7. Point estimation

- Binomial distribution
- Measuring and modeling uncertainty
- Gaussian distribution
- Biased vs unbiased estimators

Many slides borrowed from Prof. Minh Hoai Nguyen's previous course offering

Binomial distribution

Fruit inspection

Good!

bad!

Bernoulli model:

$$\Pr(\text{good strawberry}) = \theta \in [0, 1]$$

Fruit quality estimate

- Bernoulli model:

$$\Pr(\text{good berry}) = \theta \in [0, 1]$$

- Sample data i.i.d. $\rightarrow$ Binomial distribution

$$\Pr(\{\text{berries}\}|\theta) = \prod_k \Pr(\{\text{berry } k\}|\theta) = \theta^{\#\text{ good berries}}(1-\theta)^{\#\text{ bad berries}}$$

- **Key assumption:** Quality of berry i doesn't depend on berry j
  - Is this assumption reasonable?
  - When is i.i.d. not reasonable?

Point estimate

Data

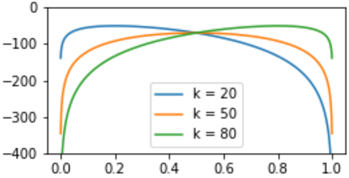
$$\mathcal{D} = \{\text{good, bad, good } \ldots\} \;\rightarrow\; \underbrace{k \text{ good berries, } m-k \text{ bad berries}}_{\text{summary statistics}}$$

Log likelihood

$$\log(\underbrace{\Pr(\mathcal{D}|\theta)}_{\text{likelihood}}) = \log\left(\theta^k(1-\theta)^{m-k}\right) = k\log(\theta) + (m-k)\log(1-\theta)$$

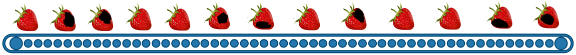
Maximum likelihood solution

$$\theta_{\mathbf{MLE}} := \underset{\theta}{\operatorname{argmax}}\ \log(\Pr(\mathcal{D}|\theta)) \overset{\text{set derivative to } 0}{=} \frac{k}{m \cancel{-k}}$$



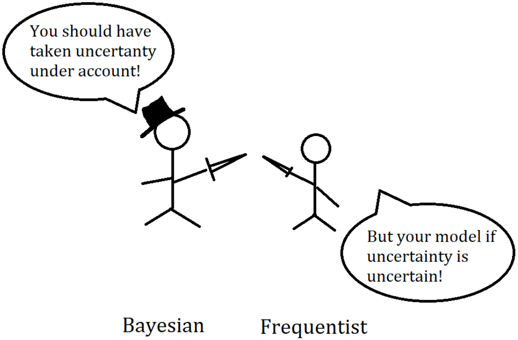
Measuring and modeling uncertainty

Example: strawberry inspection



k		$\hat{\theta}$
1	bad	0
2	bad	0
3	good	$1/3$
4	good	$2/4$
5	bad	$2/5$
6	good	$3/6$
7	good	$4/7$
8	good	$5/8$
9	good	$6/9$
10	good	$7/10$

estimate gets better with more data



How good is point estimate?

For $x_1, ..., x_m$ sampled i.i.d., with $0 \le x_i \le 1$,

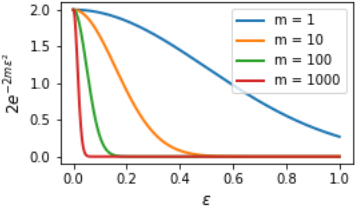
$$\Pr\left(\frac{1}{m}\sum_{i=1}^{m} x_i - \mathbb{E}[X] \ge t\right) \le e^{-2mt^2} \qquad \text{(Hoeffding's inequality)}$$

Point estimate of true θ^*:

$$\hat{\theta}_m = \frac{\text{\# good berries}}{\text{\# total berries}} = \frac{k}{m}$$

Hoeffding:

$$\Pr(|\hat{\theta}_m - \theta^*| > \epsilon) \le 2e^{-2m\epsilon^2}$$



Point estimate of Binomial distribution

$$\Pr(|\hat{\theta}_m - \theta^*| > \epsilon) \leq 2e^{-2m\epsilon^2} \leq \delta \iff m \geq \underbrace{\frac{\log(2/\delta)}{2\epsilon^2}}_{=\text{poly}(1/\epsilon, 1/\delta)}$$

Provably approximately correct (PAC)

We say a task is PAC-learnable if, after N observations,

$$\Pr(|\text{guess} - \text{truth}| < \epsilon) \geq 1 - \delta$$

where $N = \text{poly}(\frac{1}{\epsilon}, \frac{1}{\delta})$ (at most polynomial in $\frac{1}{\epsilon}, \frac{1}{\delta}$).

