

연구노트 (생존분석)

유재성¹

최종 수정일자: 2017/12/29

Contents

I 서론	8
1 서론	8
2 임상적 연구 결과 데이터	9
3 신뢰성 공학	10
II 생존 자료의 분석	12
4 생존 분석 방법의 분류	14
5 생존분석 관련 소프트웨어 패키지	15
III 생존 함수	16
6 반응변수	16
7 생존 함수(survival function), 신뢰도 함수(reliability function)	17
8 위험 함수(hazard function), 위험률 함수(hazard rate function; 고장률 함수 failure rate function)	18
9 누적 위험 함수(cumulative hazard function), 누적 고장률 함수(cumulative failure rate function)	19
10 생존 함수간 관계	20
11 위험률 함수와 생존함수의 1:1 대응	21
11.1 역공식을 이용한 위험함수와 생존함수의 관계	21
11.2 분포의 비모수적 분류	23
11.2.1 비모수적 군	23
11.2.2 비모수적 군들의 상호관계	24
IV 척도 함수	25
12 '고장(재발) 발생까지 시간 길이의 길고 짧음'에 따른 척도 함수	25
12.1 생존 함수(Survival Function)	25
12.2 평균 수명(mean time to failure)	25
12.3 평균 고장 간격(mean time between failure)	26
12.4 백분위수	26
12.5 Bearing Life	26
13 '고장(재발) 발생까지 시간 길이의 길고 짧음'에 따른 척도 함수	27
13.1 평균 잔여 수명 함수	27
13.2 가용도 함수	28
13.3 위험률(고장률)	29
13.3.1 비수리계(non-repairable system)의 위험률(고장률)	29
13.3.2 수리계(repairable system)의 위험률(고장률)	29
13.3.3 평균 위험률(평균 고장률, average failure rate)	29
V 수명 자료(Life Data), 중도 절단 자료(Censored Data)	30
14 우증도절단(right censoring)	32
14.1 완전 자료(complete data)	33
14.2 Type 1 우증도절단 자료(정시증단자료)	33
14.3 Type 2 우증도절단 자료(정수증단자료)	33
14.4 임의 우증도절단 자료	34
14.4.1 Kaplan-Meier Estimation, Product Limit Estimation	34
14.4.2 Life-table Method	35

¹고려대학교 컴퓨터학과 박사과정. praster1@gmail.com

VI 모수적 방법을 이용한 생존함수의 추정

37

16 확률분포	37
16.1 연속확률분포	38
16.1.1 Alpha Distribution(알파 분포)	39
16.1.2 Arcsine Distribution	40
16.1.3 Beta Distribution(베타 분포)	41
16.1.4 Birnbaum-Saunders Distribution	44
16.1.5 Burr Distribution of Type XII	45
16.1.6 Cauchy Distribution	46
16.1.6.1 Half-Cauchy Distribution	47
16.1.7 Chi(χ) Distribution	48
16.1.8 Chi-square(χ^2) Distribution(카이제곱 분포)	49
16.1.9 Cosine Distribution	51
16.1.9.1 Ordinary Cosine Distribution	51
16.1.9.2 Raised Cosine Distribution	52
16.1.10 Dhillon's Distribution(딜론 분포)	53
16.1.10.1 Dhillon's I Distribution	54
16.1.10.2 Dhillon's II Distribution	55
16.1.11 Exponential Distribution(지수 분포)	56
16.1.11.1 Exponential Distribution with Location Parameter	63
16.1.11.2 Exponentiated Exponential Distribution	64
16.1.11.3 Reflected Exponential Distribution	65
16.1.12 Extreme Value Distribution(극치 분포)	66
16.1.12.1 최대극치분포(The Maximum Extreme Value Distribution)와 최소극치분포(The Minimum Extreme Value Distribution)	66
16.1.12.2 Type I 최소값 극치분포(Gumbel 최소값 분포)	67
16.1.12.3 Type I 최대값 극치분포(Gumbel 최대값 분포)	70
16.1.12.4 Type II 극치분포(Frechet 분포)	73
16.1.12.5 Type II 극치분포(Frechet 분포) with Location Parameter	76
16.1.13 F Distribution	77
16.1.14 Gamma Distribution(감마 분포)	78
16.1.14.1 Gamma Distribution with 2 Parameters(2-모수 감마 분포)	78
16.1.14.2 Gamma Distribution with Location Parameter	85
16.1.14.3 Log-gamma Distribution	87
16.1.14.4 Generalized Gamma Distribution	88
16.1.15 Gompertz Distribution	89
16.1.16 Gompertz-Makeham Distribution	90
16.1.18 Hjorth Distribution(호스 분포)	91
16.1.19 Hyperbolic Secant Distribution	92
16.1.20 Laplace Distribution	95
16.1.20.1 Log-Laplace Distribution	96
16.1.21 Linear Hazard Rate Distribution(선형 증가 분포)	97
16.1.21.1 Generalized Linear Hazard Rate Distribution	98
16.1.22 Logistic Distribution	101
16.1.22.1 Log-logistic Distribution	102
16.1.22.2 Half-logistic Distribution	103
16.1.22.3 Generalized Logistic Distribution	104
16.1.23 Lomax Distribution	105
16.1.23.1 Generalized Lomax Distribution	106
16.1.24 Makeham Distribution(메이크햄 분포)	107
16.1.25 Maxwell-Boltzmann Distribution	108
16.1.26 Muth Distribution	110
16.1.27 Normal(Gaussian) Distribution	111
16.1.27.1 Log-normal Distribution(대수 정규 분포, 로그 정규 분포)	112
16.1.27.2 Log-normal Distribution with Lower Threshold	113
16.1.27.3 Log-normal Distribution with Upper Threshold	116
16.1.27.4 Inverse Normal(Gaussian) Distribution	117
16.1.27.5 Half-normal Distribution	118
16.1.27.6 Trimmed(Truncated) Normal Distribution(절사 정규 분포)	119
16.1.28 Parabolic U-shaped Distribution	120
16.1.28.1 Parabolic Inverted U-shaped Distribution	121
16.1.29 Pareto Distribution of the first kind	122
16.1.29.1 Generalized Pareto Distribution of the first kind	123
16.1.30 Power Function Distribution	124
16.1.31 Rayleigh Distribution(레이리 분포)	125
16.1.31.1 Inverse Rayleigh Distribution	127
16.1.31.2 Generalized Rayleigh Distribution	128
16.1.32 Semi-elliptical Distribution	129
16.1.33 t Distribution	130
16.1.34 Teisser Distribution	131
16.1.35 Triangular Distribution(Continuous)	132
16.1.36 Uniform Distribution(Continuous)	133
16.1.37 V-shaped Distribution	134
16.1.38 Wald Distribution	134
16.1.39 Weibull Distribution(와이블 분포)	135
16.1.39.1 Weibull Distribution with 2 Parameters(2-모수 와이블 분포)	135
16.1.39.2 Weibull Distribution with 3 Parameters(3-모수 와이블 분포)	140
16.1.39.3 Log-Weibull Distribution	142
16.1.39.4 Double Weibull Distribution	143
16.1.39.5 Inverse Weibull Distribution	144
16.1.39.6 Reflected Weibull Distribution	145
16.1.40 Wigner's Semi-circle Distribution	145
16.2 이산화률분포	146
16.2.1 Binomial Distribution	146
16.2.1.1 Positive Binomial Distribution	146
16.2.2 Extreme Value Distribution(극치 분포)	146
16.2.2.1 Weibull Distribution of Type I	146
16.2.2.2 Weibull Distribution of Type II	146
16.2.2.3 Weibull Distribution of Type III	147
16.2.3 Geometric Distribution	148
16.2.3.1 Positive Geometric Distribution	148
16.2.3.2 Zero-inflated Geometric Distribution	148
16.2.4 Hypergeometric Distribution	148
16.2.4.1 Positive Hypergeometric Distribution	148
16.2.5 Logarithmic Distribution	148
16.2.5.1 Right-truncated Logarithmic Distribution	148
16.2.6 Matching Distribution	148

