$Y_1, \ldots Y_n \sim N(\mu, \sigma^2)$, σ^2 is known.

$\mu \sim N(\mu_0, \frac{\sigma^2}{k_0})$ (Prior)

prior parameters

Bayes Estimators

Posterior

$$P(\mu | y_1, \ldots y_n, \sigma^2) \sim N(\mu_n, \sigma_n^2)$$

$$\mu_n = w\bar{y} + (1-w)\mu_0, \quad w = \frac{n}{n+k_0}$$

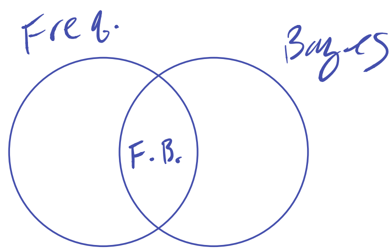
$$\sigma_n^2 = \frac{\sigma^2}{n+k_0} = w\frac{\sigma^2}{n}$$

$$E[\mu | y_1, \ldots y_n, \sigma^2] = w\bar{y} + (1-w)\mu_0$$ (estimate)

lowercase

What are the frequentist properties of $E[\mu | Y_1, \ldots Y_n, \sigma^2]$ estimator!!

Capital



$$w\bar{Y} + (1-w)\mu_0$$

Estimators: Bayes / Frequentist Unification

- Bayesian inference provides a straightforward procedure for producing estimators given your prior beliefs.

    1. Compute posterior distribution

    2. Summarize the posterior distribution with a point estimator (e.g. posterior mean or posterior mode) and a probability interval

- Frequentists provide tools for evaluating the sampling properties of an estimator.

    - Bias, variance and MSE of an estimator
    - Well-calibrated probability intervals

- Both are useful!

The Bias-Variance Tradeoff

Reminder: an estimator is a random variable, an estimate is a constant

- *Bias*: systematic sampling error of the estimator
- *Variance*: variance of the estimator (from sampling & measurement error)
- Often we evaluate an estimator in terms of mean square error: $\text{MSE}(\hat{\theta}) = E_Y(\hat{\theta} - \theta)^2$
- The Bias-Variance tradeoff: $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$

The Bias-Variance Tradeoff

- Variance of an estimator comes sampling from a population
  - If you were to repeatedly draw new samples of the same size how much would your estimates vary?
  - e.g. if $y_i \sim N(\mu, \sigma^2)$ then $\mathrm{Var}(\bar{Y}) = \sigma^2/n$

$$\hat{\mu}_{MLE} = \bar{Y}$$

$$E[\bar{Y}] = \mu \quad \text{(unbiased)}$$

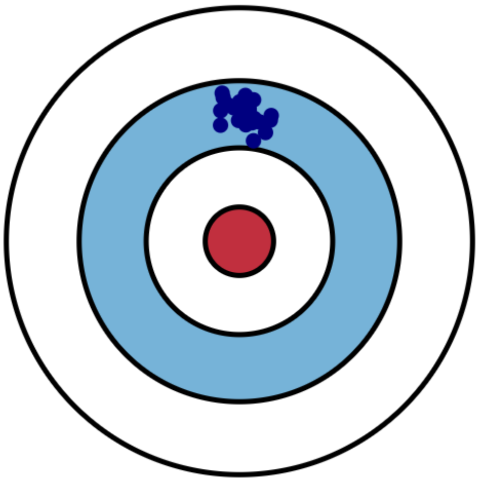
$$\mathrm{Var}(\bar{Y}) = \mathrm{Var}\left(\frac{1}{n}\sum Y_i\right) = \frac{1}{n^2}\sum \mathrm{Var}(Y_i)$$

