4 - Estimation and prediction

- Indicator functions
- The Markov inequality
- The Chebyshev inequality
- Confidence intervals
- Standard deviation
- Selecting an estimate
- Minimum probability of error
- Mean square error
- The MMSE predictor
- Cost matrices
- Minimum cost estimates

Indicator functions

Suppose $A \subset \Omega$ is an event. Define the *indicator function* $1_A : \Omega \to \mathbb{R}$ by

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

The important property is

$$\boxed{\mathbf{E}\, 1_A = \mathbf{Prob}(A)}$$

Because

$$\begin{aligned} \mathbf{E}\, 1_A &= \sum_{\omega \in \Omega} 1_A(\omega) p(\omega) \\ &= \sum_{\omega \in A} p(\omega) \end{aligned}$$

The Markov inequality

Suppose

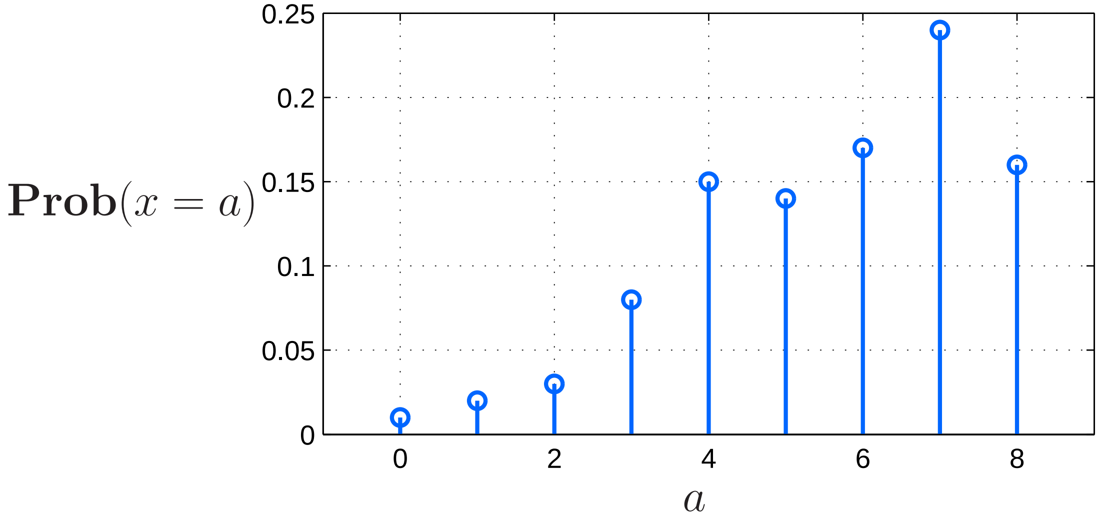
- $x : \Omega \to \mathbb{R}$ is *real-valued random variable*
- x is *nonnegative*, i.e., $x(\omega) \geq 0$ for all $\omega \in \Omega$

The *Markov inequality* bounds the probability that x is large

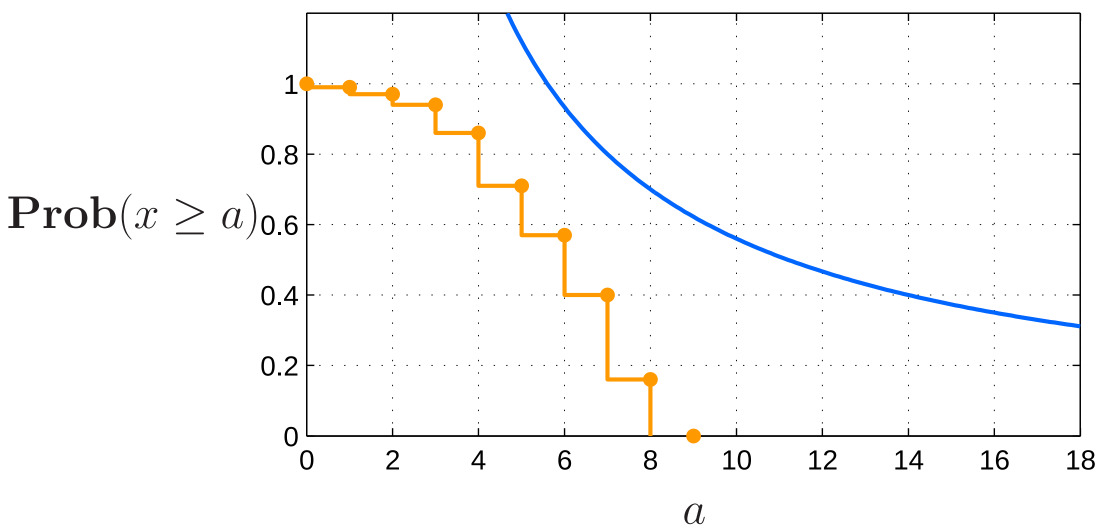
$$\boxed{\mathbf{Prob}(x \geq a) \leq \frac{1}{a} \mathbf{E}\, x}$$

