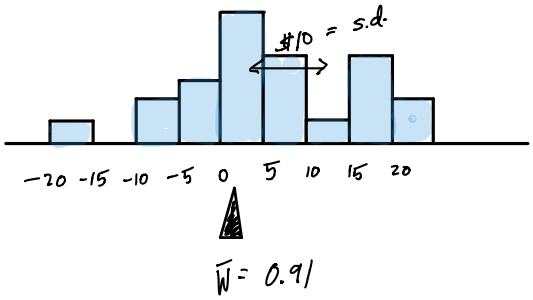
Lecture 10: Estimation

Start with a motivating example.

Set up:

- You are watching someone repeatedly play a slot machine in Las Vegas.
- They played n times and left
- You recorded the winnings of each play

Suppose the histogram of these W_i's looks like this



$\bar{W} = 0.91$

- Question: Can you use this data to determine if this machine has a positive expected payout, i.e. if

$$E(W_{n+1}) > 0$$

↪ if so, you should sit down and play as much as possible.

Let $\mu = E(W_{n+1})$. So μ is unknown and we want to use the previous play to conjecture if $\mu > 0$ or not.

Is it reasonable to think μ is near \$0.91?

If so, how wrong could this be...

... if it can be wrong by $\pm$\$10 then we can't be sure if $\mu > 0$ or not

... if it can be wrong by only $\pm$\$0.1 then we can be confident $\mu > 0$.

Case 1: $n = 100$

$$E(\bar{W}) = E(W_i) = \mu$$

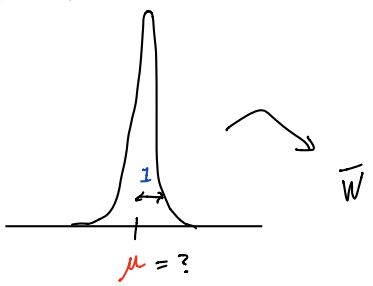
$$sd(\bar{W}) = \frac{sd(W_i)}{\sqrt{n}} = \frac{sd(W_i)}{\sqrt{100}}$$

$$\therefore \bar{W} \approx N\left(\mu, \left(\frac{sd(W_i)}{\sqrt{100}}\right)^2\right) \text{ by the CLT}$$

If $sd(W_i)$ is around \$10 then

$$sd(\bar{W}) \approx \frac{10}{\sqrt{100}} = 1$$

so $\bar{W}$ behaves like a random draw from



So we would expect $\bar{W}$ to be $\mu \pm 1$

Moving things around we "expect" μ to be $\bar{W} \pm 1$

which is 0.91 ± 1 from our sample.

$\therefore$ it is possible that $\mu \leq 0$.

Case 2 : $n = 2000$

$$E(\bar{W}) = E(W_i) = \mu$$

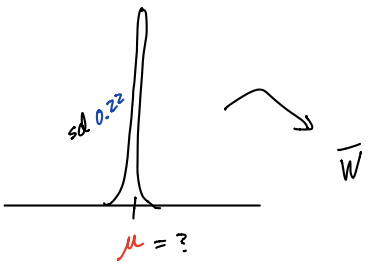
$$sd(\bar{W}) = \frac{sd(W_i)}{\sqrt{n}} = \frac{sd(W_i)}{\sqrt{2000}}$$

$$\therefore \quad \bar{W} \approx N\left(\mu, \left(\frac{sd(W_i)}{\sqrt{2000}}\right)^2\right)$$

if $sd(W_i)$ is around \$10 then

$$sd(\bar{W}) \approx \frac{10}{\sqrt{2000}} = 0.22$$

so $\bar{W}$ behaves like a random draw from



So we would expect $\bar{W}$ to be $\mu \pm 0.22$

Moving things around we "expect" μ to be $\bar{W} \pm 0.22$

which is 0.91 ± 0.22 from our sample.

