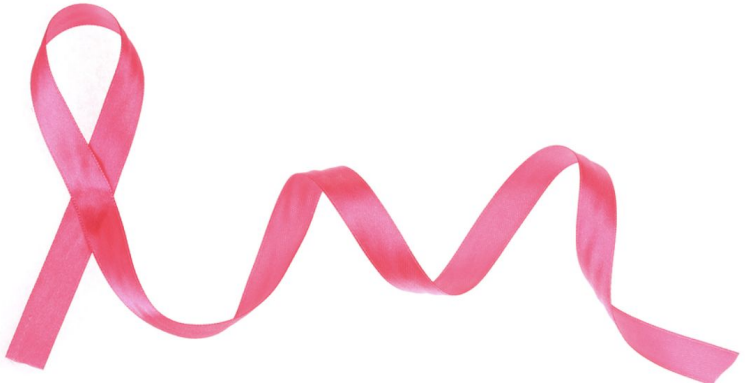
UNIVERSITY of WASHINGTON

Estimation for cancer screening models using deconvolution

Huayue (Lucia) Zou
Mentor: Antonio Olivas

Breast Cancer's Popularity



How do we know that we’re getting a Cancer or NOT?

We need a SCREENING TEST!

But with selections

WHY SELECT?

High Cost,
Spend Time,
Higher Efficiency

Test Results

	Disease	Non-disease
Test: Positive	True Positive	False Positive
Test: Negative	False Negative	True Negative

Sensitivity (\beta) = TP/TP+FN → Testing women has the cancer

Specificity = TN/TN+FP

Estimation Process

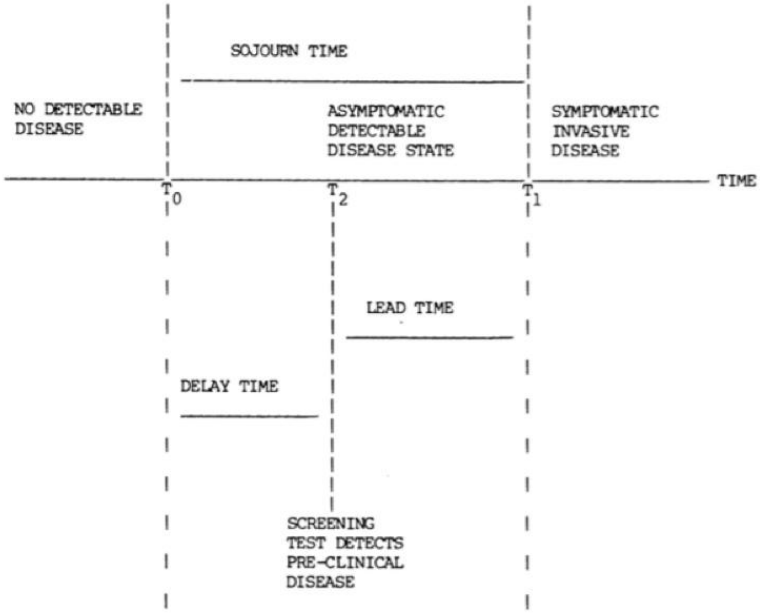


Figure 1. Schema for the progression of a chronic disease, with the intervention of an early detection screening test.

Estimation Process

Figure 1. Schema for the progression of a chronic disease, with the intervention of an early detection screening test.

Estimating on 3 parameters;

3 questions we're interested in

3 Parameters

- Gamma (γ)
- Lambda (λ)
- Beta (β)

3 Questions

- WHEN should I start my screening?
- HOW FREQUENTLY should I take screening?
- HOW GOOD is the testing (test β)?