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

$$MSE(\hat{\mu}_{MLE}) = \frac{\sigma^2}{n}$$

Bias

The expected difference between the estimate and the response

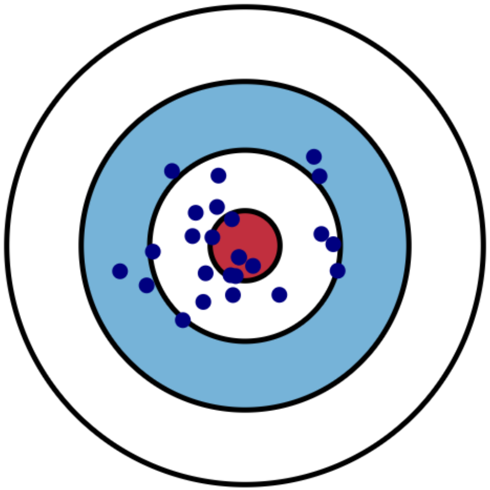


Statistical definition of bias:

$$E_Y[\hat{\theta} - \theta]$$

Variance

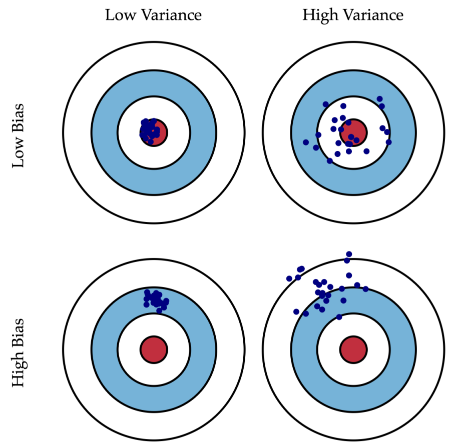
How variable is the prediction about its mean?



Statistical definition of variance:

$$E_Y[\hat{\theta} - E_Y[\hat{\theta}]]^2$$

Bias and Variance



$$\underbrace{\mathrm{MSE}(\hat{\theta})}_{\text{accuracy}} = \underbrace{\mathrm{Var}(\hat{\theta})}_{\text{variance}} + \underbrace{\left(E[\hat{\theta} - \theta]\right)^2}_{\text{bias squared}}$$

The Bias-Variance Tradeoff

- The prior distribution (usually) makes your estimator biased...
- But the prior distribution also (usually) reduces the variance!
- Example: compute the frequentist mean and variance of the posterior mean.

$$E[\mu \mid Y_1, \ldots Y_n, \sigma^2] = \hat{\mu}_{PM} = w\bar{Y} + (1-w)\mu_0$$

Bias:

$$E[w\bar{Y} + (1-w)\mu_0 - \mu] = wE[\bar{Y}] + (1-w)E[\mu_0] - \mu$$

$$= w\mu + (1-w)\mu_0 - \mu = (1-w)(\mu_0 - \mu)$$

Var:

$$Var(\mu \mid Y_1, \ldots Y_n, \sigma^2) = Var(w\bar{Y} + (1-w)\mu_0) =$$

$$= w^2 Var(\bar{Y}) + (1-w)^2 \underbrace{\cancel{Var(\mu_0)}}_{= 0 \text{ (const)}}$$

$$= w^2 Var(\bar{Y}) \leq Var(\bar{Y}) = Var(\hat{\mu}_{MLE})$$

Example: IQ scores

sd

- Scoring on IQ tests is designed to yield a N(100, 15) distribution for the general population

- We observe IQ scores for a sample of n individuals from a particular town and estimate μ, the town-specific IQ score

  I.V.

- If we lacked knowledge about the town, a natural choice would be $\mu_0 = 100$

- Suppose the true parameters for this town are $\mu = 112$ and $\sigma = 13$

  - The town is smarter on average than the general population

How does $\bar{Y}$ compare to

$$\hat{\mu}_{PM} = w\bar{Y} + (1-w)100 \text{ ??}$$

Example: IQ scores

- What is the mean squared error of the MLE? MSE of the posterior mean?