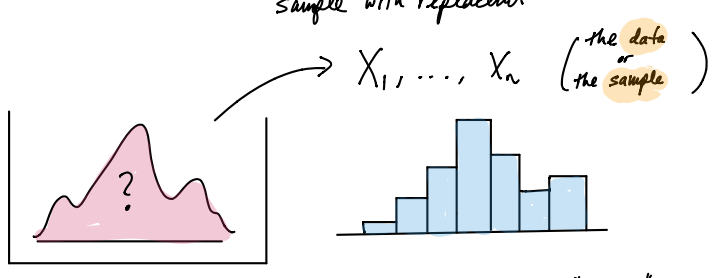
$\therefore$ it is likely that $\mu > 0$.

How likely ... well if $\mu \leq 0$ we just observed a $\bar{W}$ that is more than $\frac{0.91}{0.22} = 4.1$ s.d.s away from what we expect. This happens with probability $\approx P(\underbrace{Z}_{N(0,1)} \geq 4.1) = \underbrace{2 \times 10^{-5}}_{\text{i.e. 2 in 100,000 times.}}$

Basics of estimation

The previous example demonstrated the basics of estimation, confidence intervals, p-values & hypothesis testing (without all the technical fuss).

The general set up is as follows



Some population of numbers. You want to investigate some parameter θ describing the shape of the PMF or PDF

Now use the "shape" of this distribution of numbers to estimate θ, call it $\hat{\theta}$.

$\hat{\theta}$ is a function of the data $X_1, \ldots, X_n$. i.e. $\hat{\theta}(X_1, \ldots, X_n)$.

The main problem is figuring out how wrong $\hat{\theta}$ could be.

The basic tools used for this are MF1, MF2 and the CLT.

Example: estimating a population proportion.

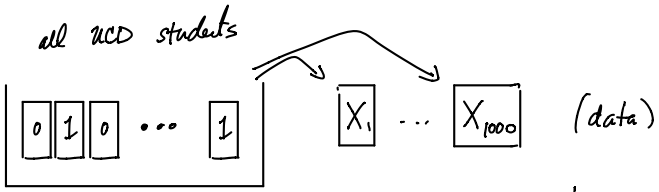
Interview 1000 UCD students

data: 34% said they like stats

What does this tell us about all UCD students?

Could 34% misrepresent the whole campus by as much as ± 20% ?

Here is the box model picture of what the data represents



A ticket for each student

1 means likes stats

0 not likes stats

let θ = proportion of ones in the box.

Let $\hat{\theta}$ = proportion of ones in the data = 0.34

Note $\hat{\theta} = \frac{X_1 + \cdots + X_{1000}}{1000} = \bar{X}$

$\hat{\theta} = 0.34$ from the data

how close do we expect this to be to θ?

$\hat{\theta} - \theta$ = error

$(\hat{\theta} - \theta)^2$ = squared error

$E[(\hat{\theta} - \theta)^2]$ = expected squared error

called $MSE(\hat{\theta})$

$\sqrt{E[(\hat{\theta} - \theta)^2]}$ = root mean squared error

called $RMSE(\hat{\theta})$

$RMSE(\hat{\theta})$ tells you the typical error when using $\hat{\theta}$ to estimate θ.

For this example notice that

$$E(\hat{\theta}) = E\left(\frac{X_1 + \cdots + X_{1000}}{1000}\right)$$

$$\overset{\text{MF1}}{=} E(X_1)$$

$$= 1 \cdot \theta + 0 \cdot (1-\theta)$$

$$= \theta$$

so $E(\hat{\theta}) = \theta$ which means if a "million" other people interviewed 1000 random UCD student and got their own values for $\hat{\theta}$ ... the average value of these $\hat{\theta}$ values would be the true θ.

Another way to say this ...
the estimate $\hat{\theta}$ has no systematic bias.