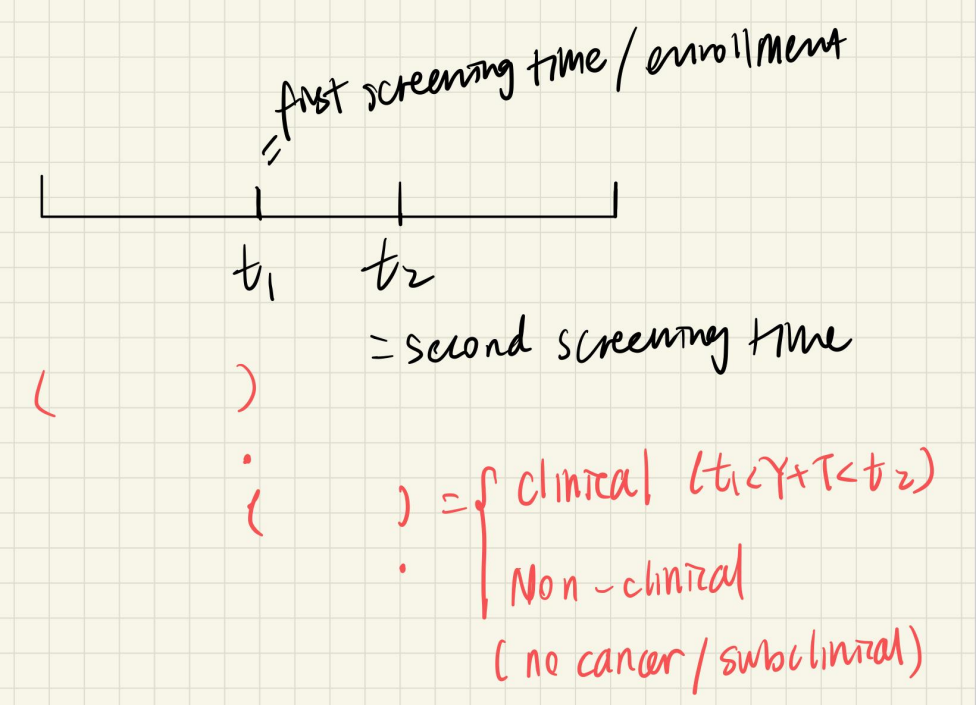
Exponential Distribution

- $X \sim Exp(\lambda)$, Range of $X = (0, \infty)$
- PDF (Probability Density Function) of X:
  - $f(x) = \lambda e^{(-\lambda x)}, x >= 0$
- CDF (Cumulative Distribution Function) of X:
  - $F(x) = 1 - e^{(-\lambda x)}$

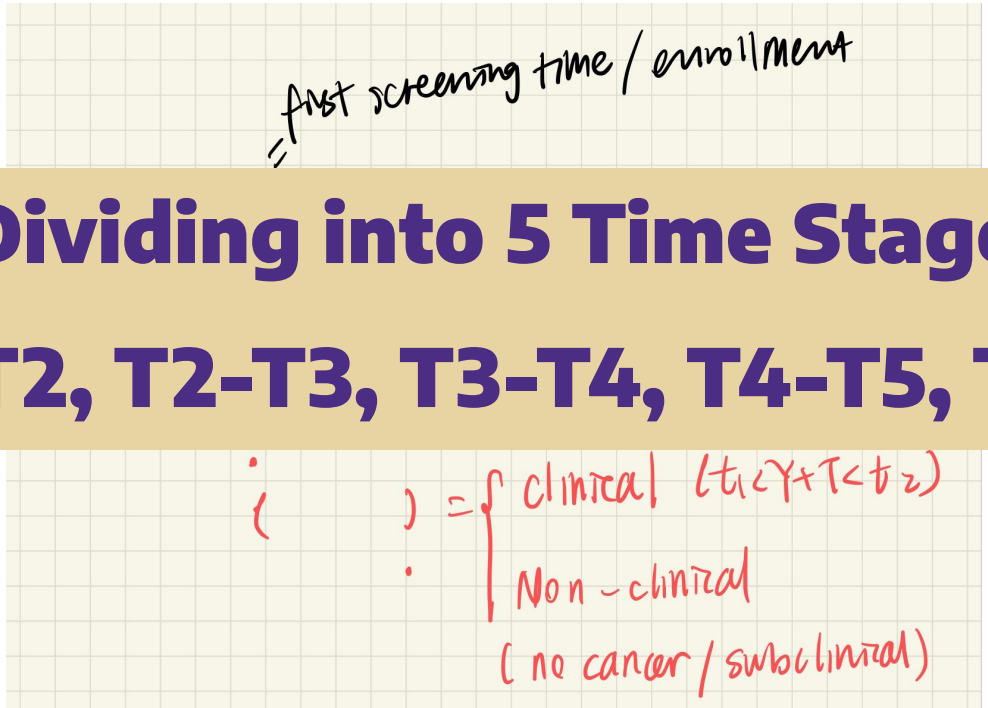
Functions Setting

- Suppose:
- Time to Onset Breast Cancer AS Y:
  - $Y \sim Exp(\gamma)$
- Sojourn Time of Breast Cancer AS T:
  - $T \sim Exp(\lambda)$