16.2.7 Negative Binomial Distribution	148
16.2.8 Negative Hypergeometric Distribution	148
16.2.9 Occupancy Distribution	148
16.2.10 Poisson Distribution	148
16.2.10.1 Positive Poisson Distribution	148
16.2.11 Polya Distribution	148
16.2.12 Runs Distribution	148
16.2.13 Salvia-Bollinger's Distribution	148
16.2.13.1 Salvia-Bollinger's DHR Distribution	148
16.2.13.2 Salvia-Bollinger's Generalized DHR Distribution	148
16.2.13.3 Salvia-Bollinger's Generalized IHR Distribution	148
16.2.14 Triangular Distribution	148
16.2.14.1 Right-angled and Negatively Skew Triangular Distribution	148
16.2.14.2 Right-angled and Positively Skew Triangular Distribution	148
16.2.14.3 Right-angled and Symmetric Triangular Distribution	148
16.2.15 Uniform Distribution(Discrete)	148
16.2.16 Yule Distribution	148
16.2.17 Zeta Distribution	148
16.2.17.1 Zeta Distribution of Zipf	148
16.2.17.2 Zeta Distribution of Haight	148

VII Cox 비례위험모형 149

17 Cox 비례위험모형의 원리	150
18 Covariate가 1개인 비례위험모형	151
18.1 데이터 구조 - covariate가 1개인 경우	151
18.2 준모수적 Cox 비례위험모형	151
18.3 회귀계수에 대한 추론	152
19 Covariate가 여러 개인 비례위험모형	154
19.1 데이터 구조 - covariate가 여러 개인 경우	154
19.2 준모수적 Cox 비례위험모형	154
19.3 회귀계수와 누적위험함수의 추정	154
19.3.1 비례위험모형	155
19.3.2 부분기능도함수	155
19.3.3 정보행렬(information matrix)	156
20 동점 처리(handling ties)	157
20.1 Exact Method	157
20.2 Breslow's Approximation	158
20.3 Efron's Approximation	159

VIII Performance Evaluation Metrics 160

21 반응 변수가 binary data인 경우	161
21.1 Receive operating characteristic curve(ROC Curve) & Area under the curve(AUC)	161
21.2 Net reclassification improvement(NRI)	162
21.3 Mean Square Error(MSE) & Mean Absolute Error(MAE)	163
21.4 Cox and Snell의 결정계수(R^2) & Nagelkerke의 결정계수(R^2)	163
21.5 Brier Score	163
21.6 Hosmer-Lemeshow Test	163
22 반응 변수가 생존 시간인 경우	164
22.1 Ignoring time to event (logistic C)	164
22.2 Chambless and Diao의 C statistic	164
22.3 Gonen and Heller의 K	164
22.4 Harrell의 Concordance Index(C-index)	164
22.5 Heagerty의 intergrated AUC	165
22.6 Uno의 C-index	165
22.7 Pencina and Uno의 NRI	165
22.8 Ctd Index	165

IX Results of Survey: Read Paper and Run Software 166

23 Neural Network with Life Data	167
23.1 Peter M. Ravdin et al.(1992), A practical application of neural network analysis for predicting outcome of individual breast cancer patients	167
23.1.1 데이터셋	167
23.1.2 요약	167
23.1.3 의견	168
23.2 Knut Liestol et al.(1994), Survival analysis and neural nets	169
23.2.1 데이터셋	169
23.2.2 요약	169
23.2.3 의견	170
23.3 David Faraggi and Richard Simon(1995), A Neural Network Model for Survival Data	171
23.3.1 데이터셋	171
23.3.2 요약	171
23.3.3 의견	172
23.4 Stephen F. Brown et al.(1997), On the use of artificial neural networks for the analysis of survival data	173
23.4.1 데이터셋	173
23.4.2 요약	173
23.4.3 의견	175
23.5 Elia Biganzoli et al.(1998), Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach	176
23.5.1 데이터셋	176
23.5.2 코드	176
23.5.3 요약	176
23.5.4 의견	178
23.6 Anny Xiang et al.(2000), Comparison of the performance of neural network methods and Cox regression for censored survival data	179
23.6.1 데이터셋	179
23.6.2 요약	180
23.6.3 의견	180

23.7	Jared L. Katzman et al.(2016), Deep Survival: A Deep Cox Proportional Hazards Networks	181
23.7.1	데이터셋	181
23.7.2	코드	181
23.7.3	요약	181
23.7.4	의견	181
23.8	Travers Ching et al.(2016), Cox-nnet: an artificial neural network Cox regression for prognosis prediction	182
23.8.1	데이터셋	182
23.8.2	코드	182
23.8.3	요약	182
23.8.4	의견	182
23.9	Margaux Luck et al.(2017), Deep Learning for Patient-Specific Kidney Graft Survival Analysis	183
23.9.1	데이터셋	183
23.9.2	요약	183
23.9.3	의견	183
24	Life Data Analysis without Neural Network	184
24.1	Torsten Hothorn et al.(2006), Survival Ensembles	184
24.1.1	데이터셋	184
24.1.2	Additional Materials	184
24.1.3	요약	184
24.1.4	의견	185
24.2	Hemant Ishwaran et al.(2008), Random Survival Forests	186
24.2.1	데이터셋	186
24.2.2	Additional Materials	186
24.2.3	요약	186
24.2.4	의견	186

X Survey and Research Process

187

25	데이터셋 요약	187
25.1	Synthetic Data	188
25.1.1	생존 시간의 생성 by Bender et al.(2005)	188
25.1.2	Anny Xiang et al.(2000)의 intersection을 고려한 데이터	189
25.1.3	Linear Risk Experiment(Linear) by Jared L. Katzman et al.(2016)	190
25.1.4	Nonlinear Risk Experiment(Gaussian) by Jared L. Katzman et al.(2016)	191
25.2	Real Data	192
25.2.1	Drzewiecki and Andersen(1982)의 피부암 데이터(Melanoma)	192
25.2.1.1	변수 요약	192
25.2.2	Byar and Green(1980)의 전립선암 데이터(Byar)	194
25.2.2.1	변수 요약	194
25.2.3	Head and Neck cancer trial from Efron B.(1988)	197
25.2.4	Veteran's administration lung cancer data from Kalbfleisch and Prentice (1980)	198
25.2.4.1	변수 요약	198
25.2.4.2	분석	200
25.2.5	Worcester Heart Attack Study(WHAS)	205
25.2.5.1	변수 요약	205
25.2.6	Molecular Taxonomy of Breast Cancer International Consortium(Metabric)	207
25.2.6.1	변수 요약	207
25.2.7	Simulated Treatment Data(Treatment)	209
25.2.7.1	변수 요약	209
25.2.8	Hormone Treatment Recommendations for Breast Cancer(GBSG)	211
25.2.8.1	변수 요약	211
25.2.9	Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT)	213
25.2.9.1	변수 요약	213
25.2.10	BLCA: Bladder Urothelia Carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)	216
25.2.10.1	변수 요약	216
25.2.11	BRCA: Breast invasive carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)	218
25.2.11.1	변수 요약	218
25.2.12	HNSC: Head and Neck squamous cell carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)	220
25.2.12.1	변수 요약	220
25.2.13	KIRC: Kidney renal clear cell carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)	222
25.2.13.1	변수 요약	222
25.2.14	LGG: Brain Lower Grade Glioma from Broad Institute TCGA Genome Data Analysis Center (2014)	224
25.2.14.1	변수 요약	224
25.2.15	LIHC: Liver hepatocellular carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)	226
25.2.15.1	변수 요약	226
25.2.16	LUAD: Lung adenocarcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)	228
25.2.16.1	변수 요약	228
25.2.17	LUUSC: Lung squamous cell carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)	230
25.2.17.1	변수 요약	230
25.2.18	OV: Ovarian serous cystadenocarcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)	232
25.2.18.1	변수 요약	232
25.2.19	STAD: Stomach adenocarcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)	234
25.2.19.1	변수 요약	234
25.2.20	Acute Myeloid Leukemia from Bullinger et al. (2004)	236
25.2.20.1	변수 요약	236
25.2.21	Node-positive breast Cancer from German Breast Cancer Study Group (GBCS2)	237
25.2.21.1	변수 요약	237
25.2.22	Primary biliary cirrhosis data from Fleming and Harrington(1991) (pbc)	240
25.2.22.1	변수 요약	240
25.2.23	데이터셋 활용 시 참고사항	245
25.2.23.1	H5 확장자를 R에서 불러오기	245