For estimates that do have bias we can quantify it by

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

For this example $Bias(\hat{\theta}) = 0$

Now

$$RMSE(\hat{\theta}) = \sqrt{E(\hat{\theta} - \theta)^2}$$

$$= \sqrt{E(\hat{\theta} - E(\hat{\theta}))^2}$$

$$= sd(\hat{\theta})$$

$$= sd(\bar{X})$$

$$= \frac{sd(X_1)}{\sqrt{1000}}$$

$$= \frac{\sqrt{\theta(1-\theta)}}{\sqrt{1000}}$$

So the typical error when using $\hat{\theta}$ to estimate θ is

(1) $\dfrac{\sqrt{\theta(1-\theta)}}{\sqrt{1000}}$ ← but this depends on θ which we don't know.

There are two sensible ways to proceed.

(*) plug in $\hat{\theta} = 0.34$ into (1) to get an estimate of (1):

Typical error using $\hat{\theta}$

$$= \frac{\sqrt{\theta(1-\theta)}}{\sqrt{1000}}$$

$$\approx \frac{\sqrt{0.34(1-0.34)}}{\sqrt{1000}} = 0.015$$

$$\therefore \quad \hat{\theta} \approx \theta \pm \frac{\sqrt{\theta(1-\theta)}}{\sqrt{1000}}$$

$$\approx \theta \pm 0.0149$$

$$\therefore \quad \theta \approx \hat{\theta} \pm 0.0149 \approx 0.34 \pm 0.0149$$

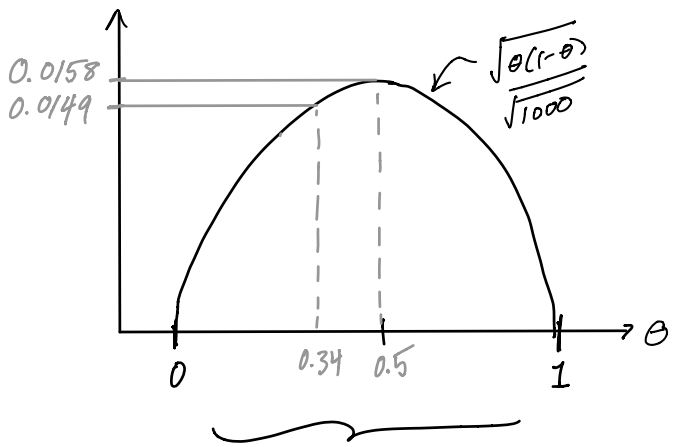
Conclusion:
- There is almost no way the true value of $\theta > 0.4$
- if $\theta > 0.4$ then $\hat{\theta}$ was observed to be more than 4 s.d.s below what we expect, since

$$P(Z < -4) \approx 3 \times 10^{-5}$$

3 times in 100,000

This number is the p-value for testing the null hypothesis: $\theta \geq 0.4$

(**) the conservative approach:



$$\therefore \quad \text{Typical error using } \hat{\theta}$$

$$= \frac{\sqrt{\theta(1-\theta)}}{\sqrt{1000}}$$

$$< \frac{\sqrt{0.5(1-0.5)}}{\sqrt{1000}} = 0.0158$$

$$\therefore \quad \theta \approx 0.34 \pm \underbrace{0.0158}_{\text{conservative error estimate.}}$$

Same conclusion: i.e. the true value of θ is probably

not less that $0.276 = 0.34 - 3\times(0.0158)$

&

not greater than $0.403 = 0.34 + 3\times(0.0158)$

The interval $(0.276, 0.403)$ is a 99.7% confidence interval for θ.

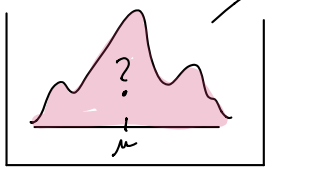
Comparing two populations

The basic reasoning used in the two previous examples works in more complicated settings.

Here is an example that tests the difference btwn two treatments.

