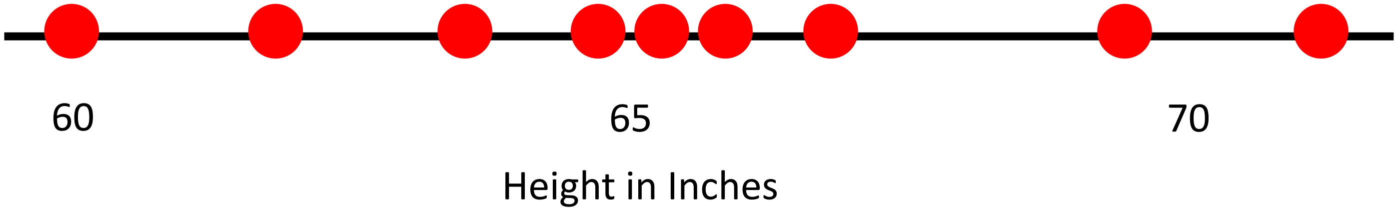
Plan for the week

- M: Maximum Likelihood Estimation
  - Naive Bayes
- W: Clustering
  - *k*-means clustering
  - Hierarchical clustering
- Miscellaneous Algorithms
  - Regression Trees, Logistic Regression, etc.

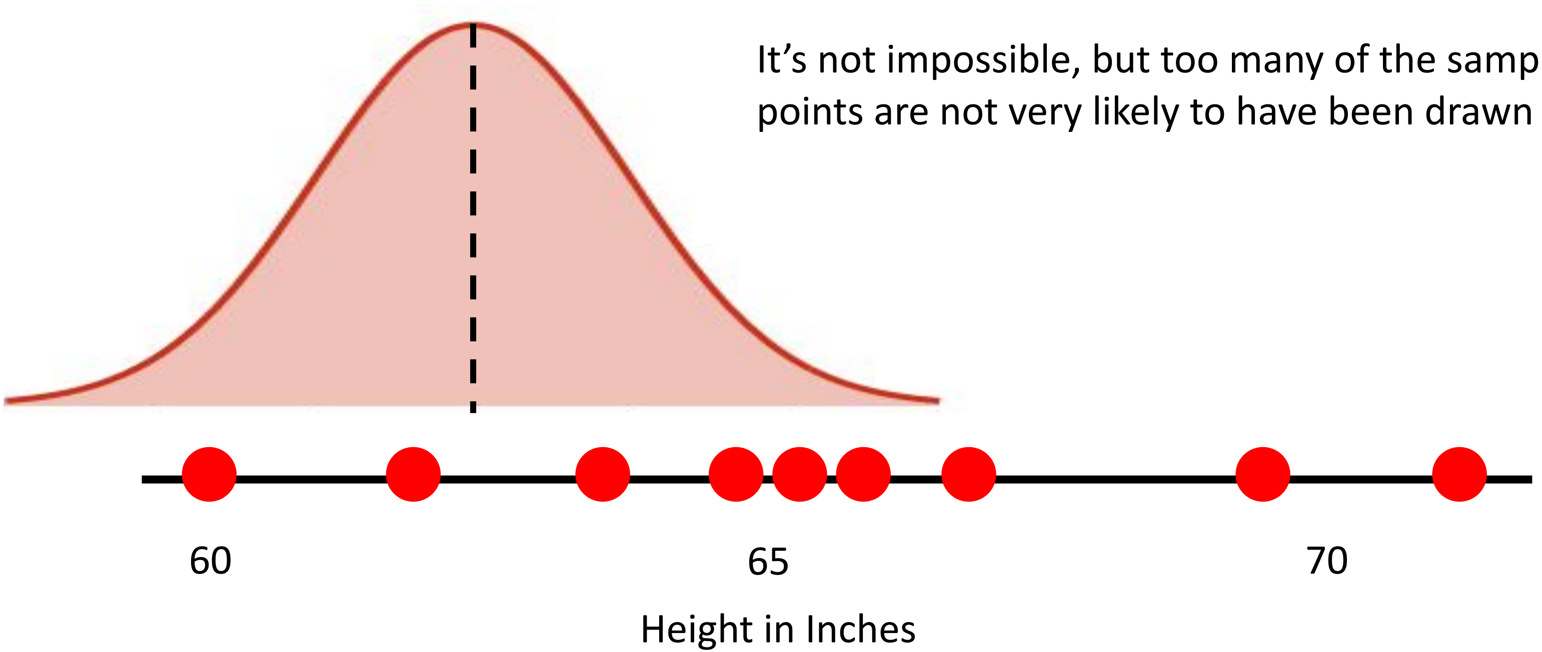
Maximum Likelihood Estimation

MLE: Intuition

We have data and suspect it is normally distributed.
What would be a good estimate of the mean of the distribution?