List of Tables

1	Regression과 생존분석의 차이점	12
2	Ping Wang et al.에 의해 소개된 생존분석 관련 소프트웨어 패키지	15
3	심근경색 환자 자료	35
4	Alpha 분포함수에 기반한 척도 함수	39
5	Arcsine 분포함수에 기반한 척도 함수	40
6	Beta 분포함수에 기반한 척도 함수	41
7	Birnbaum-Saunders 분포함수에 기반한 척도 함수	44
8	Burr 분포함수에 기반한 척도 함수	45
9	Cauchy 분포함수에 기반한 척도 함수	46
10	Half-Cauchy 분포함수에 기반한 척도 함수	47
11	χ 분포함수에 기반한 척도 함수	48
12	카이제곱분포함수에 기반한 척도 함수	49
13	Ordinary Cosine 분포함수에 기반한 척도 함수	51
14	Raised Cosine 분포함수에 기반한 척도 함수	52
15	Dhillon's I 분포함수에 기반한 척도 함수	54
16	Dhillon's II 분포함수에 기반한 척도 함수	55
17	지수분포함수에 기반한 척도 함수	56
18	Exponential 분포함수에 기반한 척도 함수	63
19	Exponentiated exponential 분포함수에 기반한 척도 함수	64
20	Reflected Exponential 분포함수에 기반한 척도 함수	65
21	Type I 극치분포(Gumbel 최소값 분포)함수에 기반한 척도 함수	67
22	Type I 극치분포(Gumbel 최대값 분포)함수에 기반한 척도 함수	70
23	Type II 극치분포(Frechet 분포)함수에 기반한 척도 함수	73
24	Type II 극치분포(Frechet 분포) with Location Parameter 함수에 기반한 척도 함수	76
25	F 분포함수에 기반한 척도 함수	77
26	2모수 감마 분포함수에 기반한 척도 함수	78
27	3모수 감마 분포함수에 기반한 척도 함수	85
28	Log-gamma 분포함수에 기반한 척도 함수	87
29	Generalized Gamma 분포함수에 기반한 척도 함수	88
30	Generalized Exponential 분포함수에 기반한 척도 함수	89
31	Gompertz 분포함수에 기반한 척도 함수	90
32	Gompertz-Makeham 분포함수에 기반한 척도 함수	91
33	호스 분포함수에 기반한 척도 함수	92
34	Hyperbolic Secant 분포함수에 기반한 척도 함수	95
35	Laplace 분포함수에 기반한 척도 함수	96
36	Log-Laplace 분포함수에 기반한 척도 함수	97
37	선형증가 분포함수에 기반한 척도 함수	98
38	Generalized linear hazard rate 분포함수에 기반한 척도 함수	101
39	Logistic 분포함수에 기반한 척도 함수	102
40	Log-logistic 분포함수에 기반한 척도 함수	103
41	Half-logistic 분포함수에 기반한 척도 함수	104
42	Generalized logistic 분포함수에 기반한 척도 함수	105
43	Lomax 분포함수에 기반한 척도 함수	106
44	Generalized Lomax 분포함수에 기반한 척도 함수	107
45	메이크램 분포함수에 기반한 척도 함수	108
46	Maxwell-Boltzmann 분포함수에 기반한 척도 함수	110
47	Much 분포함수에 기반한 척도 함수	111
48	Normal 분포함수에 기반한 척도 함수	112
49	로그정규분포함수에 기반한 척도 함수	113
50	Log-normal with lower threshold 분포함수에 기반한 척도 함수	116
51	Log-normal with upper threshold 분포함수에 기반한 척도 함수	117
52	Inverse-normal 분포함수에 기반한 척도 함수	118
53	Half-normal 분포함수에 기반한 척도 함수	119
54	절사정규분포 함수에 기반한 척도 함수	120
55	Parabolic U-shaped 분포함수에 기반한 척도 함수	121
56	Parabolic Inverted U-shaped 분포함수에 기반한 척도 함수	122
57	Pareto 분포함수에 기반한 척도 함수	123
58	Power Function 분포함수에 기반한 척도 함수	124
59	레이리 분포함수에 기반한 척도 함수	125
60	Inverse Rayleigh 분포함수에 기반한 척도 함수	127
61	Generalized Rayleigh 분포함수에 기반한 척도 함수	128
62	Semi-elliptical 분포함수에 기반한 척도 함수	129
63	t 분포함수에 기반한 척도 함수	130
64	Teisser 분포함수에 기반한 척도 함수	131
65	Triangular 분포함수에 기반한 척도 함수	132
66	Uniform 분포함수에 기반한 척도 함수	133
67	V-shaped 분포함수에 기반한 척도 함수	134
68	2모수 와이블 분포함수에 기반한 척도 함수	135
69	3모수 와이블 분포함수에 기반한 척도 함수	140
70	Log-Weibull 분포함수에 기반한 척도 함수	142
71	Double Weibull 분포함수에 기반한 척도 함수	143
72	Inverse Weibull 분포함수에 기반한 척도 함수	144
73	Reflected Weibull 분포함수에 기반한 척도 함수	145
74	Type III 극치분포함수에 기반한 척도 함수	147
75	재분류표	162
76	Peter M. Ravdin et al.에서 사용한 데이터셋의 변수	167
77	Drzewiecki and Andersen(1982)의 피부암 데이터 요약	193
78	Byar and Green(1980)의 전립선암 데이터 요약	195
79	Byar and Green(1980)의 전립선암 데이터 요약(cont'd)	196
80	Raw data for Arm A (Table 1) from Efron (1988)	197
81	Raw data for Arm B (Table 2) from Efron (1988)	197
82	Veteran's administration lung cancer data from Kalbfleisch and Prentice (1980) 데이터 요약	199
83	WHAS 데이터 요약(cont'd)	206
84	Metabric 데이터 요약	208
85	Treatment 데이터 요약	210
86	Metabric 데이터 요약	212
87	Support 데이터 요약	214
88	Support 데이터 요약 (Cont'd)	215
89	BLCA 데이터의 clinical part 요약	217
90	BRCA 데이터의 clinical part 요약	219
91	HNSC 데이터의 clinical part 요약	221

92	KIRC 데이터의 clinical part 요약	223
93	LGG 데이터의 clinical part 요약	225
94	LIHC의 데이터의 clinical part 요약	227
95	Support 데이터 요약 (Cont'd)	229
96	LUSC 데이터의 clinical part 요약	231
97	OV 데이터의 clinical part 요약	233
98	STAD 데이터의 clinical part 요약	235
99	Byar and Green(1980)의 전립선암 데이터 요약	238
100	Byar and Green(1980)의 전립선암 데이터 요약(cont'd)	239
101	PBC 데이터 요약	241
102	PBC 데이터 요약(Cont'd)	242
103	PBC 데이터 요약(Cont'd)	243
104	PBC 데이터 요약(Cont'd)	244