Example: the Markov inequality

Look at the following pmf



The Markov inequality gives an upper bound on $\mathbf{Prob}(x \geq a)$ using the mean



The Markov inequality

$$\boxed{\mathbf{Prob}(x \geq a) \leq \frac{1}{a}\mathbf{E}\,x}$$

- Recall the mean is the *center of mass* of x.

  Since x is nonnegative, the mean gives a bound on the probability that x takes large values.

- The Markov inequality is a *prediction* of the outcome.

- To use it, we need
  - to know x is nonnegative
  - to know the mean $\mathbf{E}\,x$

- We *don't* need to know the pmf!

Proof of the Markov inequality

First, a useful fact: if y and z are random variables, and

$$y(\omega) \leq z(\omega) \qquad \text{for all } \omega \in \Omega$$

Then $\mathbf{E}\, y \leq \mathbf{E}\, z$

Because

$$\begin{aligned} \mathbf{E}\, y &= \sum_{\omega \in \Omega} y(\omega) p(\omega) \\ &\leq \sum_{\omega \in \Omega} z(\omega) p(\omega) \\ &= \mathbf{E}\, z \end{aligned}$$

Proof of the Markov inequality

Define the function $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1 & \text{if } x \geq a \\ 0 & \text{otherwise} \end{cases}$$

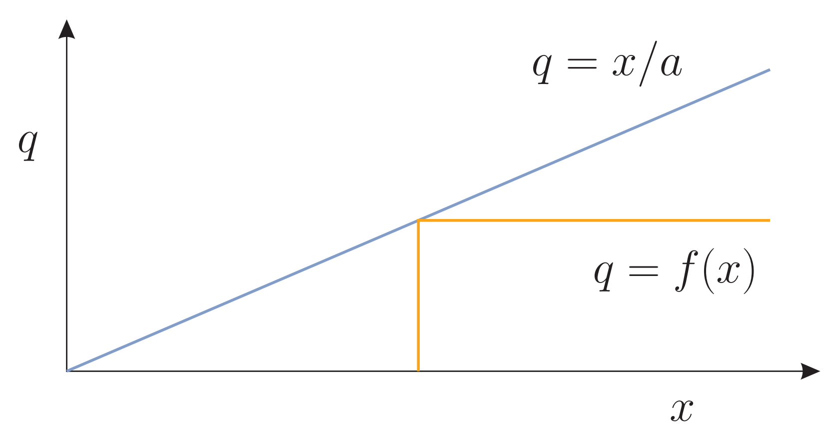
Let y be the random variable $y = f(x)$. Then $\mathbf{E}\, y = \mathbf{Prob}(x \geq a)$

Let z be the random variable $z = x/a$. Then $\mathbf{E}\, z = \dfrac{1}{a}\mathbf{E}\, x$

Then $y(x(\omega)) \leq z(x(\omega))$ for all $\omega \in \Omega$ and so

$$\mathbf{E}\, y \leq \mathbf{E}\, z$$

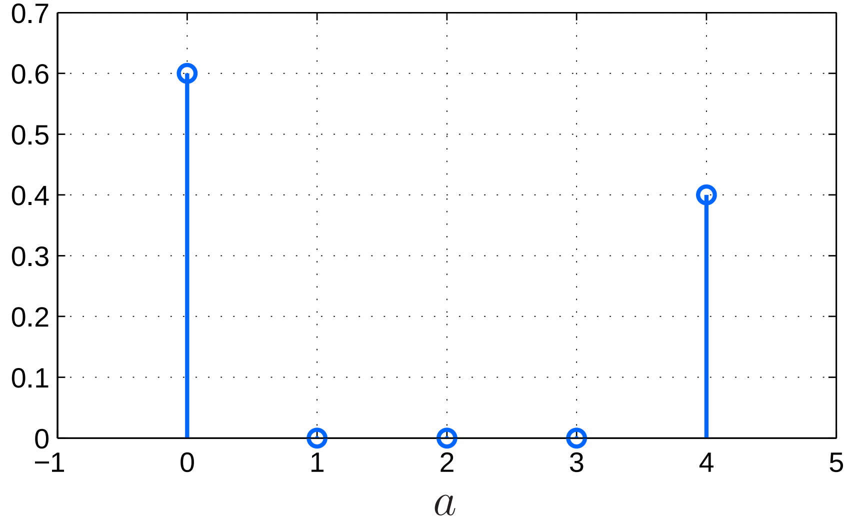
hence $\mathbf{Prob}(x \geq a) \leq \dfrac{1}{a}\mathbf{E}\, x$