Example:

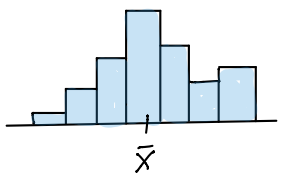
- Suppose you have developed a new drug for lowering cholesterol and want to test if it effective
- You get a random sample of 250 people and give them the drug, then measure the % reduction of cholesterol after 6 months.
  Let $X_1, X_2, \dots, X_{250}$ denote the % reduction for each patient.



Distribution of % reduction in cholesterol if the new drug is given to everyone.

$\mu = ?$
$\sigma = ?$

$X_1, \dots, X_{250}$



The % reduction in the sample

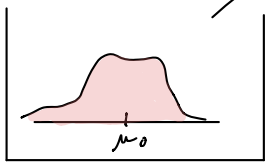
$\bar{X} = 7.2\%$

$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} = 4.1\%$

This might suggest $\mu > 0$ but we need to rule out a placebo effect

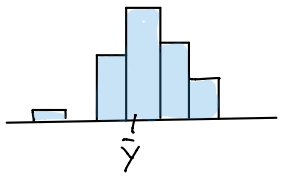
- To account for the Placebo effect you Sample 500 people (called the control group) and give them a sugar pill.
  Let $Y_1, Y_2, \dots, Y_{500}$ denote the % reduction in the control group after 6 months

$Y_1, Y_2, \dots, Y_{500}$



Distribution of % reduction in cholesterol when taking a sugar pill

$\mu_0 = ?$
$\sigma_0 = ?$



The % reduction in the sample

$\bar{Y} = 3.9\%$

$\hat{\sigma}_0 = \sqrt{\frac{1}{500}\sum_{i=1}^{500}(Y_i - \bar{Y})^2} = 2.5\%$

Question 1) Give a range of plausible values for $\mu - \mu_0$?

Answer: start with an estimate of $\mu - \mu_0$.

$\bar{X} - \bar{Y}$ is a reasonable estimate of $\mu - \mu_0$.

Note: $E(\bar{X} - \bar{Y}) \overset{MF1}{=} E(\bar{X}) - E(\bar{Y})$

$= \mu - \mu_0$

$\therefore \text{Bias}(\bar{X} - \bar{Y}) = 0$

Also the X's are independent of the Y's so that

$$MSE(\bar{X}-\bar{Y}) = E\left[(\bar{X}-\bar{Y}-\overbrace{(\mu-\mu_0)}^{E(\bar{X}-\bar{Y})})^2\right] = var(\bar{X}-\bar{Y})$$

$$\overset{MF2}{=} var(\bar{X}) + var(\bar{Y}) \quad \text{by independence}$$

$$= \frac{\sigma^2}{250} + \frac{\sigma_0^2}{500}$$

$$\approx \frac{\hat{\sigma}^2}{250} + \frac{\hat{\sigma}_0^2}{500}$$

$$= \frac{4.1^2}{250} + \frac{2.5^2}{500}$$

$$= 0.0797$$

$\therefore$ The typical error when using $\bar{X}-\bar{Y}$ to estimate $\mu-\mu_0$ is about $\sqrt{0.0797} = 0.282 \approx RMSE(\bar{X}-\bar{Y})$.

i.e. $\bar{X}-\bar{Y} \approx \mu-\mu_0 \pm 2(0.282)$ approximately 95% of the time

i.e. $\mu-\mu_0 \approx \underbrace{\bar{X}-\bar{Y}}_{\text{we got } 7.2-3.9=3.3 \text{ for this value}} \pm 2(0.282)$ approximately 95% of the time

i.e. $\mu-\mu_0 \in \underbrace{(2.735, 3.865)}_{3.3 \pm 2(0.282)}$ with 95% confidence.

Also

$\mu-\mu_0 \in \underbrace{(2.452, 4.147)}_{3.3 \pm 3(0.282)}$ with 99% confidence

So most likely $\mu > \mu_0$, i.e. the drug works better than a placebo

Question 2)

Quantify the amount of evidence that $\mu-\mu_0 > 0$ from the data.

