/2} \cdot s/\sqrt{n}$

Comparing $\mu_1 - \mu_2$

Pop 1: $\sim N(\mu_1, \sigma_1)$ Pop 2: $\sim N(\mu_2, \sigma_2)$
r.s. of n_1 r.s. of n_2

We have seen $\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$

$$\Rightarrow \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

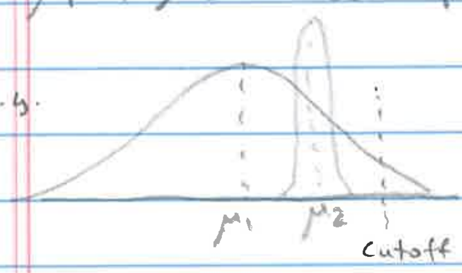
If $\sigma_1 = \sigma_2 = \sigma$ then $\dfrac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$

<u>note:</u>

when comparing μ_1 vs. μ_2, the assumption is often made that $\sigma_1 = \sigma_2$. This seems like a restrictive assumption but it is robust, i.e. if σ_1 and σ_2 are "close" then the confidence coefficient should be close.

Furthermore, if $\sigma_1 << \sigma_2$ or vice versa, comparing μ_1 vs. μ_2 is not appropriate.

e.g.



<u>Goal</u>: Get as many students as possible past "elite" cut off score

$\mu_2 > \mu_1$ but we want method I

Even if we assume $\sigma_1 = \sigma_2 = \sigma$, the common value σ will be unknown in practice when estimating $\mu_1 - \mu_2$.

"pooled" estimator $\boxed{S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$

i.e. wt. avg. of S_1^2 and S_2^2

Let $W = \dfrac{(n_1+n_2-2)\cdot S_p^2}{\sigma^2} = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2}$

$$= \frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2}$$

$$\chi^2_{(n_1-1)} + \chi^2_{(n_2-1)} \sim \chi^2_{(n_1+n_2-2)}$$

We know from ch. 7 ... $\dfrac{Z}{\sqrt{W_{(v)}/v}} \sim t_{(v)}$

$$T = \frac{\dfrac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\dfrac{\frac{(n_1+n_2-2)\cdot S_p^2}{\sigma^2}}{\cancel{(n_1+n_2-2)}}}}$$

$$= \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \cdot \sqrt{\frac{\sigma^2}{S_p^2}}$$

$$= \boxed{\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}} \sim t_{(n_1+n_2-2)} \quad \text{Pooled T-test}$$

8.90

a) $S_p = \sqrt{\dfrac{14(42)^2 + 14(45)^2}{28}} = \sqrt{\dfrac{24,696 + 28,350}{28}} = 43.52$

$t_{.025,28} = 2.048$

from above result ...

$$P\left[-t_{\alpha/2} \le \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \le t_{\alpha/2}\right] = 1 - \alpha$$

manipulating the inequalities as in the one-sample case...

$$P\left[(\bar{Y}_1-\bar{Y}_2)-t_{\alpha/2}\cdot s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1-\mu_2 \le (\bar{Y}_1-\bar{Y}_2)+t_{\alpha/2}\cdot s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$$

$$= 1-\alpha$$

$\Rightarrow 100(1-\alpha)\%$ CI for $\mu_1-\mu_2$:

$$\boxed{(\bar{Y}_1-\bar{Y}_2) \pm t_{\alpha/2}\cdot s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Back to example...

$$(534-446) \pm 2.048\sqrt{\frac{1}{15}+\frac{1}{15}}$$

$$88 \pm .75 \Rightarrow \boxed{(87.25,\ 88.75)}$$

c) We are 95% confident that the avg. verbal scores majors in Language / Literature is between 87.25 and 88.75 pts higher than the avg. verbal score for engineering majors.

d) $\sigma_1 = \sigma_2$ → seems reasonable w/ $S_1 = 42$ and $S_2 = 45$

normal pops. → probably at least approx. normal for national test scores.