The Markov inequality

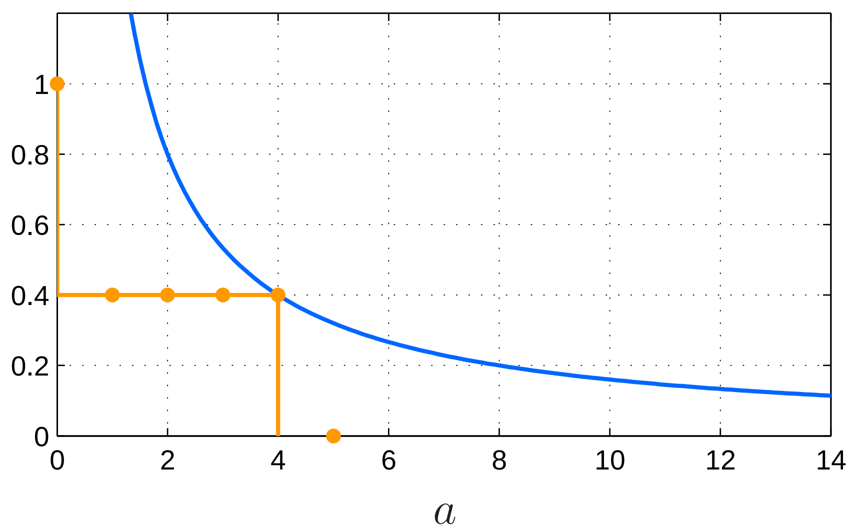
An example where the bound is tight at a point

$\mathbf{Prob}(x = a)$



The Markov inequality gives an upper bound on $\mathbf{Prob}(x \geq a)$

$\mathbf{Prob}(x \geq a)$



The Chebyshev inequality

Suppose $x : \Omega \to \mathbb{R}$. The *Chebyshev inequality* is

$$\boxed{\mathbf{Prob}\big(|x - \mathbf{E}\, x| \geq a\big) \leq \frac{1}{a^2}\,\mathbf{cov}(x)}$$

- Variance $\mathbf{cov}(x)$ gives a bound on the probability that x is far from the mean $\mathbf{E}\, x$.
- Again, we do not need to know the pmf.
- For *any pmf with finite variance*, as a becomes large

  the probability that $|x - \mathbf{E}\, x| \geq a$ decreases faster than $1/a^2$

The Chebyshev inequality

The proof is similar to that for the Markov inequality. Let $\mu = \mathbf{E}\,x$, and define

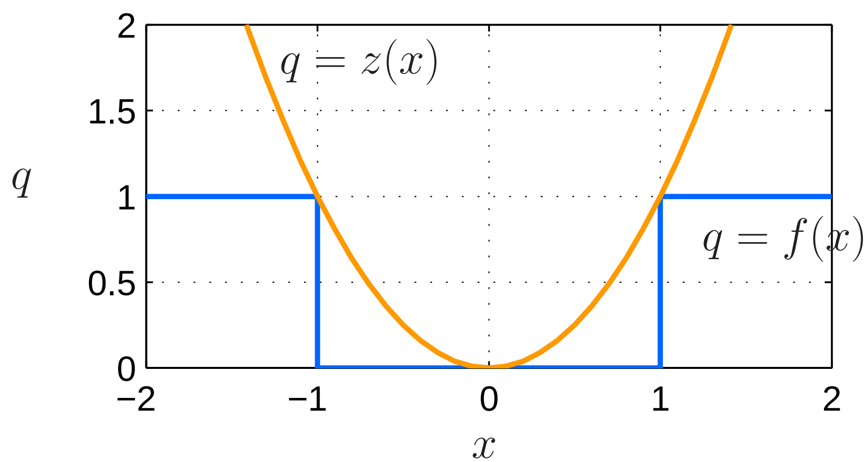
$$f(x) = \begin{cases} 1 & \text{if } |x-\mu| \geq a \\ 0 & \text{otherwise} \end{cases}$$

and let $y = f(x)$ so that $\mathbf{E}\,y = \mathbf{Prob}(|x-\mu| \geq a)$.