Gaussian distribution

Extension to Gaussian distribution

High school statistics: fitting a bell curve

student	score
a	99
b	87
c	56
d	88
e	74
f	61
g	85
h	78

- Sample mean

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

- Sample variance

$$\hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{\mu})^2$$

Bell curve fit

$$p(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}\right)$$

and $\mathrm{Pr}(\text{score } < x) = \int_{-\infty}^{x} p(x)dx$

Do these point estimates make sense?

$$p(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}\right)$$

- Neg. Log likelihood

$$-\log\ p(\mathcal{D}|\mu,\sigma^2) = \underbrace{\sum_{i=1}^{m} \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{m}{2}\log(1/\sigma^2) + \frac{m}{2}\log(2\pi)}_{\text{convex in } \mu}$$

- Maximum likelihood mean

$$0 = \frac{\partial \log\ p(\mathcal{D}|\mu,\sigma^2)}{\partial \mu} =$$

Therefore,

$$\mu_{\mathbf{MLE}} =$$

Do these point estimates make sense?

$$p(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}\right)$$

- Neg. Log likelihood

$$-\log\ p(\mathcal{D}|\mu,\sigma^2) = \underbrace{\sum_{i=1}^{m} \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{m}{2}\log(1/\sigma^2) + \frac{m}{2}\log(2\pi)}_{\text{convex in } \mu}$$

- Maximum likelihood mean

$$0 = \frac{\partial \log\ p(\mathcal{D}|\mu,\sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i - \mu)$$

Therefore,

$$\mu_{\mathbf{MLE}} =$$

Do these point estimates make sense?

$$p(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}\right)$$

- Neg. Log likelihood

$$-\log\ p(\mathcal{D}|\mu,\sigma^2) = \underbrace{\sum_{i=1}^{m} \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{m}{2}\log(1/\sigma^2) + \frac{m}{2}\log(2\pi)}_{\text{convex in } \mu}$$

- Maximum likelihood mean

$$0 = \frac{\partial \log\ p(\mathcal{D}|\mu,\sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i-\mu)$$

Therefore,

$$\mu_{\mathbf{MLE}} = \frac{1}{m}\sum_{i=1}^{m} x_i \quad \text{(sample mean)}$$

Do these point estimates make sense?

$$p(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}\right)$$

- Neg. Log likelihood

$$-\log\ p(\mathcal{D}|\mu,\sigma^2) = \underbrace{\sum_{i=1}^{m} \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{m}{2}\log(1/\sigma^2) + \frac{m}{2}\log(2\pi)}_{\text{convex in } 1/\sigma^2}$$

- Maximum likelihood inverse variance

$$0 = \frac{\partial \log\ p(\mathcal{D}|\mu,\sigma^2)}{\partial(1/\sigma^2)} =$$

Therefore,

$$\sigma^2_{\mathbf{MLE}} =$$

Do these point estimates make sense?

$$p(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}\right)$$

- Neg. Log likelihood

$$-\log\ p(\mathcal{D}|\mu,\sigma^2) = \underbrace{\sum_{i=1}^{m} \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{m}{2}\log(1/\sigma^2) + \frac{m}{2}\log(2\pi)}_{\text{convex in } 1/\sigma^2}$$

- Maximum likelihood inverse variance

$$0 = \frac{\partial \log\ p(\mathcal{D}|\mu,\sigma^2)}{\partial(1/\sigma^2)} = \frac{1}{2}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{m}{2}\sigma^2$$

Therefore,

$$\sigma^2_{\mathbf{MLE}} =$$

Do these point estimates make sense?

$$p(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}\right)$$

- Neg. Log likelihood

$$-\log\ p(\mathcal{D}|\mu,\sigma^2) = \underbrace{\sum_{i=1}^{m} \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{m}{2}\log(1/\sigma^2) + \frac{m}{2}\log(2\pi)}_{\text{convex in } 1/\sigma^2}$$

- Maximum likelihood inverse variance

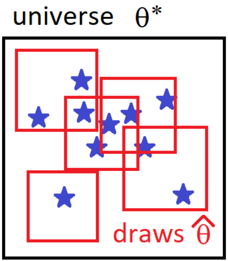
$$0 = \frac{\partial \log\ p(\mathcal{D}|\mu,\sigma^2)}{\partial(1/\sigma^2)} = \frac{1}{2}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{m}{2}\sigma^2$$

