Estimating Probability Distribution/Density Functions

Examples of p.d.f. estimation

- Suppose we "randomly" select a set of cancer patients who have tumors removed.
- For each one we see if their cancer recurs or not, and we want to estimate the probability that a new patient's cancer will recur.

	recur	not recur
number of patients	47	151

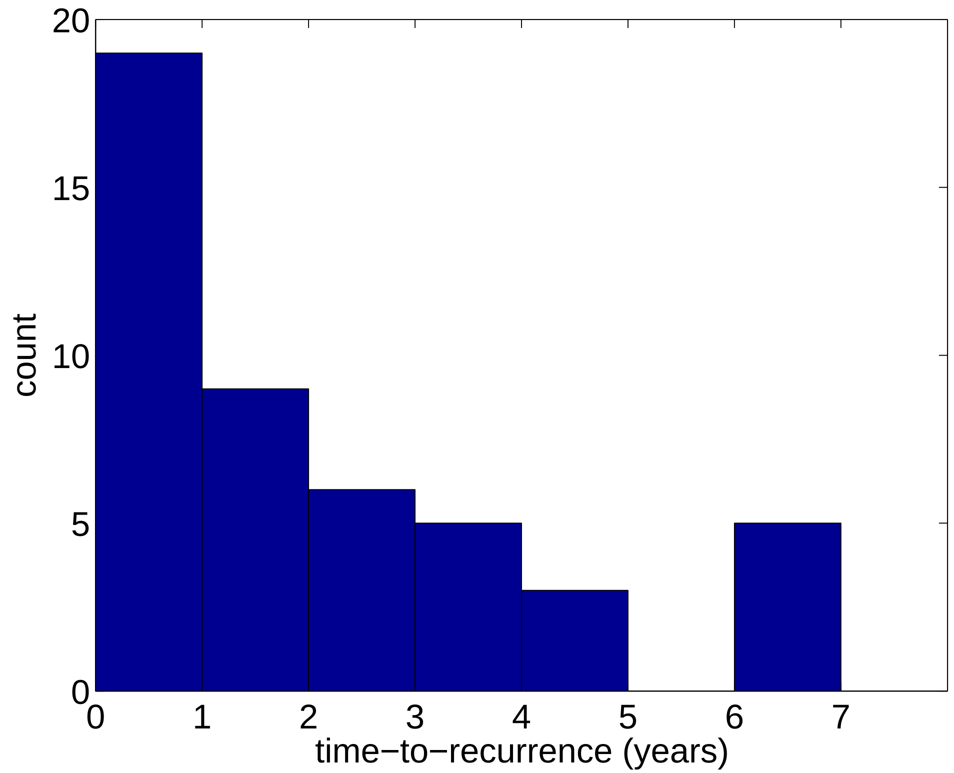
- Suppose we also measured the size of the tumor cells, and want to estimate the joint probability of cell size > 17.4 and recurrence.

	recur	not recur
cell size > 17.4	31	16
cell size ≤ 17.4	66	85

Examples of p.d.f. estimation (2)

- Suppose we measure the time-to-recurrence, for the patients whose cancer recurs. We want to predict the time-to-recurrence for a new patient.

patient	t-to-r (months)
1	27
2	77
3	77
4	36
5	10
6	10
7	9
⋮	⋮



In this lecture

- Estimating p.d.f.'s of discrete and continuous random variables.
- The principle of maximum likelihood.
- We mainly discuss parametric p.d.f. estimation.

P.d.f. estimation for binary r.v.'s

- Suppose we observe m independent binary r.v.'s, $X_1, X_2, \ldots, X_m$, each equal to one with probability p. (These are called *Bernoulli* r.v.'s.)
- Suppose m_1 come out as ones and $m_0 = m - m_1$ come out as zeros.
- How to estimate p?
- An obvious estimate is $p = \frac{m_1}{m} = \frac{m_1}{m_0+m_1}$.
  It turns out this is the maximum likelihood estimate of p.

	recur	not recur
number of patients	47	151
probability	$0.24 = \frac{47}{47+151}$	$0.76 = \frac{151}{47+151}$

