BMEG 802 – Advanced Biomedical Experimental Design and Analysis

Maximum Likelihood Estimation

Joshua G. A. Cashaback, PhD

Recap

- ANCOVA
  - covariates
  - can use for any combination of between and within designs.

Today

- Maximum Likelihood Estimation (MLE)
  - Probability Distribution Function
  - Likelihood function
  - 3 Ways to find the Maximum Likelihood Estimation
    - Analytical (Calculus)
    - Brute Force (Grid Search)
    - Optimization (Gradient Descent)

Maximum Likelihood Estimation

- Tool for parameter estimation
- good approach for cases when OLS (ordinary least squares) assumptions are violated
- e.g. for non-linear models with non-normal data
- in MLE, we estimate the parameters of a model that maximize the likelihood of your data

Probability Density Function

- assume an observed **data** vector

$$y = (y_1, y_2, ..., y_n)$$

- Goal of MLE: identify the population (the model) that is **most likely** to have generated the data

Probability Density Function

- Here we assume population (model) is associated with a corresponding probability distribution
- Each probability distribution is characterized by a unique value of the model's parameter(s)
- As model parameters change, different probability distributions are generated
- Model $=$ the family of probability distributions indexed by the model's parameter(s)

Probability Density Function

- $f(y|w)$ is the probability density function (PDF) specifying the probability of observing data y, given model parameter(s) w
  - note: w may be a parameter vector, $w = (w_1, w_2, ..., w_n)$
    - e.g. for a normal PDF: $w = (\mu, \sigma)$

Probability Density Function

- If observations yi are i.i.d. (indepedent and identically distributed), then the PDF for the data as a whole, $y = (y_1, y_2, ..., y_n)$ given the parameter vector $\mathbf{w} = (w_1, w_2, ..., w_n)$, can be expressed as the multiplication of PDFs for individual observations:

$$f(y_1, y_2, ..., y_n|\mathbf{w}) = f_1(y_1|\mathbf{w})f_2(y_2|\mathbf{w}), ..., f_n(y_n|\mathbf{w})$$

Or, more concisely $f(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^{n} f_n(y_n|\mathbf{w})$\$

PDF Example with a Normal Distribution

- Let's say our data vector Y is made up of 3 observations: $y_1 = 80, y_2 = 110, y_3 = 130$
- We want to compute the PDF for a Normal distribution:

$$f(y_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)}{2\sigma^2}}$$

Let's assume $\mu = 100, \sigma = 15$

$$f(80|\mu = 100, \sigma = 15) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(80-\mu)}{2\sigma^2}} = 0.010934$$

$$f(110|\mu = 100, \sigma = 15) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(110-\mu)}{2\sigma^2}} = 0.021297$$

$$f(130|\mu = 100, \sigma = 15) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(130-\mu)}{2\sigma^2}} = 0.003599$$

$$f(y_1, y_2, y_3|\mu,\sigma) = f(y_1|\mu,\sigma)f(y_2|\mu,\sigma)f(y_3|\mu,\sigma) = (0.010934)(0.021297)(0.003599) = .000000838$$

Binomial Distribution Example

- y is the number of successes in a sequence of 10 Bernoulli trials (e.g. tossing a coin 10 times)
- a Bernoulli trial is an experiment whose outcome is random and can be either of two possible outcomes: success or failure.
- Binomial Distribution PDF:

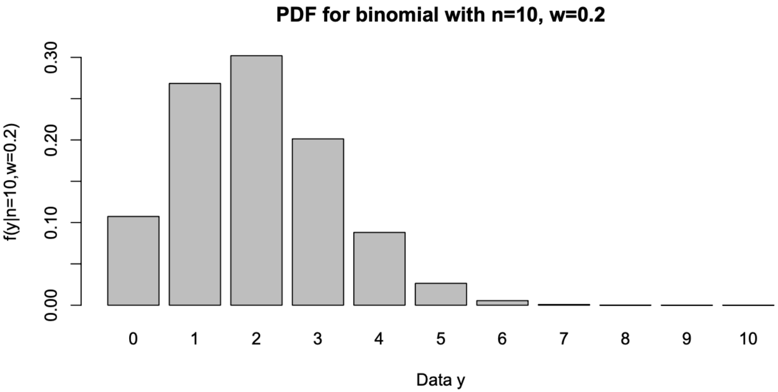
$$f(y|n,w) = \frac{n!}{y!(n-y)!} w^y (1-w)^{n-y}$$