List of Figures

1	Ping Wang et al.에 의해 소개된 생존 분석이 현실에서 성공적으로 활용되고 있는 사례	8
2	Clinical Research의 분류	9
3	Ping Wang et al.에 의해 소개된 생존 분석 방법 분류	14
4	$F(t)$ 와 $R(t)$ 의 관계	17
5	이론적 생존함수(왼쪽)과 추정된 생존함수(오른쪽)	17
6	생존 함수를 간 관계도	20
7	일반적인 위험률 곡선의 유형	22
8	육조 곡선(bath-tub curve)	22
9	비모수적 군들의 상호 관계	24
10	수리가 가능한 시스템	28
11	신뢰성 보증 사이클	30
12	"고장"이 일어나는 유형 및 매커니즘	30
13	우종도절단(right-censored)	32
14	(a): 완전 자료(complete data), (b): Type 1 우종도절단, (c): Type 2 우종도절단, (d): 임의 우종도절단	32
15	랜덤 중단 자료	34
16	베타분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	41
17	카이제곱분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	49
18	지수분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	56
19	Type I 극치분포(Gumbel 최소값 분포)에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	67
20	Type I 극치분포(Gumbel 최대값 분포)에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	70
21	Type II 극치분포(Frechet 분포)에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	73
22	감마분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	78
23	3모수 감마분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	85
24	호스분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	92
25	선행증가분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	98
26	메이크램분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	108
27	로그정규분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	113
28	레일리분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	125
29	2모수 와이블분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	135
30	3모수 와이블분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수	140
31	ROC curve	161
32	Peter M. Ravdin et al. 페이퍼에서의 데이터 전처리	167
33	Peter M. Ravdin et al. 페이퍼에서의 신경망 모형	168
34	Liestbl et al.(1994)에서 소개하는 One-layer network와 Two-layer network	169
35	Single hidden layer neural network to single output	171
36	Coohong et al.(1993)의 방법으로 추정된 observed failure의 평균(O 표시)와 censoring time의 평균(X 표시). 실선은 true mean이고 점선은 true median이다.	174
37	Kaplan-Meier 방법으로 추정된 observed failure의 평균(O 표시)와 censoring time의 평균(X 표시). 실선은 true mean이고 점선은 true median이다.	174
38	Feed forward neural network model for partial logistic regression(PLANN)	178

List of Notes

Note 1:	진공관 사례	10
Note 2:	신뢰성 분석의 주요 역사	11
Note 3:	MIL-HDBK-339A에 따른 신뢰성 예측 모형의 기원	11
Note 4:	생존 분석의 주요 역사[47]	13
Note 5:	위험률 함수 $h(t)$ 의 유도	18
Note 6:	누적함수를 이용하여 생존함수 구하기	19
Note 7:	평균 위험률(평균 고장률)	23
Note 8:	$T \geq 0$ 인 경우 평균수명의 유도	25
Note 9:	평균잔여수명함수의 유도	27
Note 10:	평균잔여수명함수와 위험률 함수, 생존 함수간의 1:1 관계	27
Note 11:	라오-블랙웰(Rao-Blackwell) 정리	37
Note 12:	지수분포에 기반한 k차 적률	57
Note 13:	지수분포를 따르는 완전자료에 대한 모수 추정	60
Note 14:	지수분포를 따르는 Type II 우종도절단자료(정수중단자료)에 대한 모수 추정	61
Note 15:	지수분포를 따르는 Type I 우종도절단자료(정시중단자료)에 대한 모수 추정	62
Note 16:	최대 극치 분포(The Maximum Extreme Value Distribution)	66
Note 17:	최소 극치 분포(The Minimum Extreme Value Distribution)	66
Note 18:	감마 함수	80
Note 19:	감마분포에 기반한 생존함수(신뢰도함수)	80
Note 20:	감마분포에 기반한 위험률함수(고장률함수)	80
Note 21:	감마분포에 기반한 k차 적률	80
Note 22:	감마분포를 따르는 완전자료에 대한 모수 추정	82
Note 23:	Digamma 함수	83
Note 24:	감마분포를 따르는 Type II 우종도절단자료(정수중단자료)에 대한 모수 추정	84
Note 25:	2모수 와이블분포에 기반한 평균잔여수명함수	137

Note 26: 2모수 와이블분포에 기반한 k 차 적률	137
Note 27: 100 p 백분위수	137
Note 28: 2모수 와이블분포를 따르는 완전자료에 대한 모수 추정	138
Note 29: 2모수 와이블분포를 따르는 Type II 우종도절단자료(정수중단자료)에 대한 모수 추정	139
Note 30: Cox 비례위험모형 추정량의 통계적 성질	156
Note 31: AUC를 이용하여 두 모형을 비교할 경우, 통계적 유의성 검정	161
Note 32: Model _k ($k = 1, 2$)에서 구해진 예측확률 p_k 를 cut point를 기준으로 작성한 분류표와 이에 따른 NRI 계산	162
Note 33: C-index를 이용하여 두 모형을 비교할 경우, 통계적 유의성 검정	165
Note 34: C-index의 95% 신뢰구간 구하기	165

List of Datasets

Dataset 1: Nichols Institute Oncology Research Network의 breast cancer 환자 데이터	167
Dataset 2: Drzewiecki and Andersen(1982)의 피부암 데이터	169
Dataset 3: Byar and Green(1980)의 전립선암 데이터	171
Dataset 4: Anny Xiang et al.(2000)의 Synthetic Data	179
Dataset 5: T. Hothorn et al.(2006)의 Acute Myeloid Leukemia from Bullinger et al. (2004)	184
Dataset 6: Node-positive Breast Cancer	184
Dataset 7: Raw data for Arm A (Table 1) from Efron (1988)	197
Dataset 8: Raw data for Arm A (Table 1) from Efron (1988)	197

List of Codes

Code 1: 베타분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	42
Code 2: 카이제곱분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	50
Code 3: 지수분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	57
Code 4: Type I 극치분포(Gumbel 최소값 분포)에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	68
Code 5: Type I 극치분포(Gumbel 최대값 분포)에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	71
Code 6: Type II 극치분포(Frechet 최대값 분포)에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	74
Code 7: 감마분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	79
Code 8: 3모수 감마분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	86
Code 9: 호스분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	93
Code 10: 선형증가분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	99
Code 11: 메이크램분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	109
Code 12: 로그정규분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	114
Code 13: 절사정규분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	120
Code 14: 레일리분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	126
Code 15: 2모수 와이블분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	136
Code 16: 3모수 와이블분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)	141
Code 17: Drzewiecki and Andersen(1982)의 피부암 데이터 불러오기(R 코드)	192
Code 18: Byar and Green(1980)의 전립선암 데이터 불러오기(R 코드)	194
Code 19: Raw data for Arm A (Table 1) from Efron (1988) 불러오기(R 코드)	197
Code 20: Raw data for Arm B (Table 2) from Efron (1988) 불러오기(R 코드)	197
Code 21: Veteran's administration lung cancer data from Kalbfleisch and Prentice (1980) 데이터 불러오기(R 코드)	198
Code 22: Byar and Green(1980)의 전립선암 데이터 불러오기(R 코드)	237
Code 23: Byar and Green(1980)의 전립선암 데이터 불러오기(R 코드)	240

List of Examples

Example 1: Kaplan-Meier 방법에 의한 생존함수 추정	34
Example 2: Life-table Method: 심근경색 환자의 임상시험결과	35

Part I

서론

1 서론

생존분석(survival analysis)에서는, 어떤 연구에 들어온 시간부터 어떤 사건이 발생할 때까지의 시간 구간(time interval) 데이터에 관심이 있다.

관심 있는 **반응변수는, 사건이 발생할 때까지 걸리는 시간이다.**

생존분석의 주 관심사는 다음 두 가지로 요약할 수 있다.

- 생존함수(survival function)의 추정
- 생존함수(survival function) 또는 위험함수(hazard function)에 영향을 주는 공변량(covariate) 또는 예측변수를 찾아내어 그 연관 정도를 추정
*공변량(covariate) 또는 예측변수 :
 - 일반적으로 연구자가 통제할 수 있는 처리(treatment)
 - 또는 실제 관심요인은 아니지만, 반응변수에 유의한 영향을 줄 수 있는 요인

생존분석이 다른 통계분석방법들과 구별되는 가장 큰 특징은, 종도절단자료(censored data)를 포함한다는 것이다.
생존분석 자료는 유한한 시간 내에서 일어나는 사건만 측정할 수 있다.