Also let $z = (x-\mu)^2/a^2$ so that $\mathbf{E}\,z = \mathbf{cov}(x)/a^2$

Then $y\big(x(\omega)\big) \leq z\big(x(\omega)\big)$ for all $\omega \in \Omega$, hence $\mathbf{E}\,y \leq \mathbf{E}\,z$.

e.g., when $a = 1$ and $\mu = 0$



Confidence intervals

The Chebyshev bound also can be written as

$$\mathbf{Prob}\Big(x \in [\mathbf{E}\,x - a, \mathbf{E}\,x + a]\Big) \geq 1 - \frac{\mathbf{cov}(x)}{a^2}$$

- The interval $[\mathbf{E}\,x - a, \mathbf{E}\,x + a]$ is called a *confidence interval*.
- a is called the *half-width*
- $1 - \mathbf{cov}(x)/a^2$ is called the *confidence level*

Confidence intervals and standard deviation

Denote the standard deviation of x by $\sigma = \mathbf{std}(x)$. Then

$$\mathbf{Prob}\Big(x \in [\mathbf{E}\,x - a, \mathbf{E}\,x + a]\Big) \geq 1 - \big(\sigma/a\big)^2$$

Some examples:

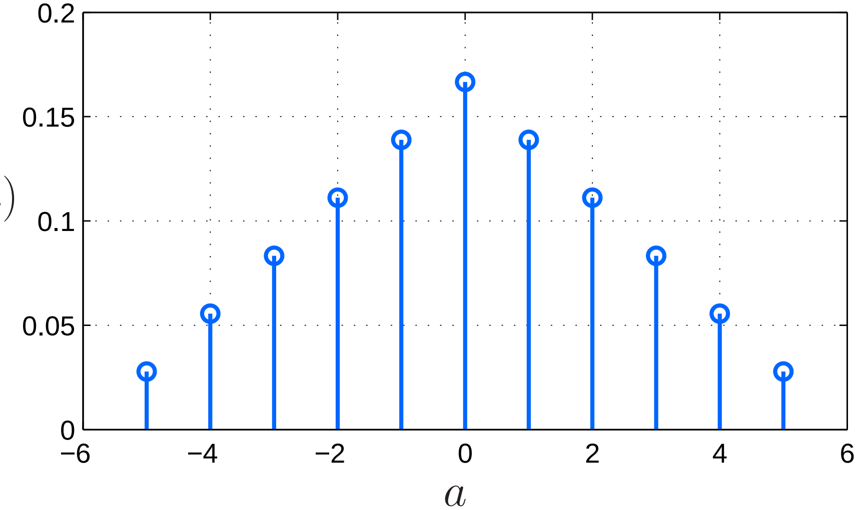
- Pick $a = 3\sigma$; then the probability that x lies within 3σ of the mean is at least 0.88
- Choosing $a = 6\sigma$ gives probability 0.97

Note: we need to know only σ, nothing else about the pdf of x! (but the bound may be loose)

Confidence Intervals

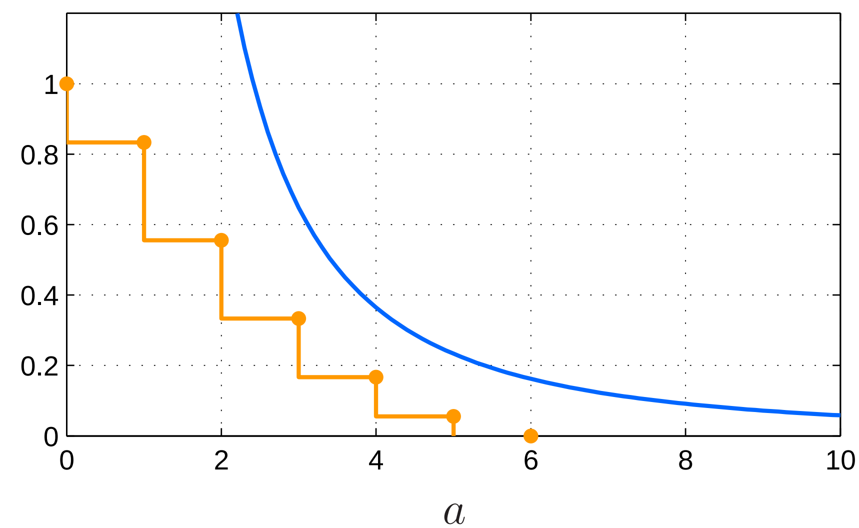
There is a trade-off between the width of the confidence interval and the confidence level