- assume probability of a success on any one trial is 0.2 (a biased coin)
- parameter vector w is n=10, w=0.2

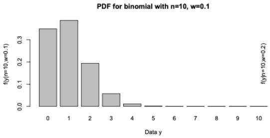
$$f(y|n=10, w=0.2) = \frac{10!}{y!(10-y)!} 0.2^y (1-0.2)^{10-y}; (y = 0, 1, ..., 10)$$

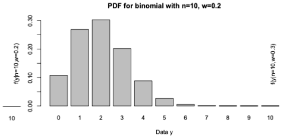
Binomial Distribution Example

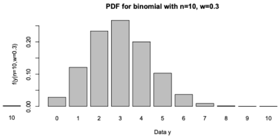
Binomial Distribution

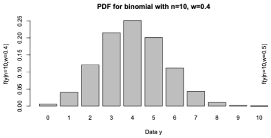


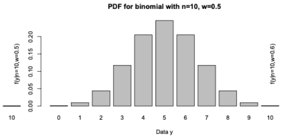
Binomial Distribution - Varying a Parameter

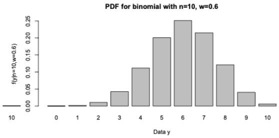


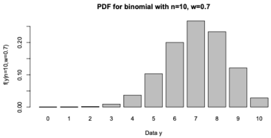








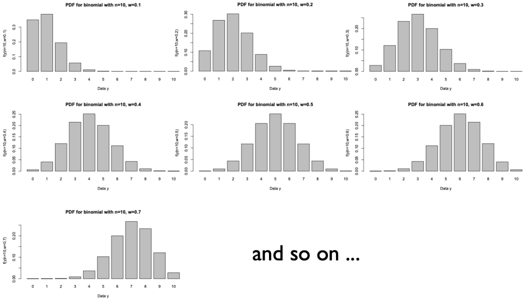




and so on ...

Binomial Distribution - A Model

The collection of all such PDFs generated by varying the parameter across its range defines a **model**



and so on ...

Likelihood Function

Likelihood Function

- Given a set of parameter values, the corresponding PDF will show that some data are more probable than other data
- In fact we have already observed the data

Likelihood Function

- We are faced with the inverse problem
- Given the observed data, and a model of the process by which the data was generated
  - find the one PDF, among all the probability densities that the model prescribes, that is **most likely to have produced the data**

Likelihood Function

- we define the likelihood function by reversing the roles of the data vector y and the parameter vector w in $f(y|w)$:

$$\mathcal{L}(w|y) = f(y|w)$$

$\mathcal{L}(w|y)$ represents the likelihood of the parameter w given the observed data y

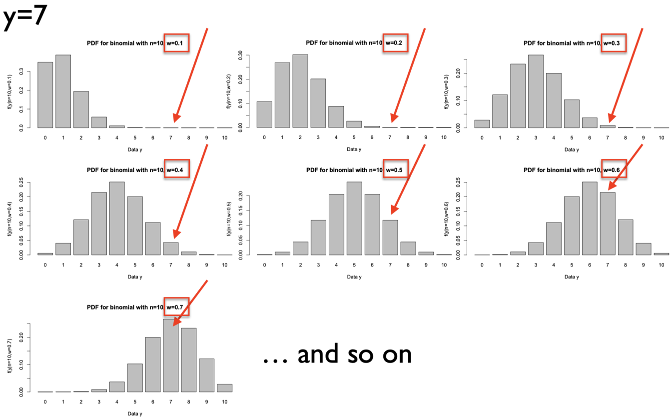
- note: a likelihood function does not need to sum to 1.0
- For our one-dimensional binomial example the likelihood function for y=7 and n=10 is

$$\mathcal{L}(w|n=10, y=7) = \frac{10!}{7!(10-7)!} w^7 (1-w)^{10-7}; (0 \leq w \leq 1)$$

But, what is the value of w???