Ping Wang et al.[51]은 생존 분석이 어디에 활용될 수 있는지를 Figure 1과 같이 정리하고 있다. 참고하도록 하자.

Application	Event of interest	Estimation	Features
Healthcare [Miller Jr 2011] [Reddy and Li 2015]	Rehospitalization Disease recurrence Cancer survival	Likelihood of hospitalization within t days of discharge.	Demographics: age, gender, race. Measurements: height, weight, disease history, disease type, treatment, comorbidities, laboratory, procedures, medications.
Reliability [Lyu 1996] [Modarres et al. 2009]	Device failure	Likelihood of a device being failed within t days.	Product: model, years after production, product performance history. Manufactory: location, no. of products, average failure rate of all the products, annual sale of the product, total sale of the product. User: user reviews of the product.
Crowdfunding [Rakesh et al. 2016] [Li et al. 2016a]	Project success	Likelihood of a project being successful within t days.	Projects: duration, goal amount, category. Creators: past success, location, no. of projects. Twitter: no. of promotions, backings, communities. Temporal: no. of backers, funding, no. of retweets.
Bioinformatics [Li et al. 2016d] [Beer et al. 2002]	Cancer survival	Likelihood of cancer within time t .	Clinical: demographics, labs, procedures, medications. Genomics: gene expression measurements.
Student Retention [Murtaugh et al. 1999] [Ameri et al. 2016]	Student dropout	Likelihood of a student being dropout within t days.	Demographics: age, gender, race. Financial: cash amount, income, scholarships. Pre-enrollment: high-school GPA, ACT scores, graduation age. Enrollment: transfer credits, college, major. Semester performance: semester GPA, % passed credits, % dropped credits.
Customer Lifetime Value [Zeithaml et al. 2001] [Berger and Nasr 1998]	Purchase behavior	Likelihood of a customer purchasing from a given service supplier within t days.	Customer: age, gender, occupation, income, education, interests, purchase history. Store/Online store: location, customer review, customer service, price, quality, shipping fees and time, discount.
Click Through Rate [Yin et al. 2013] [Barbieri et al. 2016]	User clicking	Likelihood of a user clicking the advertisement within time t .	User: gender, age, occupation, interests, users click history. Advertisement (ad): time of the ad, location of the ad on the website, topics of the ad, ad format, total click times of the ad. Website: no. of users of the website, page view each day of the website, no. of websites linking to the website.
Unemployment Duration in Economics [Kiefer 1988]	Getting a job	Likelihood of a person finding a new job within t days.	People: age, gender, major, education, occupation, work experience, city, expected salary. Economics: job openings, unemployment rates every year.

Figure 1: Ping Wang et al.에 의해 소개된 생존 분석이 현실에서 성공적으로 활용되고 있는 사례

2 임상적 연구 결과 데이터

앞선 Figure 1로부터, 생존 분석을 다양한 분야에서 활용할 수 있음을 소개했지만, 앞으로 이 노트에서는 의학에서의 임상적 연구 결과 데이터를 다루는 경우로 한정한다.

본래 이 연구에서는 의학에서의 임상적 연구 결과 데이터만 다루는 것으로 한정하였으나, 연구 방향을 parametric model로 변경함에 따라 이러한 제한을 하지 않기로 하였다.

임상 연구는 크게 laboratory research와 clinical research로 분류되고, 송경일과 최종수(2007)[2]에 따르면, clinical research는 다음과 같이 분류된다.

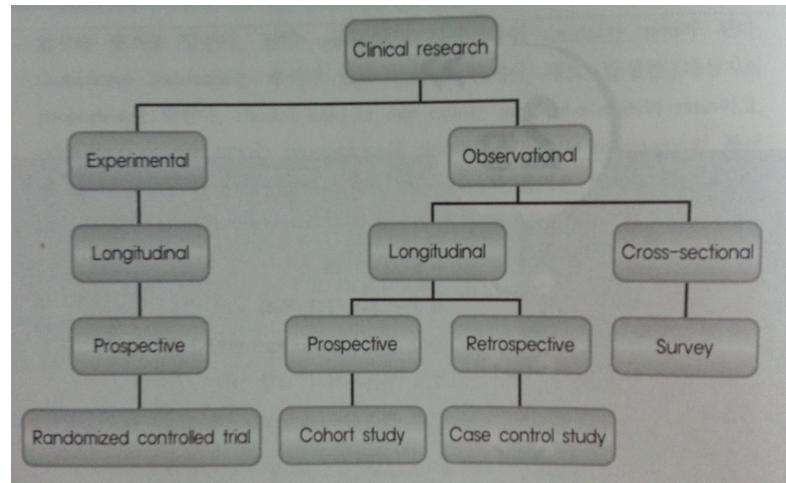


Figure 2: Clinical Research의 분류

- **Experimental Study**에서는 대상자에서 약물 투여와 같은 treatment 혹은 intervention 후 그 효과에 대해 연구한다. 혼히 longitudinal study이며, intervention한 후 일정한 시간이 지난 후 반응을 연구한다.
- **Observation Study**는 연구자가 대상자의 존재하는, 혹은 기준에 존재하고 있는 상황을 관찰하고 기술하고 분석한다.
 - **Cohort Study**는 특정 연구대상을 선정하고 그 대상으로부터 특정 질병의 발생에 관여하리라 의심되는 어떤 특성이나 위험인자에 폭로된 정보를 수집한 후, 특정 질병의 발생을 시간 경과에 따라 prospective(전향적으로)하게 추적조사 혹은 관찰함으로써 특정 요인에 폭로되지 않은 집단에 비해 폭로된 집단에서의 질병의 발생률을 비교하는 방법이다.
 - **Case Control Study(환자대조군 연구)**는 후향적 연구이며, 관찰된 결과를 관심있는 사건이 발생한 case와 발생하지 않은 control로 나누어서, 사건의 결과를 발생하게 한 가능한 위험인자를 찾아내는 방법이다.
 - **Cross Sectional Study(단면조사 연구)**는 현재 연구시점의 한 시점에서 일정한 인구 집단을 대상으로 모집단을 대표할 수 있는 표본을 추출하여 조사한 후, 속성에 따라 이를 표본집단을 의심되는 원인에 폭로된 집단과 의심되는 원인에 폭로되지 않은 집단으로 나누어 차이점을 분석하는 것이다.

3 신뢰성 공학

신뢰성(reliability)은 다음과 같이 정의된다.

- 성능, 시스템의 성공적인 임무 수행, 파손이나 고장의 부재를 의미한다.
- 1. 구성품, 장치, 설비, 아이템 또는 시스템이
- 2. 주어진 사용 조건 하에서
- 3. 규정된 기간 동안
- 4. 의도한 기능을 수행할 확률

제품의 신뢰성을 생각할 때, 제조사 입장과 사용자 입장은 분리하여 다음과 같이 나누어 정의하기도 한다.

- **고유 신뢰성(inherent reliability)**: 제조사 측에서 보증하는 제품 본래의 신뢰성

제품 기획단계에서 목표품질을 설정하고, 규칙(sepcification)을 결정하여, 부품 재료의 선택, 구입, 설계, 시작, 시험, 검사, 제조 등을 거쳐 제품 생산 전체 공정에 관계한다.

- **사용 신뢰성(use reliability)**: 포장, 수송, 보관, 설치환경, 취급 조작, 보전 기술, 보전 방식 등 사용 과정에서 나타나는 신뢰성으로, 인간의 요소가 신뢰성에 밀접하게 관계한다.