$\mathbf{Prob}(x = a)$



The Chebyshev bound gives an upper bound on $\mathbf{Prob}\big(|x - \mathbf{E}\,x| \geq a\big)$

$\mathbf{Prob}(|x - \mathbf{E}\,x| \geq a)$



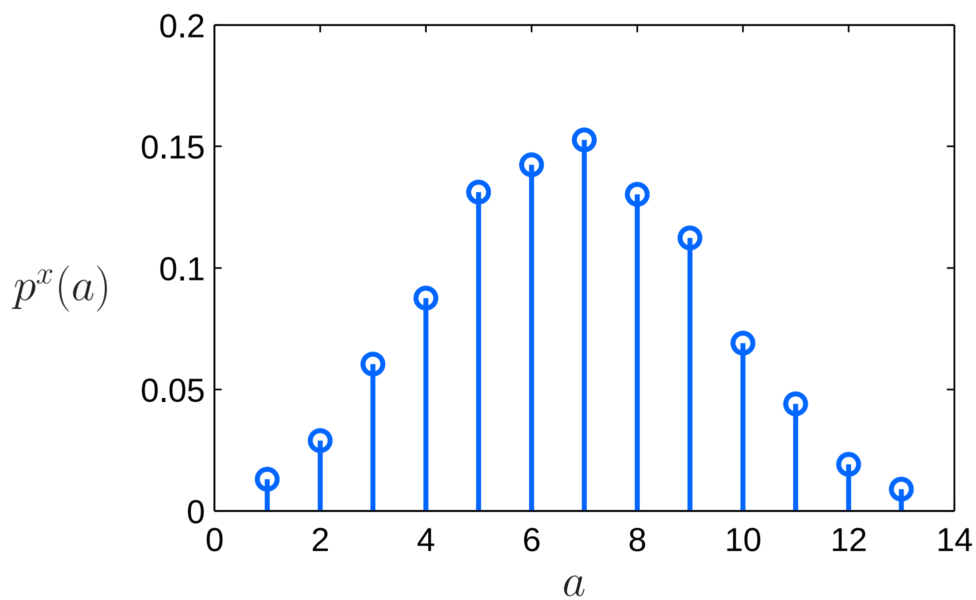
Selecting an estimate

Suppose $x : \Omega \to \mathbb{R}$ is a random variable, with induced pmf $p^x : \mathbb{R} \to [0, 1]$.

The induced pmf p^x may be

- The frequencies of letter usage in a book, rock sample types, etc.
- How often an aircraft passes within range of a radar system i.e., $\Omega = \{0, 1\}$
- Discretized force exerted by wind disturbances on a vehicle; (usually $\Omega = \mathbb{R}^n$)

We want to predict the outcome of the experiment; which $x_{\text{est}} \in \mathbb{R}$ should we pick?



Selecting an estimate

Some possible choices

- We could pick the mean. A disadvantage is that the sample space Ω is a finite set, so the mean may not equal $x(\omega)$ for any $\omega \in \Omega$; then the prediction is always wrong.