Dividing Stages



Dividing Stages



Dividing into 5 Time Stages:

T1-T2, T2-T3, T3-T4, T4-T5, T5-T6

3 Cases in Each Stage

- **Subclinical (Screen-detected):**
  - The women enroll in the program & has positive test at Tx
  - The onset time is actually less than the entry time
- **Clinical (Clinical-detected):**
  - The onset time is between this time stage
  - Has symptoms
- **Not-detected:**
  - move to the next stage as the entry probabilities

How to Calculate the Probabilities in Each Stage?

FIRST ENROLLING PROB.

P(Y+T>t1)

- **By integrating**

```
prob_enroll = function(g,l,t1)
  (g/(g-l))*exp(-l*t1)-(l/(g-l))*exp(-g*t1)
prob_enroll(g=0.0025, l=0.4, t1=50)
P0<-prob_enroll(g=0.0025, l=0.4, t1=50)
```

```
[1] 0.8880472
```

T1-T2 (t1 < Y+T < t2):

Subclinical (Screen-detected Case)

- **P(Y <= t1 given (Y+T > t1))*β**
- **Apply integrating process**

```
prob_screen_t1 = function(g,l,t1,beta)
  (exp(-l*t1)-exp(-g*t1)+(l/(g-l))*(exp(-l*t1))-(l/(g-l))*(exp(-g*t1)))*(beta)
prob_screen_t1(g=0.0025,l=0.4,t1=50,beta=0.8)
P1<-prob_screen_t1(g=0.0025,l=0.4,t1=50,beta=0.8)
```

```
[1] 0.004440236
```

WHY MULTIPLYING β?

- Sensitivity (β) = TP/TP+FN
  - Actual probability of there is a cancer
- We catch it for Y <= t1 given (Y+T > t1), it's actually happened in the right period! So we multiply by β

T1-T2 (t1 < Y+T < t2):

Clinical (Clinical-detected Case)

- **P(Y <= t1 given (t1 < Y+T < t2))*(1-β) +**
  P(t1 < Y <= t2 given (t1 < Y+T < t2))
- **Apply integrating process**

```
prob_clinical_t = function(g,l,t1,t2,beta)
  (exp(-g*t1)-exp(-g*t2)-(exp(-l*t2)*(g/(l-g)*(exp((l-g)*t2)-exp((l-g)*t1)))))+
  (exp(-l*t1)-exp(-l*t2))*((g/(l-g))*exp((l-g)*t1)-(g/(l-g)))*(1-beta)
prob_clinical_t(g=0.0025,l=0.4,t1=50,t2=51,beta=0.8)
P2<-prob_clinical_t(g=0.0025,l=0.4,t1=50,t2=51,beta=0.8)
```

```
[1] 0.0007534884
```

WHY MULTIPLYING (1-β)?

- Sensitivity (β) = TP/TP+FN
  - Actual probability of there is a cancer
- When the time stage is given that t1 < Y+T < t2, but we catch it when Y <= t1, which is not belonged to this period: we've missed it before;

  So we multiply by (1-β)!

SECOND ENROLLING PROB.

P(Y+T>t1) - P1(Subclinical) - P1(Clinical)

```
[1] 0.8828535
```

T2-T3 (t2 < Y+T < t3):

Subclinical (Screen-detected Case)

- **HAS CHANGED! It's different from T1-T2 Case.**
- **P(Y <= t1 given (Y+T > t2))*(1-β)*β +**

  P(t1 < Y < t2 given (Y+T > t2))*β
- **AND now, we need to multiply by β and (1-β)*β!**

WHY MULTIPLYING?