신뢰성(reliability)이라는 용어가 처음 쓰인 시점은 1816년의 Coleridge에 의한 것으로 알려져 있다.^[57] 제2차 세계대전 이전까지만 해도, 모든 과학 분야에서 이루어지는 실험은 동일 조건에서 동일한 결과가 반복적으로 얻어져야 했으므로, "신뢰성"은 "반복성(repeatability)"의 의미로 쓰였다. 그러던 것이, 1924년에 미국 벨 연구소 소속의 Walter A. Shewhart 박사가, 프로세스 관리에 통계를 사용하면서, 제품의 품질 관리에 확기적인 전기를 마련하였고, 1939년에는 스웨덴 물리학자 와이블(waloddi Weibull)이 재료의 파괴 강도를 분석할 때, 금속 및 복합 재료의 강도나 피로도, 전자 및 기계 부품의 통계적 모형에 "와이블 분포(Weibull distribution)"를 고안해 활용하였다.^[62] 1940년대 들어, 미국이 제2차 세계대전에 휘말리게 되고, 이를 계기로 "신뢰성"은 미국방성을 통해 체계화하면서 크게 발전하였다. 초창기 때 군수품에 대한 요구는 "임무 수행 시 지정된 기간 동안 잘 작동하기만을 기대하는 수준"이었으나, 제2차 세계대전 당시에 "피로 누적 손상(Cumulative Damage in Fatigue)처럼 전자 부품들이 안고 있는 고질적인 문제들이 세상에 드러나기 시작했다.

다음 Note 1은 "신뢰성"에 대한 개념이 부상한 직접적인 계기가 된 "진공관(Vacuum Tube)"의 사례를 옮긴 것이다.

Note 1. 진공관 사례

'신뢰성'의 시작은, 한 개의 진공관으로부터 비롯되었다. 당시는 제차 세계대전이 한창인 시기였다. 미국은 극동 전략용으로서 남방 기지에 많은 군용 항공기를 배치했으나, 극동에 선박으로 수송된 항공기용 전자 기기의 60%가 목적지에 도착했을 때 고장을 일으켰으며, 창고에 보관되 있던 기기 및 예비 부품의 50%가 사용되기도 전에 사용 불가 상태에 놓이곤 하였다. 또, 공군 폭격기용 전자 기기가 20시간 동안 무고장이 거의 없었을 정도로 고장 빈도가 높았다. 당시 고장 보고서와 부품 교환 기록에 따르면 무선 통신 기기의 고장 발생은 시간당 14%에 이른 것으로 알려져 있다.

이에 그 이유를 조사한 결과, 진공관의 고장이 원인인 것으로 판명되었다. 진공관은 항공기의 전자 분야에서 매우 중요한 기능을 담당하고 있었다. 이 결과는 미국 정부의 전쟁 수행 능력에 중대한 문제가 되었음은 말할 나위도 없었고, 결국 정부와 민간이 일체가 되어 긴급하게 대책을 강구하기에 이른다. 진공관이 제조 시점부터 불량이었다는 보고를 근거로 생산 중 검사를 충분히 강화시킨 뒤, 도면대로 완성된 진공관을 극동 쪽으로 다시 보냈다. 그러나 고장은 여전히 계속되었다. 수 차례 반복해도 결과는 마찬가지였다. 공장에서 양품이었던 진공관이 출하 후 사용 중, 또는 사용 전에 고장 난다는 사실은, 종래의 제조 기술이나 제조 검사의 한계를 넘어서는 또 다른 **무언가**의 고려가 더 필요하다는 결론에 이르렀다.

무언가란, 바로 "고장을 일으키지 않도록 하는 특성"에 있었으며, 이 특성이 "신뢰성"으로 정의되었다.

1950년대부터는 신뢰성 연구 활동이 본격화되기 시작했다. 1950년 12월 미 국방성 연구 개발국(RDB, Research and Development Board)은, 가능한 한 보전 횟수가 적은 신뢰성 높은 기기를 얻기 위한 방법을 제안할 목적으로, 전자 기기의 신뢰성 연구를 위한 비공식 그룹(Ad Hoc Committee)을 만들었는데, 그들은 1952년 5월에 다음과 같은 6개 항의 권고안을 발표하였다.

- 기기 작동과 부품 고장에 관한 보다 정확한 데이터를 얻기 위해 조직적인 데이터 수집 계획을 수립하고, 그 계획에 따라 규칙적으로 작업하며, 보전을 실시하는 사람들에 의한 적절한 보고 시스템을 확립해야 한다.
- 높은 신뢰성을 갖는 부품을 개발해야 한다.
- 기기와 부품에 대하여 양적인 신뢰도 요구를 설정해야 한다.
- 신규 기기 개발자들에 대한 신뢰성 교육 계획을 수립해야 한다.
- 신규로 설계, 제조되는 기기는 양산에 들어가기 전에 실험실 또는 현장 시험에 의하여 평가되어야 한다.
- 부품을 공급하고 있는 하청 업체들을 잘 관리하여야 한다.

1952년 8월에 RBD는 Ad Hoc Committee이 제출한 권고안을 받아들여, 전자 기기의 신뢰성에 관한 자문 그룹인 AGREE(Advisory Group on Reliability of Electronics Equipment)을 설치하였다. 이 그룹은 신뢰도 측정법, 규격서 작성법, 수송이나 보관의 문제 등 9개 그룹으로 나뉘어 5년 동안 활동하였고, 1957년 6월에 발표한 보고서 내용에 "신뢰성 공학"의 제 개념들이 정립되어 "MIL-STD"²의 근간을 이루며 오늘날까지 신뢰성 규격의 기초가 되었다.

²United States Military Standard, 위키피디아 https://en.wikipedia.org/wiki/United_States_Military_Standard 참고

AGREE에서는 다음과 같은 세 개의 작업 지침을 권고한 바 있다.

- 부품의 신뢰성을 향상시킬 것
 - 공급자를 위해 품질과 신뢰성 요구 사항들을 정립할 것
 - 시장 데이터를 수집하고 고장의 근본 원인들을 규명할 것
-

Note 2. 신뢰성 분석의 주요 역사

- 1939 Wallodi Weibull이 ”고장률”을 모델링하기 위한 ”와이블 분포”를 제의함.
 - 1941~1945 제2차 세계대전 당시, ”신뢰성”의 중요성을 인식: 항공기용 통신 장비 등에서 60~75%의 진공관 고장 발생
 - 1943 진공관 개발위원회(VTDC) 설치. 이후 1946년에 PET, 1953년에 AGET로 이름이 변경됨.
ARINC에서 3군의 약 3만개의 진공관을 검사하고, 고장난 진공관을 수집하여 분석한 뒤 고장 형태와 그 원인을 밝혀냄.
 - 1948 진공관 문제를 해결하기 위해 IEEE에서 the Reliability Society가 설립됨.
IEC에 의해 ”부품에 대한 기본적 기후와 기계적 내력의 시험 절차”라는 표준 보고서가 완성됨.
 - 1950 W. Edwards Deming 교수의 ”통계적 품질 관리 기법(SQC)”은, 일본 신뢰성 발전의 획기적인 전기를 마련함.
미 국방성 연구 개발국(RDB)에서 Ad Hoc Committee이 조직됨.
 - 1952 미 국방성(DOD)의 특별 위원회 AGREE(전자 장비의 신뢰성에 대한 자문단)이 설립됨.
 - 1955 B. Epstein이 지수분포의 수명분포로서의 유용성을 증명함. 신뢰도의 척도로 평균 고장 시간(MTBF; mean time between failure)이나 고장률(failure rate)을 사용함
 - 1958~1959 AGREE가 ”군사전자장비의 신뢰성(MIL-STD-441)”, ”신뢰성과 수명요구조건, 전자장비(MIL-R-266670)”를 발행함
 - 1963 AGREE가 무기 체계의 신뢰성 예측을 보고하기 위한 ”신뢰성 모델링과 예측(MIL-STD-756)”을 발행함.
 - 1965 AGREE가 ”시스템, 장비 개발, 생산을 위한 신뢰성 프로그램(MIL-STD-785)”를 발행함.
 - 1982 AGREE가 ”전자 장비의 신뢰성 예측(MIL-HDBK-217)”을 발행함.
 - 1991 ”소프트웨어”가 신뢰성에 중요한 요소로 떠오르기 시작함. 카네기 멜론 대학(CMU)의 소프트웨어공학 연구소(SEI)가 미 국방성(DOD)의 지원을 받아, CMMI를 발표함. (연구를 시작한 것은 1986년)
 - 1994 Ford사가 생산품과 생산 과정 엔지니어링에 대해 ”신뢰성”과 ”Robustness 설계 방식”을 통합하는 심포지엄에 신뢰성 전문가들을 초청함.
 - 1995 미 국방성(DOD)이 MIL-HDBK-217 등을民間에 공개
자동차 기술자 협회(SAE)와 자동차 산업 활동 그룹(AIAG)이 ”잠재적 고장모드와 영향 분석(FMEA)”에 대한 규격 발표
 - 20세기~ 컴퓨터의 향상된 엔지니어링 도구로, 하드웨어 테스트 필요성이 줄고, 가상현실 발전으로 견본 제작이 빨라짐.
RAMS(Reliability and Maintainability Symposium)이 발전함.
-