Answer: if it was actually the case that $\mu-\mu_0 \leq 0$ (i.e. $\mu \leq \mu_0$) then we just observed $\bar{X}-\bar{Y}$ to have a z-score greater than 11.74

since if $\mu-\mu_0 \leq 0$ then

$$z = \frac{\bar{X}-\bar{Y}-(\mu-\mu_0)}{\sqrt{\frac{\sigma^2}{250}+\frac{\sigma_0^2}{500}}} > \frac{3.3-0}{\sqrt{\frac{\sigma^2}{250}+\frac{\sigma_0^2}{500}}} \approx \frac{3.3}{\underbrace{\sqrt{0.0797}}_{=11.74}}$$

The probability of that happening is

$$P(Z > 11.74) = \underbrace{3.6 \times 10^{-32}}_{\text{The p-value for testing the null hypothesis } \mu-\mu_0 \leq 0.}$$

$\therefore$ the data gives conclusive evidence that $\mu-\mu_0 > 0$.

Bias – Variance trade off

Another simple looking but deep formula:

For any estimate $\hat{\theta}$

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

$var(\hat{\theta})$: intrinsic sampling variability.

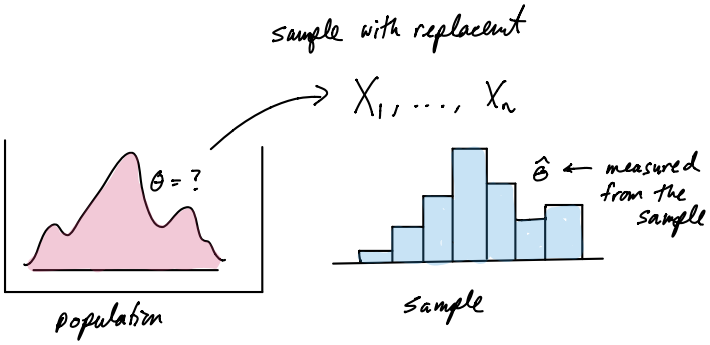
$Bias(\hat{\theta})$: systematic error

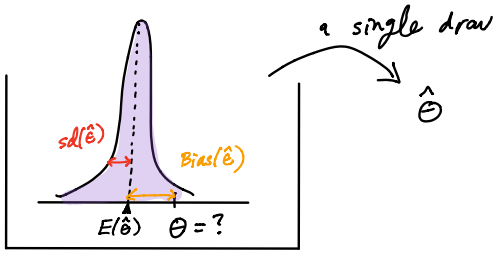
$\therefore$ If $Bias(\hat{\theta}) = 0$ then

$$MSE(\hat{\theta}) = var(\hat{\theta}) \quad \text{and}$$

$$RMSE(\hat{\theta}) = sd(\hat{\theta})$$

Also this decomposition gives a nice geometric picture:

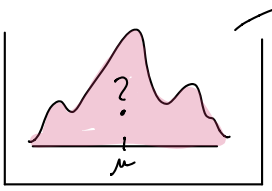




Box model for the randomness in $\hat{\theta}$

$$RMSE(\hat{\theta}) = \sqrt{sd(\hat{\theta})^2 + bias(\hat{\theta})^2}$$

Example:



Population of numbers

$\mu = ?$

$\sigma = ?$

X_1, X_2, X_3

3 samples with replacement

Consider two estimates of μ:

$$\hat{\mu}_1 = \frac{X_1 + X_2 + X_3}{3}$$

$$\hat{\mu}_2 = \frac{X_1 + X_2 + X_3}{2}$$

Which estimate has a smaller MSE?