Therefore,

$$\sigma^2_{\mathbf{MLE}} = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu)^2 \quad \text{(sample variance)}$$

Biased vs unbiased estimators

Biased estimates



Definition: $\hat{\theta}$ is an <u>unbiased estimate</u> of θ^* if $\mathbb{E}_{\mathcal{D}}[\hat{\theta}] = \theta^*$

Sample mean $\hat{\mu}_{\mathcal{D}}$ is unbiased

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\hat{\mu}_{\mathcal{D}}] \quad &= \quad \mathbb{E}_{\mathcal{D}}\left[\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} x\right] \\
&\overset{\text{chain rule}}{=} \quad \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{x|\mathcal{D}}\left[\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} x\Big|\mathcal{D}\right]\right] \\
&= \quad \mathbb{E}_{\mathcal{D}}\left[\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} \underbrace{\mathbb{E}_{x|\mathcal{D}}[x]}_{\text{i.i.d., } =\mu}\right] \\
&= \quad \mu
\end{aligned}
$$

Sample variance

$$\mathbb{E}_{\mathcal{D}}[\hat{\sigma}^2_{\mathcal{D}}] = \mathbb{E}_{\mathcal{D}}\Bigg[\underbrace{\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}(x-\hat{\mu}_{\mathcal{D}})^2}_{=:A}\Bigg]$$

Decompose:

$$\begin{aligned}
A &= \frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}(x-\hat{\mu}_{\mathcal{D}}+\mu-\mu)^2 \\
&= \frac{1}{|\mathcal{D}|}\left(\sum_{x\in\mathcal{D}}\left((x-\mu)^2+(\mu-\hat{\mu}_{\mathcal{D}})^2\right)+2\sum_{x\in\mathcal{D}}(x-\mu)(\mu-\hat{\mu}_{\mathcal{D}})\right) \\
&= \underbrace{\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}(x-\mu)^2}_{\text{variance of } x} - \underbrace{(\mu-\hat{\mu}_{\mathcal{D}})^2}_{\text{variance of } \hat{\mu}_{\mathcal{D}}}
\end{aligned}$$

Variance of sample mean

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[(\hat{\mu}_{\mathcal{D}} - \mu)^2] \quad &= \quad \mathbb{E}_{\mathcal{D}}\left[\left(\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}(x-\mu)\right)^2\right] \\
&\overset{\substack{\text{chain rule} \\ \text{i.i.d. } \bar{x}\in\mathcal{D}}}{=} \quad \mathbb{E}_{\mathcal{D}}\Bigg[\frac{1}{|\mathcal{D}|^2}\Bigg(\underbrace{(\bar{x}-\mu)^2}_{=\sigma^2} + \Bigg(\sum_{\substack{x\in\mathcal{D}\\x\neq\bar{x}}}(x-\mu)\Bigg)^2 \\
&\qquad + \underbrace{(\bar{x}-\mu)}_{\mathbb{E}[\bar{x}]=\mu}\Bigg(\sum_{\substack{x\in\mathcal{D}\\x\neq\bar{x}}}(x-\mu)\Bigg)\Bigg)\Bigg] \\
&\overset{\text{recursively}}{=} \quad \mathbb{E}_{\mathcal{D}}\left[\frac{1}{|\mathcal{D}|^2}(\sigma^2+\cdots+\sigma^2)\right] = \frac{\sigma^2}{|\mathcal{D}|}
\end{aligned}$$

Sample variance $\hat{\sigma}^2_{\mathcal{D}}$ is biased

$$\mathbb{E}_{\mathcal{D}}[\hat{\sigma}^2_{\mathcal{D}}] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}(x-\mu)^2\right]}_{\text{variance of } x} - \underbrace{\mathbb{E}_{\mathcal{D}}\left[(\mu-\hat{\mu}_{\mathcal{D}})^2\right]}_{\text{variance of } \hat{\mu}_{\mathcal{D}}} = \sigma^2 + \frac{\sigma^2}{|\mathcal{D}|}$$

Summary

- Estimating quantities
- Measuring uncertainty of quantities
  - Hoeffding's inequality
  - Probably approximately correct (PAC) learning
- Maximum likelihood estimators
  - Gaussian: sample mean, sample variance
  - Biased vs unbiased