$\sigma = 13$

- $\text{MSE}[\hat{\mu}_{MLE}] = \text{Var}[\hat{\mu}_{MLE}] = \frac{\sigma^2}{n} = \frac{169}{n}$

- $\text{MSE}[\hat{\mu}_{PM}|\theta_0] = w^2 \frac{169}{n} + (1 - w)^2 144$

- Reminder: $w = \frac{n}{\kappa_0 + n}$. For what values of n and κ_0 is the MSE smaller for the posterior mean estimator than the maximum likelihood?

Bias^2

$$\frac{169}{n} \overset{?}{\gtrless} w^2 \frac{169}{n} + (1-w)^2 144$$

$\text{MSE}(\hat{\mu}_{MLE})$ $\quad$ $\text{MSE}(\hat{\mu}_{PM})$

Bias : $(1-w)(100-112)$

$\Rightarrow \text{Bias}^2 = (1-w)^2 144$.

Example: IQ scores

$\dfrac{MSE(\hat{\mu}_{PM})}{MSE(\hat{\mu}_{MLE})}$

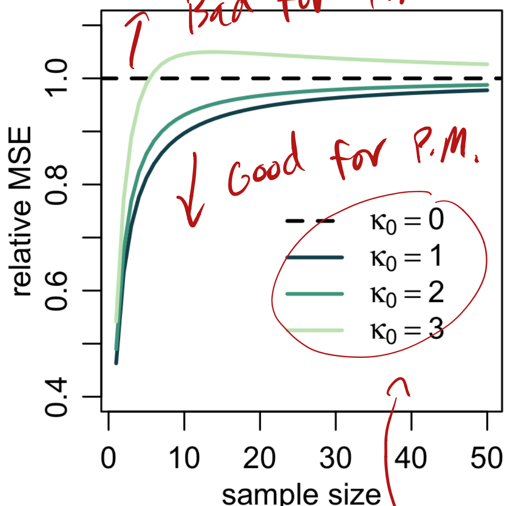
$w = \dfrac{n}{n+\kappa_0}$

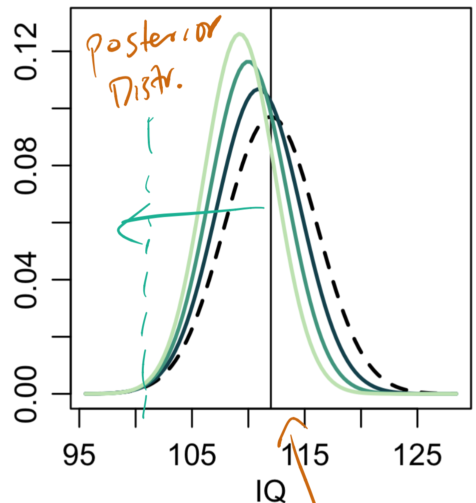
$n \to \infty, \; w \to 1$

e.g. $\hat{\mu}_{P.M} \to \hat{\mu}_{MLE}$

↑ Bad for P.M.

↓ Good for P.M.





Truth.

Strength of prior knowledge (# of pseudo-obs, determines w)

MSE_{PM}:

$w\,\dfrac{169}{n} + (1-w)^2\, 144$

Decision Theory

Why the posterior mean?

- Often times we need to make a "decision" by providing a single estimate
- The posterior provides a full distribution over θ, which can be summarized in infinitely many ways
- Specify a *loss function* which describes the cost of estimating $\hat{\theta}$ when the truth is θ $L(\hat{\theta}, \theta)$

Bayes Estimators

Penalize the difference between estimate & truth.

- The *loss function*: $L(\hat{\theta}, \theta)$
  - Squared error: $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$
  - Absolute error: $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$
- The **Bayes risk** is the posterior expected loss:

$$E_{\theta|y}[L(\hat{\theta}, \theta)] = \int L(\hat{\theta}, \theta)p(\theta \mid y)d\theta$$

Monte Carlo

1. Sample θ^s from $P(\theta | y)$
2. $L(\hat{\theta}, \theta^s)$

Average all losses.