Likelihood Function - Iterate Through Variable

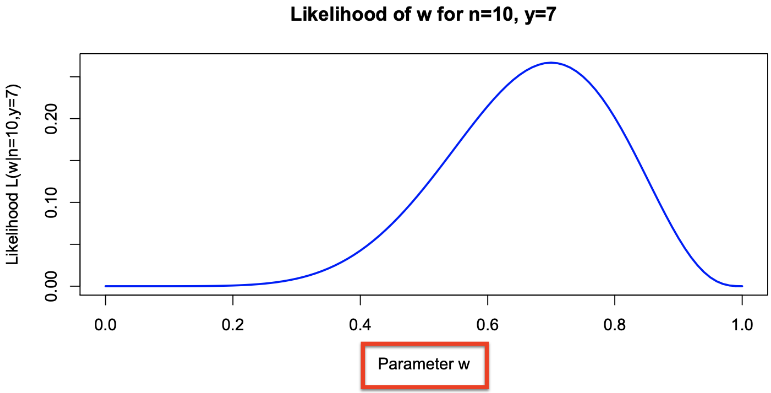
Let's try all value of w between 0 and 1



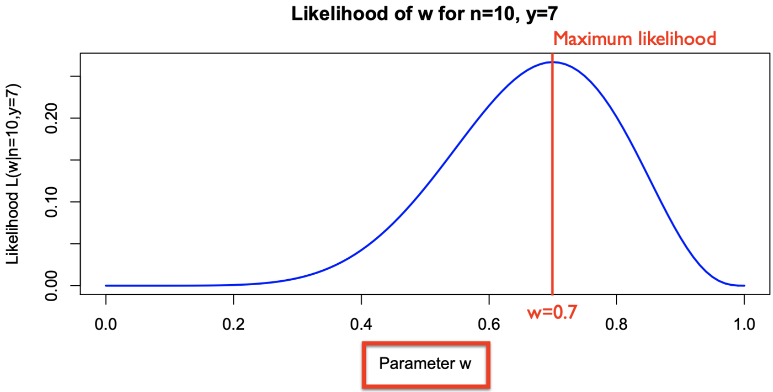
… and so on

Notice $\mathcal{L}(w|n = 10, y = 7)$ is highest when $w = 0.7$

Graphing the Likelihood Function



Graphing the Likelihood Function



$w = 0.7$ is the Maximum Likelihood Estimate!!!

Maximum Likelihood Estimate (MLE)

- find the probability distribution (the model) that makes the observed data most likely
- seek the value of the parameter vector w that maximizes the likelihood function

$\mathcal{L}(w|y)$ - the resulting parameter vector w is known as the MLE estimate

Maximum Likelihood Estimate (MLE)

Three ways of finding the MLE

1. Analytical: use calculus to solve for the parameter value(s) w that result in a peak
2. Brute Force: exhaustive search through parameter space in a grid
3. Optimization: use non-linear optimization (e.g. gradient descent) to iteratively find the peak

Numerical Considerations

- we saw before that the PDF for observed data, $y = (y_1, y_2, ..., y_n)$ given a parameter vector w, can be expressed as the **product (multiply) of PDFs for individual observations**

$$\mathcal{L}(w|y_1, y_2, ..., y_n) = \mathcal{L}_1(w|y_1)\mathcal{L}_2(w|y_2)...\mathcal{L}_n(w|y_n)$$