Note 2는 신뢰성의 역사가 주로 미국에서 일어난 시각에서 알아본 것인데, 미국이 주류를 형성한 이유도 있지만, 신뢰성과 관련된 각종 국제 표준들도 미국이 주도하는 흐름 속에서 파생되고 발전해 왔음을 부인하기 어렵기 때문이다.

다만, 다른 시각도 있는데, MIL-HDBK-338A[45]에 따르면, Predictive Reliability Model의 기원이 미국이 아닌, 제2차 세계대전 당시 독일에서 이루어진 로켓 개발과 관련이 있음을 언급하고 있다. 관련 내용을 짚기면 다음과 같다.

Note 3. MIL-HDBK-339A에 따른 신뢰성 예측 모형의 기원

’신뢰성’을 처음으로 깊이 있게 탐구한 사람들 중, 독일의 로켓 개발 엔지니어인 Wernher Von Braun 박사가 있다. 그는 제2차 세계대전 때 ”Buzz-Bomb”으로 알려진 ”V-1 로켓”을 최초로 개발했다. 당시 ”V-1 로켓”은 신뢰성 문제들에 시달리고 있었고, Braun 박사와 그의 팀은 이 문제를 해결하기 위해 고군분투하고 있었다. 그들은 로켓의 문제를 진단하고 그를 향상시키기 위해 단순한 기계적 신뢰성 개념을 적용했는데, 이는 사슬과 같이 최약 연결을 찾아 개선하면, 고장이 나지 않을 것이라고 기대하고 있었다. 그러나 최약체 부품의 신뢰도를 높여 조립했음에도 ”V-1 로켓”的 신뢰도에는 별다른 영향이 없었다.

그 때 다른 프로젝트에서 Braun 박사와 함께 연구 중이던 독일의 수학자 Eric Pieruschka, Braun 박사의 신뢰성 문제 해결에 도움을 주게 된다. 그는 Braun 박사가 갖고 있던 ”신뢰도 낮은 부품의 존재가 전체 시스템인 로켓에 악영향을 줄 것”이라는 개념이 잘못된 것임을 지적했다. 즉, Pieruschka는 ”로켓의 신뢰도는 각 부품들의 신뢰도의 곱과 같음”을 보인 것이다. 이와 같은 접근은, 최초로 문서화된 현대적 수준의 ”신뢰적 예측 모형(predictive reliability model)”이며, 후에 신뢰성 분석 선구자로 알려진 Lusser’s Law를 낳는 데 기초가 되었다.

$$R(t) = \prod_{i=1}^N R_i(t)$$

Part II

생존 자료의 분석

생존 분석에 사용되는 생존 자료는 다음과 같은 특징이 있다.

- 생존과 관계되는 결과변수는 '사망' 혹은 생존'과 같은 생존의 여부(binary outcome)와, 생존 시간(survival time)이라는 두 가지의 형태로 이루어져 있다.
- 생존 시간은 음수(negative)로 측정될 수 없고, 향상 양수(positive) 값으로 측정된다.
- 사건이 발생할 때까지 걸리는 시간인 'time to event occurrence'에 대한 분포는 보통 정규분포를 하지 않는다. 정규분포를 하지 않는 이유는, 결과가 발생하기까지의 시간 간격이 대상자마다 다르고, 연구가 종료될 때 모든 대상자에서 사건이 발생하지 않기(다양한 이유의 censoring) 때문이다.
- 중도 절단된 관찰(censored observation)이 발생하기 마련이다.

위와 같은 생존 자료의 특성상, 생존 자료는 일반적인 통계 분석 방법으로는 분석하기 어렵다. 그럼 어떤 방법으로 분석하는 것이 좋을까?

- multiple linear regression으로 분석한다면 어떨까?
 - 그러나, 생존 자료의 종속 변수는 대부분 정규분포를 하지 않기 때문에³ multiple linear regression에서 독립변수의 분포는 정규분포를 한다는 가정에 위배된다.
 - 또한, 중도 절단이 발생하므로, multiple linear regression은 중도 절단에 대한 처리 방법이 없으며, 이분형의 결과 변수를 처리할 수 없으므로 생존 자료는 분석하기 어렵다.
 - 만약 중도 절단된 자료가 전혀 없고, 모든 대상자에서 사건이 발생했다면, 그리고 생존시간도 어느 정도 정규분포를 한다면, 예측 인자와 time to event의 사이의 분석에 multiple linear regression을 이용할 수도 있다.
- binary logistic regression으로 분석한다면 어떨까?
 - logistic regression analysis도 중도 절단된 자료를 처리할 수 없으며, time to event에 대한 자료를 분석할 수 없다.
 - logistic regression은 '사건의 발생 여부'에 초점을 둔 방식이고, 중도 절단이 없다면 logistic regression을 할 수 있다. 하지만 중도 절단된 자료도 분명히 생존시간에 기여를 했는데, 이를 분석에서 제외한다면, 중요한 정보를 소실하게 되는 것이다.

방법	Linear Regression	Logistic Regression	Survival Analysis
예측 변수	Cartegorical or Continuous	Cartegorical or Continuous	Time and {Cartegorical or Continuous}
결과 변수	Normal Distribution	Binary	Binary
중도 절단 허용	No	No	Yes
수학적 모형	$Y = B_1 X + B_0$	$\log\left(\frac{p}{1-p}\right) = B_1 X + B_0$	$H(t) = H_0(t) \exp^{B_1 X + B_0}$
결과의 산출	Linear Change	Odds Ratios	Hazards Rates

Table 1: Regression과 생존분석의 차이점

생존 분석을 하는 목적은, 주어진 생존 자료로 다음과 같은 결과를 보고자 함이다.

- descriptive survival analysis: 생존함수⁴를 추정하여 모집단에서 생존 시간의 분포를 알아보기 위해서 (예: 1 year survival rate, median survival time 등)
- univariate survival analysis: 추정된 생존함수가 집단에 따라서 차이가 있는지를 비교하기 위해서 (예: 치료집단과 비치료집단 사이에 생존율의 차이가 있는가?)
- semi-parametric survival analysis⁵: Time to event와 이에 영향을 미치는 독립변수들과의 관계를 찾기 위해서 (예: 생존율이 차이가 있다면, 생존율에 영향을 미치는 중요한 인자는 무엇인가?)

³생존시간은 보통 exponential, Weibull, log normal distribution을 따른다

⁴"생존 분석"에서는 $S(t)$ "신뢰성 공학"에서는 $R(t)$ 로 주로 표현한다.

⁵cox-proportional hazards model 포함

생존 분석의 기본 가정은 다음과 같다. 데이터가 다음의 기본적인 가정을 만족하지 못하면, 생존분석을 수행할 수 없다.