- Sensitivity (β) = TP/TP+FN
  - Actual probability of there is a cancer
- When Y< t1, we then need to capture it from the period T1-T2; however, we miss it! So it is a miss; multiply by (1-β)
- However, we still catch it! So we multiply (1-β)*β
- AS FOR t1 < Y < t2 given (Y+T > t2), it's actually happened in the right period! So we multiply by β

T2-T3 (t2 < Y+T < t3):

Subclinical (Screen-detected Case)

- P(Y <= t1 given (Y+T > t2))*(1-β)*β +
  P(t1 < Y < t2 given (Y+T > t2))*β
- Apply integrating

```
[1] 0.002048046
```

```
prob_screen_t2 = function(g,l,t1,t2,beta){
  (exp(-l*t2)*g/(l-g)*(exp((l-g)*t2)-exp((l-g)*t1)))*beta+
    (exp(-l*t2)*(g/(l-g))*(exp((l-g)*t1)-1))*(1-beta)*beta
}
prob_screen_t2(g=0.0025,l=0.4,t1=50,t2=51,beta=0.8)
P4 <- prob_screen_t2(g=0.0025,l=0.4,t1=50,t2=51,beta=0.8)
```

T2-T3 (t2 < Y+T < t3):

Clinical (Clinical-detected Case)

- P(Y <= t1 given (t2 < Y+T < t3))*(1-β)^2 +
  P(t1 < Y < t2 given (t2 < Y+T < t3))*(1-β)+
  P(t2 < Y < t3 given (t2 < Y+T < t3))
- Apply integrating

```
[1] 0.0005553565
```

```
prob_clinical_t2_t3 = function(g,l,t1,t2,t3,beta){
  ((g/(l-g))*(exp(-l*t2)-exp(-l*t3))*(exp((l-g)*t1)-(1))*((1-beta)^2))+
    ((g/(l-g))*(exp(-l*t2)-exp(-l*t3))*(exp((l-g)*t2)-exp((l-g)*t1))*(1-beta))+
    (exp(-g*t2)-exp(-g*t3)-(exp(-l*t3)*(g/(l-g)*(exp((l-g)*t3)-exp((l-g)*t2)))))
}
prob_clinical_t2_t3(g=0.0025,l=0.4,t1=50,t2=51,t3=52,beta=0.8)
P5 <- prob_clinical_t2_t3(g=0.0025,l=0.4,t1=50,t2=51,t3=52,beta=0.8)
```

WHY MULTIPLYING?

- Sensitivity (β) = TP/TP+FN
  - Actual probability of there is a cancer
- When the time stage is given that t2 < Y+T < t3, but we catch it when Y <= t1, which is not belonged to this period: we've missed TWICE (before t1 & t1~ t2 stage)

  So we multiply by (1-β)^2!
- Same logic applies.

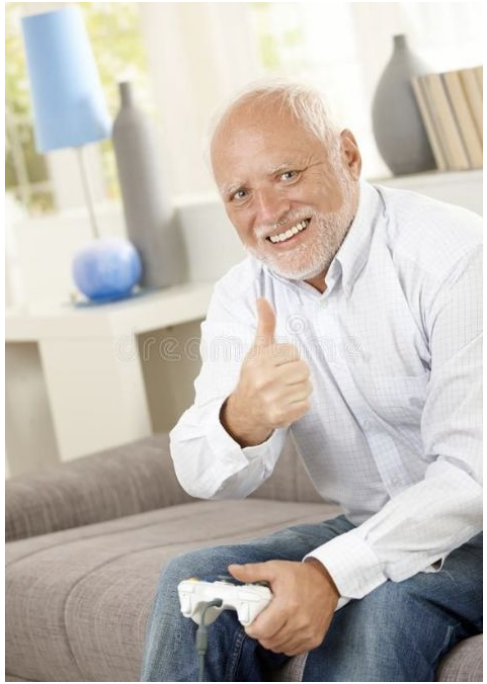
THIRD ENROLLING PROB.

P(2nd enroll) - P2(Subclinical) - P2(Clinical)

```
[1] 0.8802501
```

SAME LOGIC APPLIES....