$$bias(\hat{\mu}_1) = E(\hat{\mu}_1) - \mu \overset{MF1}{=} 0$$

$$bias(\hat{\mu}_2) = E(\hat{\mu}_2) - \mu$$

$$= E\left(\frac{X_1 + X_2 + X_3}{2}\right) - \mu$$

$$\overset{MF1}{=} \frac{3\mu}{2} - \mu = \frac{\mu}{2}$$

$$var(\hat{\mu}_1) \overset{MF2}{=} \frac{\sigma^2}{3} \quad \text{where } \sigma^2 = var(X_1)$$

$$var(\hat{\mu}_2) = var\left(\frac{X_1 + X_2 + X_3}{2}\right) = \frac{3\sigma^2}{4}$$

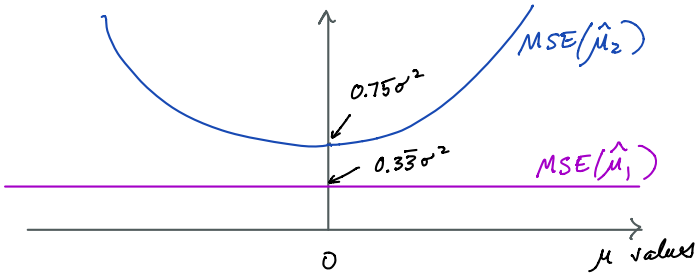
$$\therefore MSE(\hat{\mu}_1) = var(\hat{\mu}_1) + bias(\hat{\mu}_1)^2$$

$$= \frac{\sigma^2}{3}$$

$$\& \quad MSE(\hat{\mu}_2) = var(\hat{\mu}_2) + bias(\hat{\mu}_2)^2$$

$$= \frac{3\sigma^2}{4} + \left(\frac{\mu}{2}\right)^2$$

Here is a plot of the values of $MSE(\hat{\mu}_1)$ and $MSE(\hat{\mu}_2)$ as a function of the unknown μ:



So, no matter what μ is we expect $\hat{\mu}_1$ to be closer to μ than $\hat{\mu}_2$.

Now consider a 3rd estimate

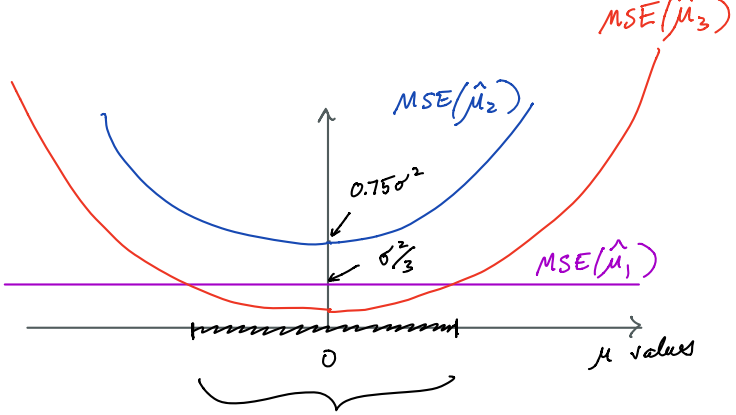
$$\hat{\mu}_3 = \frac{X_1 + X_2 + X_3}{6}$$

MF1 & MF2 gives

$$\text{bias}(\hat{\mu}_3) = -\frac{\mu}{2}$$ ← i.e $\hat{\mu}_3$ tends to be too small in magnitude

$$\text{var}(\hat{\mu}_3) = \frac{1}{12}\sigma^2$$ ← smaller variance that both $\hat{\mu}_1$ & $\hat{\mu}_2$

$$\therefore MSE(\hat{\mu}_3) = \frac{\sigma^2}{12} + \frac{\mu^2}{4}$$



So if μ happens to be in this region then $\hat{\mu}_3$ has smaller MSE.

... however if μ is not in this interval then it will be much less accurate.

$\therefore$ if you happen to have insider information that μ is in the shaded region, then use $\hat{\mu}_3$ ... but without that info $\hat{\mu}_2$ is the safest bet.