MLE: Intuition (cont’d)

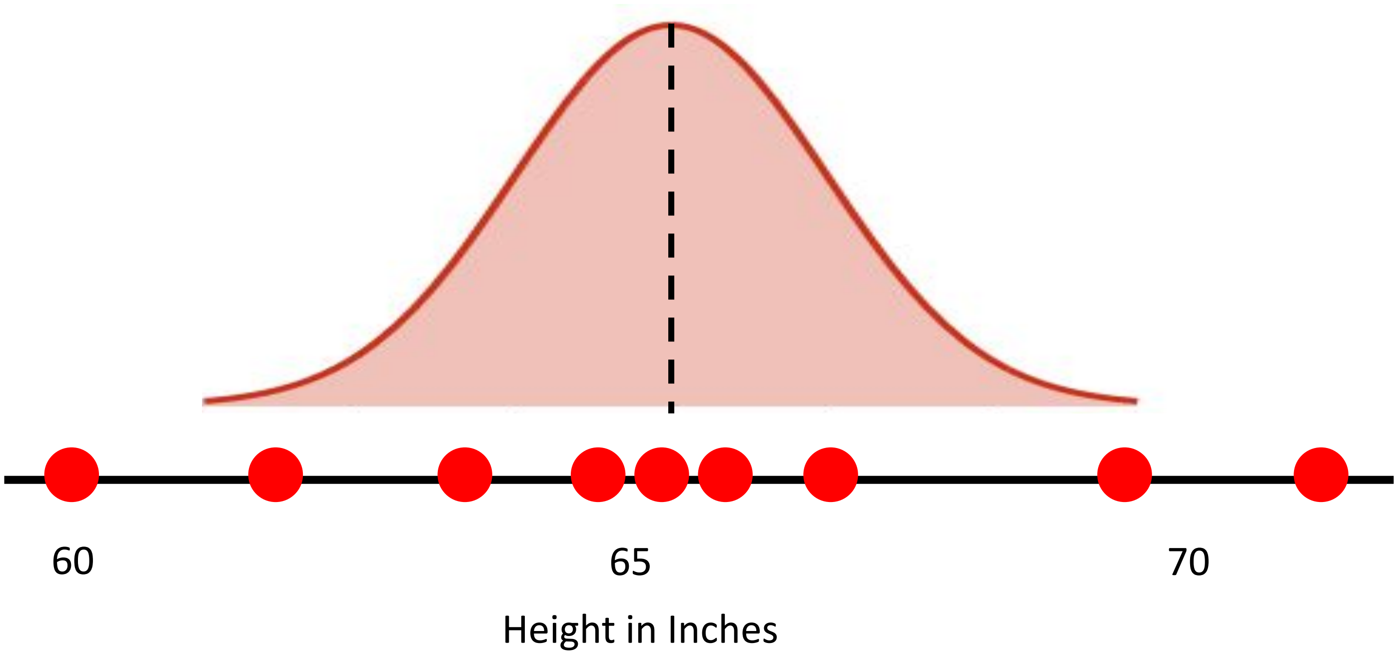


How about this distribution with a mean of about 63?

It’s not impossible, but too many of the sample data points are not very likely to have been drawn from it.

MLE: Intuition (cont'd)