T3 - T4
T4 - T5
T5 - T6



T3 - T4

```
prob_enroll_t3 <- P6
prob_enroll_t3

prob_screen_t3 = function(g,l,t1,t2,t3,beta)
  (exp(-l*t3)*g/(l-g)*(exp((l-g)*t3)-exp((l-g)*t2)))*beta+
  (exp(-l*t3)*g/(l-g)*(exp((l-g)*t2)-exp((l-g)*t1)))*(1-beta)*beta+
  (exp(-l*t3)*(g/(l-g))*(exp((l-g)*t1)-1))*((1-beta)^2)*beta
prob_screen_t3(g=0.0025,l=0.4,t1=50,t2=51,t3=52,beta=0.8)
P7 <- prob_screen_t3(g=0.0025,l=0.4,t1=50,t2=51,t3=52,beta=0.8)

prob_clinical_t3_t4 = function(g,l,t1,t2,t3,t4,beta){
((g/(l-g))*(exp(-l*t3)-exp(-l*t4))*(exp((l-g)*t1)-(1))*((1-beta)^3))+
    ((g/(l-g))*(exp(-l*t3)-exp(-l*t4))*(exp((l-g)*t2)-exp((l-g)*t1))*((1-beta)^2))+
    ((g/(l-g))*(exp(-l*t3)-exp(-l*t4))*(exp((l-g)*t3)-exp((l-g)*t2))*(1-beta))+
    (exp(-g*t3)-exp(-g*t4)-(exp(-l*t4)*(g/(l-g)*(exp((l-g)*t4)-exp((l-g)*t3)))))}
prob_clinical_t3_t4(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,beta=0.8)
P8 <- prob_clinical_t3_t4(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,beta=0.8)

P9 <- prob_enroll_t3-P7-P8
P9
```

```
[1] 0.8802501
[1] 0.001723712
[1] 0.0005276597
[1] 0.8779987
```

T4 - T5

```
prob_enroll_t4
prob_screen_t4 = function(g,l,t1,t2,t3,t4,beta)
  (exp(-l*t4)*g/(l-g)*(exp((l-g)*t4)-exp((l-g)*t3)))*beta+
  (exp(-l*t4)*g/(l-g)*(exp((l-g)*t3)-exp((l-g)*t2)))*(1-beta)*beta+
  (exp(-l*t4)*g/(l-g)*(exp((l-g)*t2)-exp((l-g)*t1)))*((1-beta)^2)*beta+
  (exp(-l*t4)*(g/(l-g))*(exp((l-g)*t1)-1))*((1-beta)^3)*beta
prob_screen_t4(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,beta=0.8)
P10 <- prob_screen_t4(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,beta=0.8)

prob_clinical_t4_t5 = function(g,l,t1,t2,t3,t4,t5,beta){
((g/(l-g))*(exp(-l*t4)-exp(-l*t5))*(exp((l-g)*t1)-(1))*((1-beta)^4))+
    ((g/(l-g))*(exp(-l*t4)-exp(-l*t5))*(exp((l-g)*t2)-exp((l-g)*t1))*((1-beta)^3))+
    ((g/(l-g))*(exp(-l*t4)-exp(-l*t5))*(exp((l-g)*t3)-exp((l-g)*t2))*((1-beta)^2))+
    ((g/(l-g))*(exp(-l*t4)-exp(-l*t5))*(exp((l-g)*t4)-exp((l-g)*t3))*(1-beta))+
    (exp(-g*t4)-exp(-g*t5)-(exp(-l*t5)*(g/(l-g)*(exp((l-g)*t5)-exp((l-g)*t4)))))}
prob_clinical_t4_t5(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,t5=54,beta=0.8)
P11 <- prob_clinical_t4_t5(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,t5=54,beta=0.8)

P12 <- prob_enroll_t4-P10-P11
P12
```

```
[1] 0.8779987
[1] 0.001676612
[1] 0.000522815
[1] 0.8757993
```