- Choose an estimator of θ based on minimizing the Bayes risk.
- An estimator $\hat{\theta}$ is said to be a **Bayes estimator** if it minimizes the Bayes risk among all estimators.

Squared error loss :

What $\hat{\theta}$ should I choose?

$$\min_{\hat\theta} E_{\theta|y}(\hat\theta-\theta)^2 = \min_{\hat\theta} \int (\hat\theta-\theta)^2 p(\theta \mid y)d\theta$$

Differentiate with respect to $\hat\theta$ and set equal to zero:

$$\frac{d}{d\hat\theta}\int(\hat\theta-\theta)^2 p(\theta|y)d\theta =$$

Do This

$$\int \frac{d}{d\hat\theta}(\hat\theta-\theta)^2 p(\theta|y)d\theta =$$

$$\int 2(\hat\theta-\theta)p(\theta|y)d\theta = 0$$

$$\cancel{2}\int \hat\theta\, p(\theta|y)d\theta = \cancel{2}\int \theta\, p(\theta|y)d\theta$$

$$\hat{\theta} \underbrace{\int P(\theta|y)\,d\theta}_{1} = \int \theta P(\theta|y)\,d\theta$$

$$\hat{\theta} = E[\theta|y]$$

Posterior Mean Definition.

Absolute loss

$$\min_{\hat\theta} E_{\theta|y}|\hat\theta-\theta| = \min_{\hat\theta}\int|\hat\theta-\theta|p(\theta\mid y)d\theta$$

Differentiate with respect to $\hat\theta$ and set equal to zero:

$$\frac{d}{d\hat\theta}\int|\hat\theta-\theta|p(\theta|y)d\theta =$$

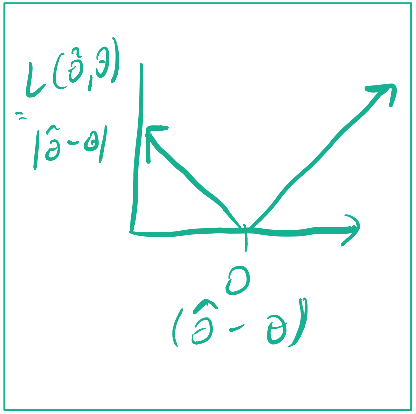
$$\int\frac{d}{d\hat\theta}|\hat\theta-\theta|p(\theta|y)d\theta =$$

$$\int \text{sign}(\hat\theta-\theta)p(\theta|y)d\theta =$$

$\hookrightarrow$ -1 if $\hat\theta \le \theta$

$+1$ if $\hat\theta \ge \theta$

Aside



$\text{Sign}(x) = +1$ if $x>0$

-1 if $x<0$

$$\underbrace{\int_{\theta=-\infty}^{\theta=\hat{\theta}} +1\, P(\theta|y)\, d\theta}_{\text{This is just a CDF!!}} + \int_{\theta=\hat{\theta}}^{\theta=+\infty} -1\, P(\theta|y)\, d\theta$$

$$P(\theta < \hat{\theta} | y) - (1 - P(\theta < \hat{\theta} | y)) = 0$$

$$2P(\theta < \hat{\theta} | y) = 1$$

$$P(\theta < \hat{\theta} | y) = 1/2$$

$\hat{\theta}$ is the posterior median

Bayes Estimator for absolute loss.

Loss functions in practice

- Squared error and absolute error are good default loss functions
    - Motivated largely by mathematical considerations
- In practice we should define a loss function specific to our problem
- Loss in dollars? Loss in "quality of life"?

Decision making: flu example

- The CDC produces estimates of the expected prevalance and severity of flu during flu season
- Assume θ represents severity of the flu

e.g. Poisson rate

- $p(\theta \mid y)$ is CDC posterior distribution based on initial data about the upcoming flu season
- $\hat{\theta}$ determines how much flu vaccine to make. How do we determine $L(\hat{\theta}, \theta)$?

+ Shelf life of vaccine
+ Too little = death
+ Too much is wasted money.