Maximum likelihood estimation of p

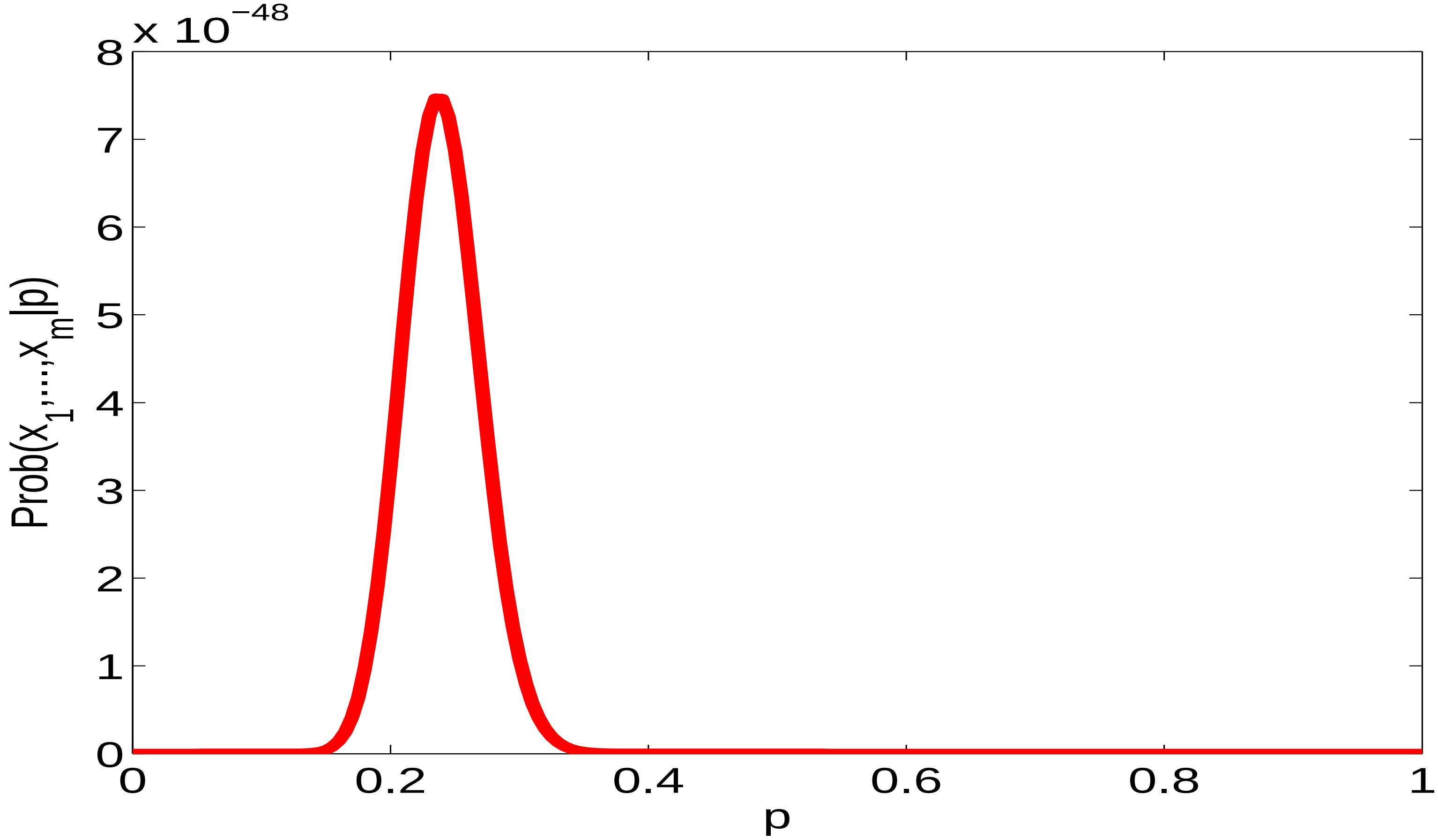
- For a particular p, the probability that we would observe the data, also called the likelihood of the data, is

$$\begin{aligned} \mathsf{P}(X_1, \ldots, X_m|p) &= \Pi_i \mathsf{P}(X_i|p) \\ &= \Pi_i \begin{cases} p & \text{if } X_i = 1 \\ 1-p & \text{if } X_i = 0 \end{cases} \\ &= \Pi_i p^{X_i}(1-p)^{(1-X_i)} \end{aligned}$$

- The *"principle" of maximum likelihood* says that the best estimate for p is the one that maximizes $\mathsf{P}(X_1, \ldots, X_m|p)$.