- One choice is to *minimize the probability of error*. We have

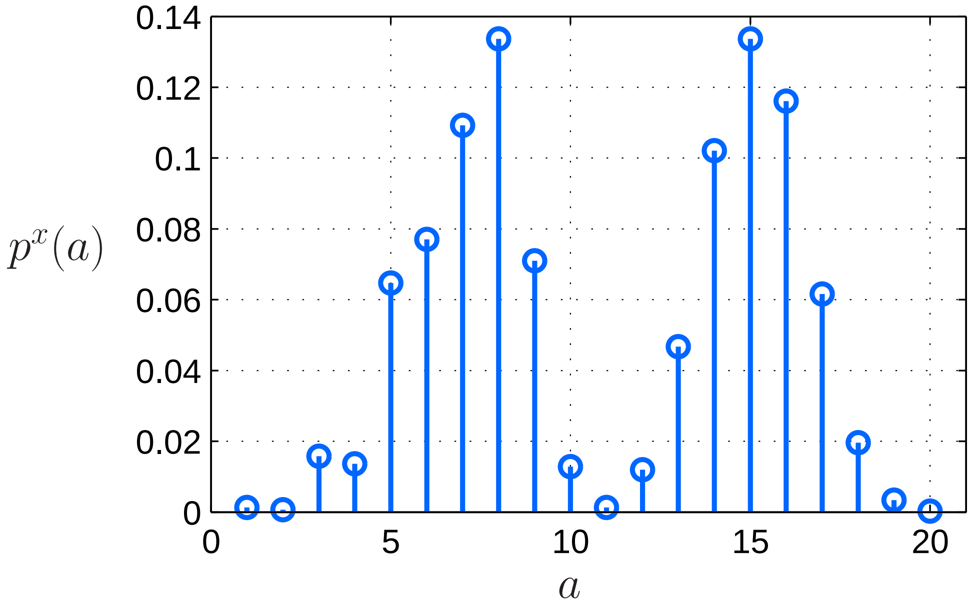
$$\begin{aligned}\text{probability of error} &= \mathbf{Prob}(x \neq x_{\text{est}}) \\ &= \sum_{a \neq x_{\text{est}}} p^x(a) \\ &= 1 - p^x(x_{\text{est}})\end{aligned}$$

  So to minimize the error probability, pick x_{est} to maximize $p^x(x_{\text{est}})$.

Problems with selecting an estimate

What's wrong with minimizing the probability of error?

- One problem is possible nonuniqueness: which peak do we want?



Usually we can handle this

- Other problems occur also. . .

Problems with selecting an estimate

- If $x : \Omega \to \mathbb{R}$, then there may be a natural choice of error

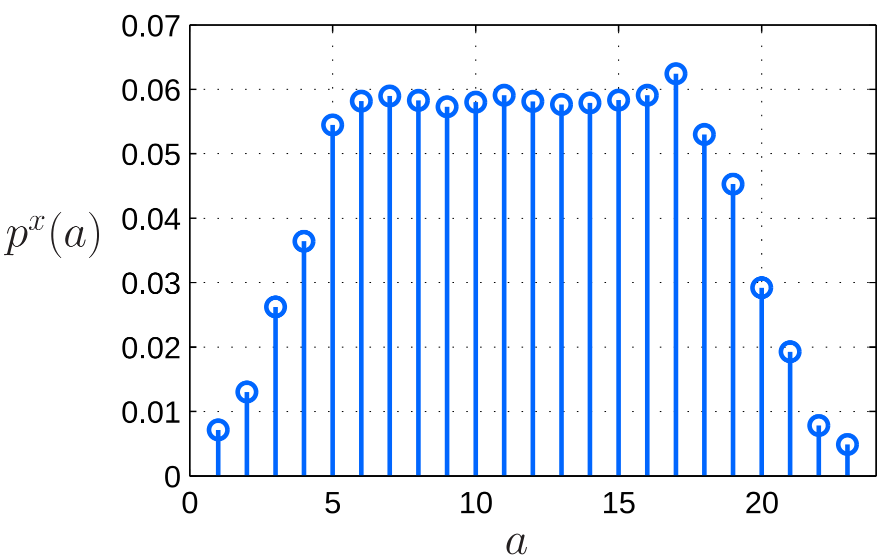
  e.g., for a radar, observing 2 aircraft is very different from observing 10 aircraft.

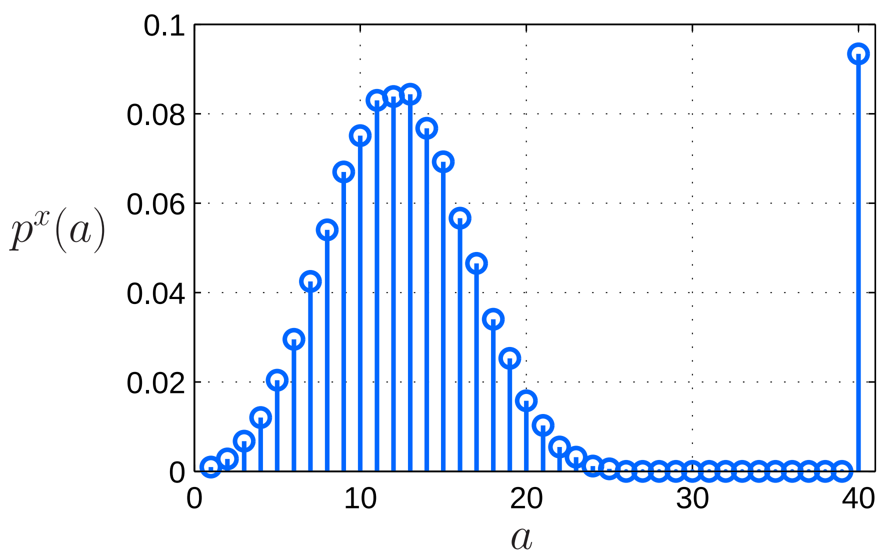
- conversely, there may be no metric;

  e.g., for character recognition, $x : \Omega \to \{a, b, c, \ldots, z\}$

  mistaking a for b is not better than mistaking a for q

If there is a metric, the minimum error estimate might be far from other good choices