Example:

Inverse variance weighted averaging

Suppose you are trying to estimate an unknown number μ from 3 random samples X_1, X_2, X_3 where

$$X_1 \sim N(\mu, 4)$$
$$X_2 \sim N(\mu, 16)$$
$$X_3 \sim N(\mu, 8)$$

Notice that if $\hat{\mu}$ is defined as

$$\hat{\mu} = \frac{X_1 + X_2 + X_3}{3} = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_4$$

Then the large variance of X_2 could make the MSE large.

A more general class of estimators of μ are found by **weighted averaging**

$$\hat{\mu} = w_1 X_1 + w_2 X_2 + w_3 X_3$$

where w_1, w_2, w_3 are non-negative and $w_1 + w_2 + w_3 = 1$

So by choosing w_2 small you can down weight X_2.

Note the requirment

$$w_1 + w_2 + w_3 = 1$$

ensures the estimate has zero bias:

$$\begin{aligned}\text{bias}(\hat{\mu}) &= E(\hat{\mu}) - \mu \\ &= E(w_1X_1 + w_2X_2 + w_3X_3) - \mu \\ &= w_1\mu + w_2\mu + w_3\mu - \mu \\ &= \underbrace{(w_1 + w_2 + w_3)}_{=1}\mu - \mu \\ &= 0\end{aligned}$$

$\therefore$ $MSE(\hat{\mu}) = \text{var}(\hat{\mu})$

and the weights w_1, w_2, w_3 which give the smallest MSE are proportional to the inverse variance:

$$w_1 = \frac{c}{\text{var}(X_1)}, \quad w_2 = \frac{c}{\text{var}(X_2)}, \quad w_3 = \frac{c}{\text{var}(X_3)}$$

where $$c = \frac{1}{\left(\frac{1}{\text{var}(X_1)} + \frac{1}{\text{var}(X_2)} + \frac{1}{\text{var}(X_3)}\right)}.$$

$$= \frac{1}{\left(\frac{1}{4} + \frac{1}{16} + \frac{1}{8}\right)} = \frac{16}{7}$$

$\therefore$ The "Best weights" are given by

$$\left.\begin{aligned} w_1 &= \frac{16/7}{4} = \frac{4}{7} \\ w_2 &= \frac{16/7}{16} = \frac{1}{7} \\ w_3 &= \frac{16/7}{8} = \frac{2}{7} \end{aligned}\right\} \quad \hat{\mu} = \frac{4}{7}X_1 + \frac{1}{7}X_2 + \frac{2}{7}X_3$$

with

$$MSE(\hat{\mu}) = \frac{c^2}{\text{var}(X_1)^2}\text{var}(X_1) + \frac{c^2}{\text{var}(X_2)^2}\text{var}(X_2) + \frac{c^2}{\text{var}(X_3)^2}\text{var}(X_3)$$

$$= c^2\underbrace{\left(\frac{1}{\text{var}(X_1)} + \frac{1}{\text{var}(X_2)} + \frac{1}{\text{var}(X_3)}\right)}_{1/c}$$

$$= c = \frac{16}{7}$$

Lets check this gives better MSE than $\hat{\mu} = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3$

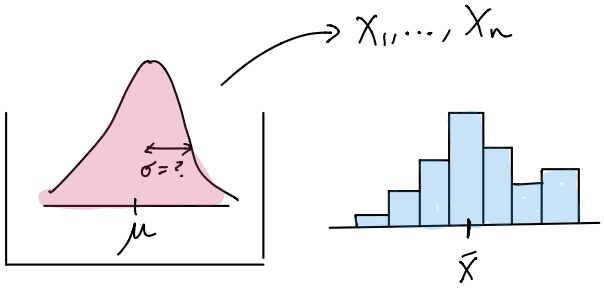
$$MSE(\hat{\mu}) = \frac{1}{9}4 + \frac{1}{9}16 + \frac{1}{9}8$$

$$= \frac{28}{9} \approx 3.1 > 2.28 \approx \frac{16}{7}$$

Example:

Suppose $X_1, X_2, \dots, X_n$ are independent samples from $N(\mu, \sigma^2)$

Picture:



Let $\hat{\sigma}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2$

resembles

$\text{var}(X) = E(X^2) - (E(X))^2$

Lets investigate the typical error using $\hat{\sigma}^2$ to estimate σ^2

$$E(\hat{\sigma}^2) \overset{MF1}{=} \left(\frac{1}{n}\sum_{i=1}^{n} E(X_i^2)\right) - E(\bar{X}^2)$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} (\sigma^2 + \mu^2)\right) - \left(\frac{\sigma^2}{n} + \mu^2\right)$$

since ...

$\text{var}(X_i) = E(X_i^2) - (E(X))^2$

σ^2 ↑ what we want ↑ μ^2

$$= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2$$

$$= \left(1 - \frac{1}{n}\right)\sigma^2$$ ← i.e. $\hat{\sigma}^2$ will systematically be somewhat too small.

$\therefore$ $\text{Bias}(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2$

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

$$= -\frac{\sigma^2}{n}$$ ← the typical size and direction of the systematic error.

This part will not covered on Midterm 3!

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

⋮ ↝ for experts ... use that

$\frac{n}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-1}$

and $\text{var}(\chi^2_{n-1}) = 2(n-1)$

$$= 2\sigma^4\left(\frac{n-1}{n^2}\right)$$

$\therefore$ $MSE(\hat{\sigma}^2) = \text{var}(\hat{\sigma}^2) + \text{Bias}(\hat{\sigma}^2)^2$

$$= 2\sigma^4\left(\frac{n-1}{n^2}\right) + \frac{\sigma^4}{n^2}$$

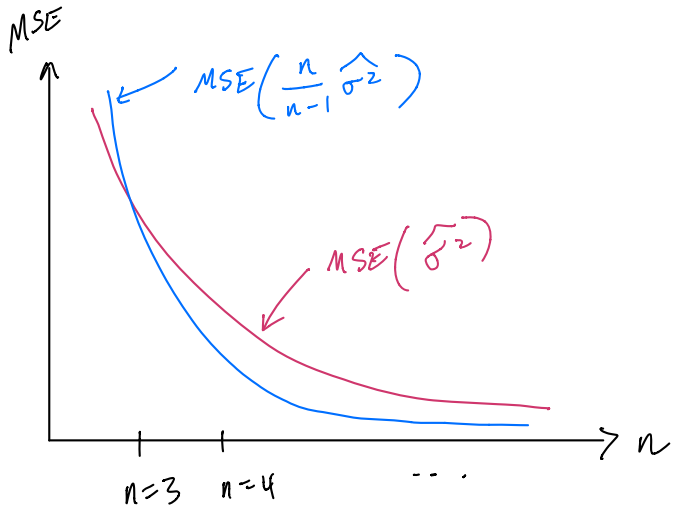
If we instead estimate σ^2 with $\frac{n}{n-1}\hat{\sigma}^2$ we get an unbiased estimate (i.e. $\text{Bias}\left(\frac{n}{n-1}\hat{\sigma}^2\right) = 0$) and

$$MSE\left(\frac{n}{n-1}\hat{\sigma}^2\right) = \text{var}\left(\frac{n}{n-1}\hat{\sigma}^2\right) + 0$$

$$= \left(\frac{n}{n-1}\right)^2 \text{var}(\hat{\sigma}^2)$$

$$= \left(\frac{n}{n-1}\right)^2 2\sigma^4\left(\frac{n-1}{n^2}\right)$$

$$= \frac{2\sigma^4}{n-1}$$



So for $n > 3$ $\frac{n}{n-1}\hat{\sigma}^2$ is typically closer to σ^2 than $\hat{\sigma}^2$