Example

With 47 recurrences and 151 non-recurrences, the probability of the data, as a function of $p =$estimated probability of recurrence is:



Maximum likelihood estimation of p (2)

- Equivalently, we can maximize $\log \mathsf{P}(X_1, \ldots, X_m | p)$ with respect to p.

$$\log \mathsf{P}(X_1, \ldots, X_m | p) = \sum_i X_i \log p + (1 - X_i) \log(1 - p)$$

- If all $X_i = 1$, then the maximum is at $p = 1 = \frac{m_1}{m_0 + m_1}$.
- If all $X_i = 0$, then the maximum is at $p = 0 = \frac{m_1}{m_0 + m + 1}$.
- Otherwise, differentiate w.r.t. p and set equal to zero

$$\begin{aligned} \frac{d}{dp} \sum_i X_i \log p + (1 - X_i) \log(1 - p) &= 0 \\ \sum_i X_i \frac{1}{p} - (1 - X_i) \frac{1}{1 - p} &= 0 \end{aligned}$$

Maximum likelihood estimation of p (3)

$$
\begin{aligned}
\frac{\sum_i X_i(1-p) - (1-X_i)p}{p(1-p)} &= 0 \\
\sum_i X_i(1-p) - (1-X_i)p &= 0 \\
m_1(1-p) - m_0 p &= 0 \\
m_1 - p(m_1 + m_0) &= 0 \\
p &= \frac{m_1}{m_0 + m_1}
\end{aligned}
$$

- In all cases, the maximum likelihood estimate of p is $\frac{m_1}{m_0+m_1}$.

Maximum likelihood estimation in general

- Let $X_1, X_2, \ldots, X_m$ be a set of random variables (discrete or continuous). We typically assume:
  - The X_i's are independent r.v.'s.
  - They have the same p.d.f., θ_{true}.
    That is $\mathsf{P}(X_i = x) = \theta_{true}(x)$ for all i.
- We want to estimate θ_{true}.
- Let H be a set of candidate distributions.
- The "best" estimate for θ_{true}, based on the data $X_1, \ldots, X_m$, is

$$\theta \in \arg\max_{\theta \in H} \mathsf{P}(X_1, \ldots, X_m | \theta)$$

Maximum likelihood for more than two discrete outcomes

- Let the X_i be discrete r.v.'s each with the same k possible outcomes.
- Let outcome k occur m_k times, across all the X_i.
- Then the maximum likelihood estimate for P(k) is just m_k/m.

number of patients	recur	not recur
cell size $>$ 17.4	31	16
cell size $\leq$ 17.4	66	85

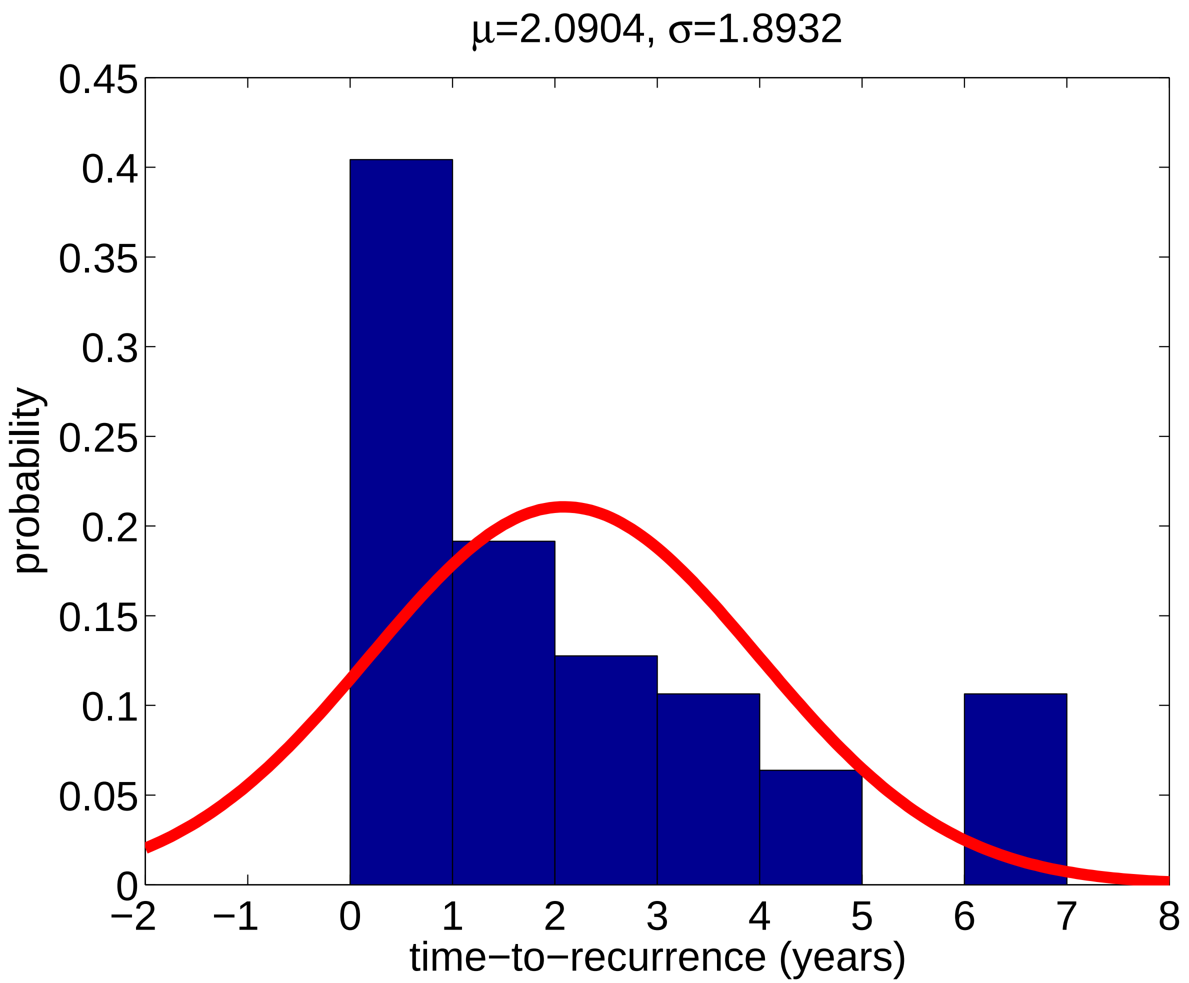
probability	recur	not recur
cell size $>$ 17.4	$0.16 = \frac{31}{198}$	$0.08 = \frac{16}{198}$
cell size $\leq$ 17.4	$0.33 = \frac{66}{198}$	$0.43 = \frac{85}{198}$