- multiplying together a lot of values that lie between 0 and 1, (as many as there are data points) will result in a very small number
- in fact the more data, the smaller the resulting product will be
- computers are not good at representing very small numbers

Numerical Considerations

- solution: take the logarithm
- this reformulates the series of products, as a series of sums
- the more data, the higher the resulting sum

$$ln[\mathcal{L}_1(w|y_1)\mathcal{L}_2(w|y_2)...\mathcal{L}_n(w|y_n)] = ln[\mathcal{L}_1(w|y_1) + \mathcal{L}_2(w|y_2)+, ..., \mathcal{L}_n(w|y_n)]$$

Numerical Considerations

- another problem: most optimization algorithms are formulated in terms of minimizing an objective function, not maximizing
- solution: rather than maximizing the log-likelihood, we will minimize the negative log-likelihood
- find w that minimizes:

$$argmin_w \left[-1.0 \left(ln \left[\mathcal{L}_1(w|y_1) + \mathcal{L}_2(w|y_2) +, ..., \mathcal{L}_n(w|y_n) \right] \right) \right]$$

An Example

- Let's say I claim I can correctly identify coffee quality between Little Goat and Starbucks coffee
- My lab designs an experiment to test me
- They give me 20 cups of coffee in random order and I have to say "Goat" or "Starbucks"
- Observed data: I get 16 correct, 4 incorrect
- what model explains the observed data?

An Example

- This experiment can be modelled as 20 Bernoulli trials (outcome of each trial is random and can be either of two possible outcomes, "success" and "failure")
- we know PDF is binomial, which has 2 parameters: n (# trials) and w (prob of a success on a given trial)
- equivalent to asking, what is the value of the parameter w?
- high w (e.g. near 1.0) means I have a good ability to discriminate
- w near 0.5 means I am flipping a coin

Likelihood Function

Likelihood Function:

$$\mathcal{L}(w|n,y) = \frac{n!}{y!(n-y)!} w^y (1-w)^{n-y}$$

Log Likelihood Function:

$$ln[\mathcal{L}(w|n,y)] = ln\left(\frac{n!}{y!(n-y)!}\right) + y \cdot ln(w) + (n-y) \cdot ln(1-w)$$

Tips: $ln(x \cdot y) = ln(x) + ln(y)$; $ln(e) = 1$; $\frac{d[ln(x)]}{dx} = \frac{1}{x}$

MLE - ANALYTICAL

MLE - ANALYTICAL

We want:

$$\frac{d}{dw}\Big(ln[\mathcal{L}(w|n,y)]\Big) = 0$$

Log Likelihood Function:

$$ln[\mathcal{L}(w|n,y)] = ln\left(\frac{n!}{y!(n-y)!}\right) + y \cdot ln(w) + (n-y) \cdot ln(1-w)$$

Taking the partial derivative of the log likelihood function:

$$\frac{d}{dw}\Big(ln[\mathcal{L}(w|n,y)]\Big) = \frac{d}{dw}\left(ln\left(\frac{n!}{y!(n-y)!}\right) + y \cdot ln(w) + (n-y) \cdot ln(1-w)\right) = 0$$

$$\frac{d}{dw}\Big(ln[\mathcal{L}(w|n,y)]\Big) = 0 + \frac{n}{w} - \frac{n-y}{1-w} = 0$$

MLE - ANALYTICAL

$$\frac{n}{w} - \frac{n-y}{1-w} = 0$$

Finding the common denominator:

$$\frac{y(1-w)}{w(1-w)} - \frac{w(n-y)}{w(1-w)} = 0$$

$$\frac{y(1-w) - w(n-y)}{w(1-w)} = 0$$

$$\frac{y - y \cdot w - w \cdot n + y \cdot w}{w(1-w)} = 0$$

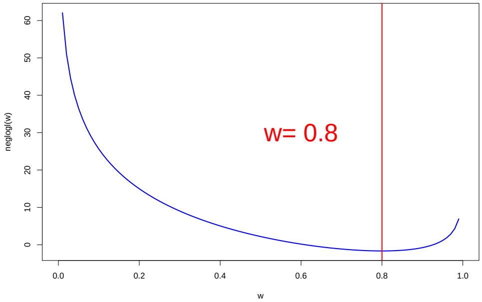
$$w = \frac{y}{n}$$

$\text{MLE} = 0.8 = \frac{16}{20}$

MLE - BRUTE FORCE

MLE - BRUTE FORCE

```r
neglogl <- function(w) {
  loglik <- log(116280) + 16 * log(w) + 4 * log(1-w)
  return(-1 * loglik)
}
w <- seq(0,1,.01) # iterate through a range of w's
plot(w, neglogl(w), type="l", col="blue", lwd=2)
imin <- which(neglogl(w)==min(neglogl(w)))
abline(v=w[imin], col="red", lwd=2)
text(.6, 30, paste("w=",w[imin]),col="red", cex = 3)
```



MLE - OPTIMIZER

MLE - OPTIMIZER

```r
neglogl <- function(w) {
  loglik <- log(116280) + 16 * log(w) + 4 * log(1-w)
  return(-1 * loglik)
}
opt <- nlm(f=neglogl, p=0.5)
```

```
## Warning in log(1 - w): NaNs produced

## Warning in nlm(f = neglogl, p = 0.5): NA/Inf replaced by maximum positive value

## Warning in log(1 - w): NaNs produced

## Warning in nlm(f = neglogl, p = 0.5): NA/Inf replaced by maximum positive value
```

```r
opt$estimate
```

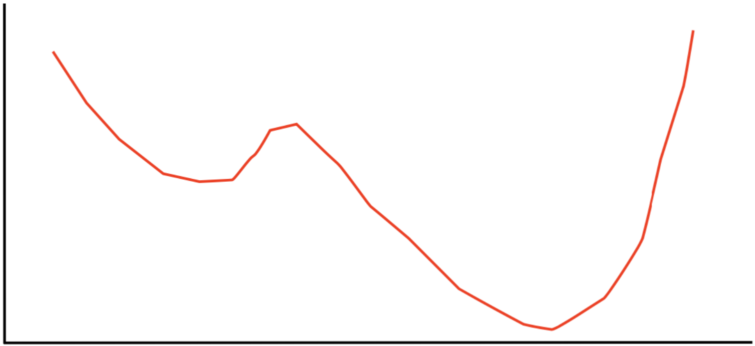
```
## [1] 0.7999995
```

Finds the Maximum Likelihood Estimate: 0.8

MLE in general

- MLE for many distributions are known (look it up)
- MLE for more complex models can sometimes be determined analytically
- Often however not possible/feasible
- Iterative optimization is a common method in these cases
  - local minima

Optimization and Local Minima



General Procedure

- If you can write an equation for the Likelihood function
- i.e. probability of obtaining your observed data, given a model with parameter(s) w
- then you can find the MLE for w
- i.e. you can find the model that is most likely to generate your data

Hypothesis Testing

- We can use the Likelihood Ratio Test to compare two models
- Little Goat and Starbucks Example
- 16 correct out of 20 trials
- our MLE for p was 0.80
- let's test this against a null hypothesis that p=0.50

Likelihood Ratio Test

- test statistic D is a ratio:

$$D = -2 \cdot ln\left(\frac{likelihood\ for\ null\ model}{(likelihood\ for\ alternative\ model}\right)$$

$$D = -2 \cdot ln(likelihood\ for\ null\ model) + (likelihood\ for\ alternative\ model)$$

Likelihood Ratio Test

- the probability distribution of test statistic D is approximately a chi-squared distribution with $df = df_2 - df_1$
- df_1 and df_2 are number of free parameters of models 1 (null) and 2 (alternative), respectively.
    - $df_1 = 0$ for the null model since assuming w is set to 0.5 (not a free parameter)
    - $df_2 = 1$ for the alternative model since assuming w is a free parameter