T5 - T6

```
prob_enroll_t5
prob_screen_t5 = function(g,l,t1,t2,t3,t4,t5,beta)
(exp(-l*t5)*g/(l-g)*(exp((l-g)*t5)-exp((l-g)*t4)))*beta+
  (exp(-l*t5)*g/(l-g)*(exp((l-g)*t4)-exp((l-g)*t3)))*(1-beta)*beta+
  (exp(-l*t5)*g/(l-g)*(exp((l-g)*t3)-exp((l-g)*t2)))*((1-beta)^2)*beta+
  (exp(-l*t5)*g/(l-g)*(exp((l-g)*t2)-exp((l-g)*t1)))*((1-beta)^3)*beta+
  (exp(-l*t5)*(g/(l-g))*(exp((l-g)*t1)-1))*((1-beta)^4)*beta
prob_screen_t5(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,t5=54,beta=0.8)
P13 <- prob_screen_t5(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,t5=54,beta=0.8)
prob_clinical_t5_t6 = function(g,l,t1,t2,t3,t4,t5,t6,beta){
((g/(l-g))*(exp(-l*t5)-exp(-l*t6))*(exp((l-g)*t1)-(1))*((1-beta)^5))+
    ((g/(l-g))*(exp(-l*t5)-exp(-l*t6))*(exp((l-g)*t2)-exp((l-g)*t1))*((1-beta)^4))+
    ((g/(l-g))*(exp(-l*t5)-exp(-l*t6))*(exp((l-g)*t3)-exp((l-g)*t2))*((1-beta)^3))+
    ((g/(l-g))*(exp(-l*t5)-exp(-l*t6))*(exp((l-g)*t4)-exp((l-g)*t3))*((1-beta)^2))+
    ((g/(l-g))*(exp(-l*t5)-exp(-l*t6))*(exp((l-g)*t5)-exp((l-g)*t4))*(1-beta))+
    (exp(-g*t5)-exp(-g*t6)-(exp(-l*t6)*(g/(l-g)*(exp((l-g)*t6)-exp((l-g)*t5)))))}
prob_clinical_t5_t6(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,t5=54,t6=55,beta=0.8)
P14 <- prob_clinical_t5_t6(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,t5=54,t6=55,beta=0.8)
P15 <- prob_enroll_t5-P13-P14
P15
```

```
[1] 0.8757993
[1] 0.001666688
[1] 0.0005210367
[1] 0.8736115
```

MOVE TO NEXT STEP: ANSWER 3 QUESTIONS

1st: Create Dataframe

Assumption:

Breast cancer happens after women are 40 years old; there is no cancer happening before 40 years

1st: Create Dataframe

Web Table 1

Screening round	No. of women	Screen-detected cases	Interval-detected cases
1	19711	142	15
2	17669	66	10
3	17347	43	9
4	17193	54	9
5	9876	28	5

CNBSS-2. Grouped data from the Canadian Breast Cancer Screening Study-2 [3]. "No. of women" is the number of women who attended all screening rounds up to and including the current round.

Source: Marc D et al., "Identification of the Fraction of Indolent Tumors and Associated Overdiagnosis in Breast Cancer Screening Trials"

1st: Create Dataframe

```r
df = data.frame(t = 9+1:5,
                n = c(19711,17669,17347,17193,9876),
                ns = c(142,66,43,54,28),
                nc = c(15,10,9,9,5))
```

*In this case we assume after 40 years, while in the functions I created before, it's begin from 50. So t = 9+1:5

2nd: MLE

WHAT IS Maximum Likelihood Estimation?

2nd: MLE

Process:

1. **Take LOG of function**
2. **Calculate probabilities**
3. **Bring it back to res. (a method used in R to optimize and find the maximum likelihood estimates)**

2nd: MLE

```
#log-likelihood function
llf = function(g,l,beta) {
  q0 = prob_enroll(g,l,df[1,1])
  p11 = prob_screen_t1(g,l,t1=df[1,1],beta)
  p12 = prob_clinical_t(g,l,t1=df[1,1],t2=df[2,1],beta)
  p13 = q0 - p11 - p12
  p21 = prob_screen_t2(g,l,t1=df[1,1],t2=df[2,1],beta)
  p22 = prob_clinical_t2_t3(g,l,t1=df[1,1],t2=df[2,1],t3=df[3,1],beta)
  p23 = p13 - p21 - p22
  p31 = prob_screen_t3(g,l,t1=df[1,1],t2=df[2,1],t3=df[3,1],beta)
  p32 = prob_clinical_t3_t4(g,l,t1=df[1,1],t2=df[2,1],t3=df[3,1],t4=df[4,1],beta)
  p33 = p23 - p31 - p32
  p41 = prob_screen_t4(g,l,t1=df[1,1],t2=df[2,1],t3=df[3,1],t4=df[4,1],beta)
  p42 = prob_clinical_t4_t5(g,l,t1=df[1,1],t2=df[2,1],t3=df[3,1],t4=df[4,1],t5=df[5,1],beta)
  p43 = p33 - p41 - p42
  p51 = prob_screen_t5(g,l,t1=df[1,1],t2=df[2,1],t3=df[3,1],t4=df[4,1],t5=df[5,1],beta)
  p52 = prob_clinical_t5_t6(g,l,t1=df[1,1],t2=df[2,1],t3=df[3,1],t4=df[4,1],t5=df[5,1],t6=df[5,1]+1,beta)
  p53 = p43 - p51 - p52
```