Maximum likelihood Gaussian fit

- Suppose the X_i are real-valued.
- Let $H =$ the set of all Gaussian distributions (any μ, any σ).
- Which μ and σ maximize the probability of the data?
- ...if you go through all the math, you find

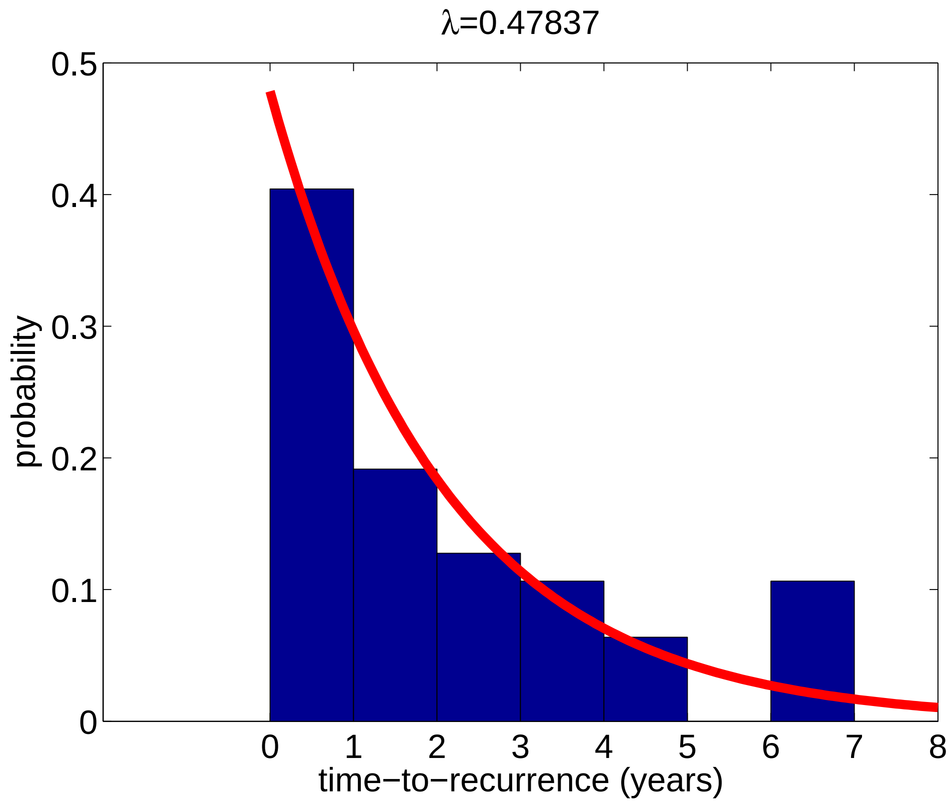
$$\mu = \frac{1}{m} \sum_i X_i \qquad \sigma^2 = \frac{1}{m} \sum_i (X_i - \mu)^2$$

Example: M.L. Gaussian fit to the time-to-recurrence data.