Problems with selecting an estimate

Suppose $\Omega = \mathbb{R}$ and $x : \Omega \to \mathbb{R}$ is a continuous random variable.

- The probability of error is always 1; i.e., the prediction is always wrong.
- There is *no estimate* that gives minimum error probability
- Here we can pick the mean, but why?

In order to select the *best estimate*, we need a *cost function*.

The mean square error

The mean square error is

$$\boxed{\mathbf{mse}(x_{\mathsf{est}}) = \mathbf{E}\Big((x - x_{\mathsf{est}})^2\Big)}$$

- A very common choice for error
- We'll use it many times in this course

The minimum mean square error (MMSE) predictor

The estimate that minimizes the MSE is the mean.

$$\boxed{x_{\text{opt}} = \mathbf{E}\, x}$$

Because

$$\begin{aligned} \mathbf{E}\big((x-a)^2\big) &= \mathbf{E}\big(x^2 - 2ax + a^2\big) \\ &= \mathbf{E}\big(x^2\big) - 2a\,\mathbf{E}\,x + a^2 \end{aligned}$$

Then differentiating with respect to a gives

$$-2\,\mathbf{E}\,x + 2a = 0$$

and hence

$$a_{\text{opt}} = \mathbf{E}\,x$$

The minimum mean square error (MMSE) predictor

An alternate proof is given by the *mean-variance decomposition*, which says

$$\mathbf{E}(x^2) = (\mathbf{E}\, x)^2 + \mathbf{E}\big((x - \mathbf{E}\, x)^2\big)$$

Apply this to the *error random variable* z

$$z = x - x_{\text{est}}$$

Then we have

$$\begin{aligned}
\mathbf{mse}(x_{\text{est}}) &= \mathbf{E}(z^2) \\
&= (\mathbf{E}\, z)^2 + \mathbf{E}\big((z - \mathbf{E}\, z)^2\big) \\
&= (\mathbf{E}\, z)^2 + \mathbf{E}\big((x - x_{\text{est}} - (\mathbf{E}\, x - x_{\text{est}}))^2\big) \\
&= (\mathbf{E}\, z)^2 + \mathbf{E}\big((x - \mathbf{E}\, x)^2\big) \\
&= (\mathbf{E}\, z)^2 + \mathbf{cov}(x)
\end{aligned}$$

The minimum mean square error (MMSE) predictor

So we have

$$\boxed{\mathbf{mse}(x_{\text{est}}) = (\mathbf{E}(x) - x_{\text{est}})^2 + \mathbf{cov}(x)}$$

- The first term is the square of the *mean error*

$$\mathbf{E}\, z = \mathbf{E}(x) - x_{\text{est}}$$

  The mean error $\mathbf{E}\, z$ is called the *bias* of the estimate x_{est}.

  The best we can do is to make this zero.

- The second term is the *covariance* of x; it is the error we *cannot remove*

Cost matrices

Suppose $x : \Omega \to V$, and $V = \{v_1, v_2, \ldots, v_n\}$.

- Exactly one outcome $\omega \in \Omega$ occurs
- Hence exactly one element of V occurs
- We'd like to predict which one.

We'll specify the cost by a *cost matrix* $C \in \mathbb{R}^{n \times n}$

$$C_{ij} = \text{cost of estimating } v_i \text{ when outcome is } v_j$$

Notice that

- for every estimate v_j and every outcome v_i, there may be a *different cost*.