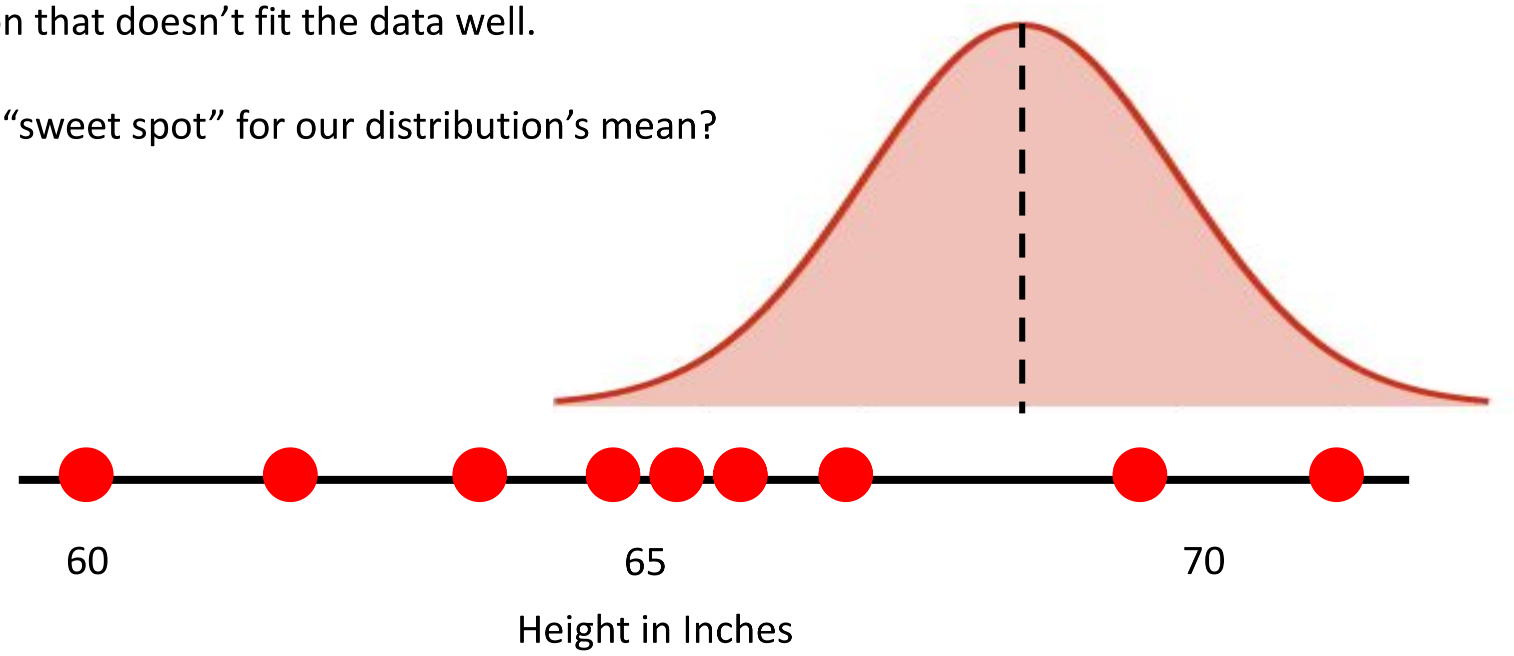
This distribution looks a lot better! It ascribes high probability to most of the points.



MLE: Intuition (cont'd)

If we increase the mean even further, we again arrive at a distribution that doesn't fit the data well.

What is the "sweet spot" for our distribution's mean?



Generic Parameter Estimation

- Assume data are generated by a probabilistic model with parameter θ.
- Data are independent and identically distributed $D = \{d_1, \ldots, d_n\}$.
- Goal is to estimate θ (i.e., the parameter) well, given data.
- MLE is one approach. (There are many others!)

MLE: Mathematical Formulation

- Find the parameter θ that maximizes the likelihood of the data: i.e., find θ s.t. $P(D \mid \theta) = P(d_1 \mid \theta) \cdots P(d_n \mid \theta)$ is maximized.
  - This simplification used the independence assumption.
- Equivalently, because the log is an increasing function, we can likewise find the parameter θ that maximizes the log likelihood of the data.
- The log of a product of terms is the sum of the logs of those terms.
- The mathematical formulation of MLE is to find θ that maximizes $\log P(D \mid \theta) = \log P(d_1 \mid \theta) + \cdots + \log P(d_n \mid \theta)$

MLE of a Bernoulli RV

- Statistical Model of the data: $D = \{d_1, \ldots, d_n\}$
    - Assume the data are generated by a Bernoulli random variable with parameter p.
    - Assume the data are independent and identically distributed (i.i.d.).
- Goal is to estimate p (i.e., the parameter) well, given data.
- The strategy is maximum likelihood estimation.
- Examples:
    - 00000100100000000100000000000101000000: it is more likely p is close to 0
    - 111101101111101110111101111011101111: it is more likely p is close to 1

What is the likelihood function?

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

$$L(x_i \mid p) = p^{x_i}(1 - p)^{1 - x_i}$$

Double check that this equation makes sense!
(Discuss with your neighbor.)

$$L(\{x_i\}_{i=1}^{n} \mid p) \quad = \quad \prod_{i=1}^{n} L(x_i \mid p)$$

Data are i.i.d.

What is the log likelihood function?

$$
\begin{aligned}
\log L(\{x_i\}_{i=1}^n \mid p) &= \log \prod_{i=1}^n L(x_i \mid p) \\
&= \sum_{i=1}^n \log L(x_i \mid p) \\
&= \sum_{i=1}^n \log \left\{ p^{x_i} (1-p)^{1-x_i} \right\} \\
&= \sum_{i=1}^n \left(x_i \log p + (1-x_i) \log(1-p) \right) \\
&= n\bar{x} \log p + n(1-\bar{x}) \log(1-p)
\end{aligned}
$$

sample proportion

What is the optimal value of *p*?

$$\frac{\partial \log L(\{x_i\}_{i=1}^n \mid p)}{\partial p} = \frac{\partial \{n\bar{x}\log p + n(1-\bar{x})\log(1-p)\}}{\partial p}$$

$$= \frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p}$$

Setting this derivative equal to zero yields:

$$\frac{n\bar{x}}{p^*} = \frac{n(1-\bar{x})}{1-p^*}$$

But then $\bar{x}(1-p^*) = p^*(1-\bar{x})$. So $p^* = \bar{x}$.

sample proportion