Maximum likelihood exponential fit

- The exponential density with parameter λ is P$(x) = \lambda e^{-\lambda x}$.
- The M.L. exponential fit is given by $\lambda = 1/\sum_i X_i$.



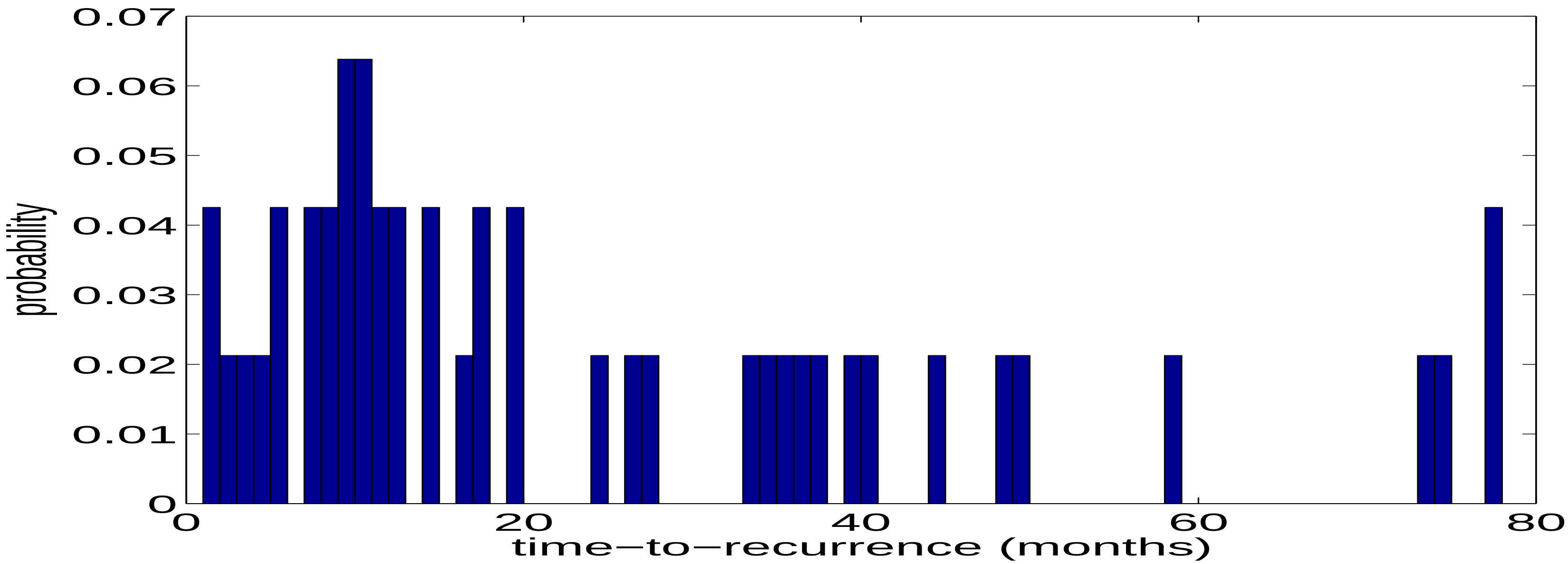
Maximum likelihood p.d.f. estimates

- For discrete r.v.'s and a variety of univariate and multivariate continuous distributions (such as Gaussian and exponential), the M.L. estimate can be computed easily from the data.
- What if some r.v.'s are discrete and some continuous?
- Problems?
  - For discrete r.v.'s, non-occurring values can be a problem. (See next slide for an example.)
  - As always, the best fit might not be very good. . .

Non-occurring values in discrete distributions

- Suppose we interpret the time-to-recurrence (reported in integer months) to be a discrete r.v.

- Max. likelihood distribution? P$(x) = ($count of $x)/m$.



- A common, quick fix to zero counts is to add *pseudocounts*. (=Dirichlet prior.)