Likelihood Ratio Test

$$\mathcal{L}(w|n,y) = \frac{n!}{y!(n-y)!} w^y (1-w)^{n-y}$$

- our data: 16 correct and 4 incorrect

$$Null\ model = -2 \cdot ln[L(w = 0.5|y = 16, n = 20)] = 16.29966$$

- MLE of w is, $w = 0.8$.

$$Alternative\ model = -2 \cdot ln[L(w = 0.8|y = 16, n = 20)] = -4.82984$$

$$D = -2 \cdot [ln(likelihood\ for\ null\ model) + ln(likelihood\ for\ alternative\ model)]$$

$$D = 16.29966 - 4.82984 = 11.46982$$

Likelihood Ratio Test

$$D = 11.46982$$

- now compute a p-value using chi-square distribution with df = 1-0 = 1

```r
pval <- pchisq(q=11.46982, df=1, lower.tail=FALSE)
pval
```

```
## [1] 0.0007073553
```

We can reject the null with a Type 1 error rate of 0.00071. Thus, Josh can detect differences in Little Goat coffee compared to Starbucks coffee compared to chance.

Beyond the Binomial

Any model (lots of examples online)

Normal Distribution:

$$p(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Linear Regression:

$$p(y_i|x_i, \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-(\beta_0+\beta_1 \cdot x_i))^2}{2\sigma^2}}$$

MLE Normal Distribution

$$p(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$p(x_i|\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\mathcal{L}(\mu,\sigma^2|x_1,x_2,\ldots,x_n) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

MLE Normal Distribution

Likelihood function:

$$\mathcal{L}(\mu, \sigma^2 | x_1, x_2, ..., x_n) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{\Sigma_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

Taking the log:

$$\ln[\mathcal{L}(\mu, \sigma^2 | x_1, x_2, ..., x_n)] = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \Sigma_{i=1}^{n} (x_i - \mu)^2$$

MLE Normal Distribution

Take the partial derivatives of the log likelihood function equal to 0.

$$\frac{\partial(ln[\mathcal{L}(\mu,\sigma^2|x_1,x_2,...,x_n)])}{\partial\mu}=\frac{\partial\Big(-\frac{n}{2}ln(2\pi)-\frac{n}{2}ln(\sigma^2)-\frac{1}{2\sigma^2}\Sigma_{i=1}^{n}(x_i-\mu)^2\Big)}{\partial\mu}=0$$

$$\frac{-2\cdot\Sigma_{i=1}^{n}(x_i-\mu)\cdot(-1.0)}{2\sigma^2}=0$$

$$+\frac{\Sigma_{i=1}^{n}(x_i-\mu)}{\sigma^2}=0$$

$$\frac{\Sigma_{i=1}^{n}x_i-n\mu}{\sigma^2}=0$$

$$\mu=\frac{\Sigma_{i=1}^{n}x_i}{n}$$

MLE Normal Distribution

$$\frac{\partial(ln[\mathcal{L}(\mu,\sigma^2|x_1,x_2,...,x_n)])}{\partial\sigma^2} = \frac{\partial\left(-\frac{n}{2}ln(2\pi)-\frac{n}{2}ln(\sigma^2)-\frac{1}{2\sigma^2}\Sigma_{i=1}^{n}(x_i-\mu)^2\right)}{\partial\sigma^2} = 0$$

$$-\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\Sigma_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$-\frac{n}{2\sigma^2}+\frac{\Sigma_{i=1}^{n}(x_i-\mu)^2}{2(\sigma^2)^2} = 0$$

$$\frac{1}{2\sigma^2}\left[\frac{\Sigma_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}-n\right] = 0$$

$$\sigma^2 = \frac{\Sigma_{i=1}^{n}(x_i-\mu)^2}{n}$$

MLE Normal Distribution

Summary:

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

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

Next Week

- Bayesian Statistics
  - Priors, Likelihoods, Posterior Distributions
  - Continually updating probabilities based on new information