2nd: MLE

```
q11 = p11/q0
q12 = p12/q0
q13 = p13/q0
q21 = p21/p13
q22 = p22/p13
q23 = p23/p13
q31 = p31/p23
q32 = p32/p23
q33 = p33/p23
q41 = p41/p33
q42 = p42/p33
q43 = p43/p33
q51 = p51/p43
q52 = p52/p43
q53 = p53/p43
```

```
c1 = df[1,3]*log(q11) + df[1,4]*log(q12) + (df[1,2]-df[1,3]-df[1,4])*log(q13)
c2 = df[2,3]*log(q21) + df[2,4]*log(q22) + (df[2,2]-df[2,3]-df[2,4])*log(q23)
c3 = df[3,3]*log(q31) + df[3,4]*log(q32) + (df[3,2]-df[3,3]-df[3,4])*log(q33)
c4 = df[4,3]*log(q41) + df[4,4]*log(q42) + (df[4,2]-df[4,3]-df[4,4])*log(q43)
c5 = df[5,3]*log(q51) + df[5,4]*log(q52) + (df[5,2]-df[5,3]-df[5,4])*log(q53)

res = -(c1+c2+c3+c4+c5)
return(res)
```

```{r}
llf(g = 0.01, l = 0.3, beta = 0.7)
```

```
[1] 2956.97
```

*Apply Values

3nd Optimize

```
x = optim(c(0.001,0.0021,0.1),function(m)llf(m[1],m[2],m[3]))
x
```

```
$par
[1] 0.003090176 0.302153689 0.805702854

$value
[1] 2542.356

$counts
function gradient
     216       NA

$convergence
[1] 0

$message
NULL
```

Conclusion

1. WHEN should I start my screening?

- # mle. g = 0.0031
- # choose the years start screening
- # Suppose we choose the prob. of breast cancer that women onset rate is 1.5% & 3%
    - P[T<=t] = 0.015/0.03
    - 1st, use CDF to calculate out the est(t*) = log(1-rate)/-gamma
    - # 1.5% log(0.985)/(-0.003148025) = 4.8 -- start 5 years after 40 years
    - # 3% log(0.97)/(-0.003148025) = 9.68 -- start 10 years after 40 years (higher prevalence, and could spend less money since screening test is expensive)

Conclusion

2. HOW FREQUENTLY should they take screening?

- # mle. l = 0.3022 --> E[Sojourn time] = 1/0.3022 = 3.3091
- avg onset--clinical 3.3 yrs
    - So we could suggest: take screening time for 2-3 yrs

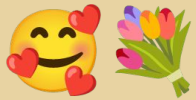
Conclusion

3. HOW GOOD is the testing?

- # mle. beta = 0.8057 --> sensitivity (cancer)
- For women with breast cancer, there is 81% probability of detecting that actually there is cancer.

FUTURE STEPS

- Dive into calculating the Confidence Interval;
- Whether the given model's data is correct?



THANK YOU FOR LISTENING!!

REFERENCES

Fancher, Tressa. “Breast Cancer Awareness Month.” Breast Cancer Awareness Month, Blogger, 16 Oct. 2017, bplolinenews.blogspot.com/2017/10/breast-cancer-awareness-month.html.

Golden, Richard. “LM101-055: How to Learn Statistical Regularities Using Map and Maximum Likelihood Estimation (Rerun).” Learning Machines 101, 16 Aug. 2016, www.learningmachines101.com/lm101-055-learn-statistical-regularities-using-map-maximum-likelihood-estimation-rerun/

Marc D et al., “Identification of the Fraction of Indolent Tumors and Associated Overdiagnosis in Breast Cancer Screening Trials”