Example: cost matrices

If $n = 4$, i.e., there are four possible outcomes, then one choice for C is

$$C = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

- Here we pick $C_{ii} = 0$ so that correct estimates have no cost.
- $C_{ij} = 1$ when $i \neq j$ so that all incorrect estimates incur the same cost

Example: cost matrices

Another choice for C is

$$C = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{bmatrix}$$

- Here the cost depends on $C_{ij} = |i - j|$
- If $V \subset \mathbb{R}$, we often assign costs of the form $C_{ij} = f(i - j)$, i.e., C_{ij} is a function only of $i - j$.
- So the matrix C is *Toeplitz*

Coding for minimum cost estimates

To represent the estimate $x_{\text{est}} \in V$, we'll use an *indicator vector* $k \in \mathbb{R}^n$

$$k_i = \begin{cases} 1 & \text{if } i = i_{\text{est}} \\ 0 & \text{otherwise} \end{cases}$$

Here i_{est} is the index of the estimate.

Also let $p^x \in \mathbb{R}^n$ be the induced pmf of x.

Minimum cost estimates

Suppose the estimator is defined by the indicator vector k.

- Then $C^T k$ is a random variable, which assigns costs to outcomes.
- Since k is an indicator vector, $C^T k$ is given by the the i_{est}'th row of C.

The *expected cost* is therefore

$$\boxed{\mathbf{E}\, C^T k = k^T C p^x}$$

We can then pick the *optimal estimator*; the one that minimizes the cost, by setting i_{est} to the index of the smallest element of Cp^x

Minimum cost estimates and minimum probability of error

Minimizing the probability of error is equivalent to choosing cost matrix

$$C = \begin{bmatrix} 0 & 1 & \dots & & 1 \\ 1 & 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & & \\ & & & 0 & 1 \\ 1 & \dots & & 1 & 0 \end{bmatrix} = \mathbf{1}\mathbf{1}^T - I$$

Then

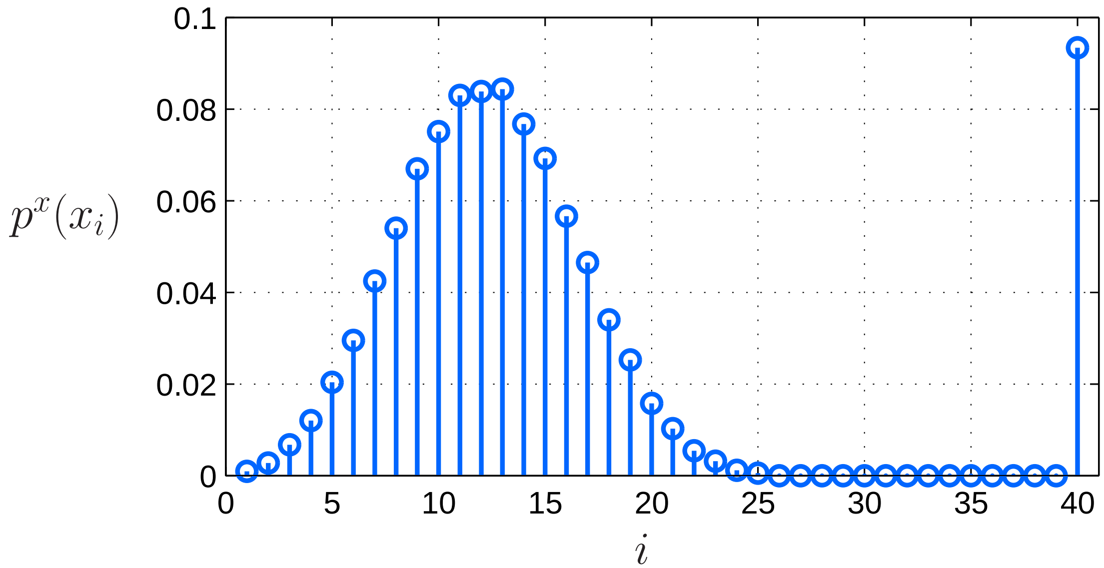
$$Cp^x = (\mathbf{1}\mathbf{1}^T - I)p^x = \mathbf{1} - p^x$$

and i_{est} selects the smallest element of $\mathbf{1} - p^x$, i.e., it selects the largest element of p^x

The cost matrix $C = \mathbf{1}\mathbf{1}^T - I$ is called the *Bayes risk*

Example: minimum cost estimates

We'll consider the distribution



and three cost matrices

$$C^{\text{min-error}} = \mathbf{1}\mathbf{1}^T - I \qquad C^{\text{abs}}_{ij} = |i - j| \qquad C^{\text{squared}}_{ij} = (i-j)^2$$

The corresponding estimates are

$$i_{\text{min-error}} = 40 \qquad i_{\text{abs}} = 13 \qquad i_{\text{squared}} = 15 \qquad \mathbf{E}\, x \approx 14.85$$