- 생존분석에서 대상자에 대해 관찰을 시작하고 종료하는 것은, 생존 여부의 발생에는 영향을 주지 않는다.
즉, 생존 여부의 발생은 완전히 무작위로 발생해야 하며, 생존의 간으성은 연구기간 동안 내내 항상 일정하다.
- 추적조사에서 탈락된 대상자나 연구에 끝까지 참여한 대상자나 같은 예후를 가진다.
즉, 생존분석에서 중도 절단의 발생은, 시간 경과에 따른 생존 여부와는 완전히 무관하게 무작위로 발생하여야(random censoring) 하며, 생존시간과는 독립적이어야 한다.
- 각 연구 대상자가 중도 절단 될 확률은, 사건이 발생한 연구 대상자가 경험할 확률과는 관계가 없다.
- 비례위험 모형에 한하여, hazards에 영향을 미치는 어떤 요인은, 연구기간 동안 항상 같은 비율로 작용한다.

Note 4. 생존 분석의 주요 역사[47]

- 1950~1960년대 임상 데이터가 폭발적으로 증가 & 발전하던 시기임.
 - 1969, 1972 Cumulative Hazard Function의 non-parametric 추정량이 Nelson에 의해서 처음으로 제안됨.
 - 1970년대 노르웨이의 통계학자들과 버클리의 통계학자들이 공동연구를 많이 함.
 - 1972 1972년에 Aalen은 석사학위논문을 통해서 Nelson과 똑같은 추정량을 석사학위논문을 통해 제안.
 - 1973 Kaplan Meier 추정량의 matrix 버전이 소개됨.
 - 1974 Breslow에 의해 Cumulative baseline hazard rate가 추정되었다.
 - 1975 Aalen은 박사학위논문을 통해 마팅게일과 더불어 hazard function과 현재의 Nelson-Aalen 추정량이라 불리우는 형태를 소개함. (Nelson이 1969년의 그 Nelson이다)
 - 1976 Aalen이 마팅게일로써의 two-sample test를 소개함.
 - 1978 Aalen은 위에서 weight function이 적용된 특별한 경우의 test score를 보여줌.
코펜하겐에서 recurrent event analysis에 관한 프로젝트가 시작되었다.
 - 1979~1980 Fleming에 의해, 마팅게일 카운팅 어프로치를 바탕으로 한 Kaplan-Meier 추정량이 발표됨.
 - 1980년대 Aalen이 additive hazards model을 제공하기 시작했다.
 - 1981 Per Kragh Anderson에 의해 Cox score function과 다양한 접근 속성을 위한 마팅게일 성질이 적용되기 시작했다.
(헷갈릴 수 있는데, Anderson-Darling의 Theodore Wilbur Anderson과 다른 사람이다.)
 - 1981 Time invariant covariates의 경우 large sample properties를 다루는 전통적인 모델이 제공되기 시작함.
 - 1985 Anderson과 Bogan에 의해 multivariate counting processes modeling으로 Cox model이 확장됨.
 - 1988, 1990 Goodness-of-fit test를 기반으로 한 Martingale residual이 정의됨
-

4 생존 분석 방법의 분류

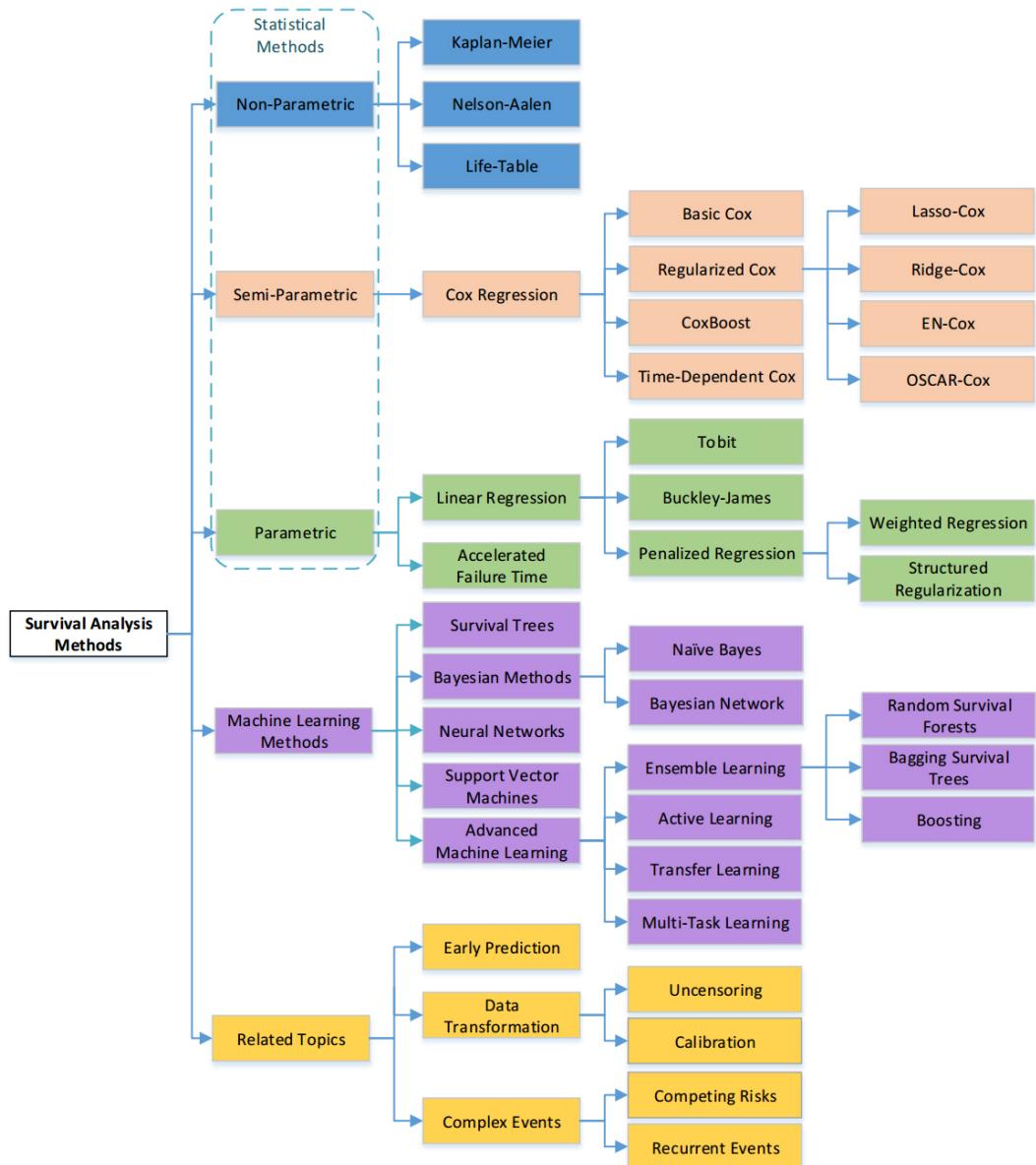


Figure 3: Ping Wang et al.에 의해 소개된 생존 분석 방법 분류

Ping Wang et al.[51]은 생존 분석 방법을 Figure 3와 같이 분류하였다.

분류의 깊이가 깊어질수록 분류 기준이 엄밀하지 않다는 점⁶은 어렵지만, 생존 분석에 대한 전체적인 큰 틀과 어떠한 방법들이 있는지를 확인할 수 있으며, CS 분야 연구자들에 의하여 나온 상당히 최근의 분류라는 점에 의의가 있다고 여겨진다.

⁶구체적으로는 다음과 같은 점을 들 수 있다.

- Naive Bayes, Bayesian Network는 Bayesian Methods가 아니다.
- Advanced Machine Learning을 따로 분류하는 대신, Survival Tree와 Random Survival Forest, Bagging Survival Trees와의 관계를 보인다던지 하면 더 좋았을 것이다.
- Complex Event에 있어, Recurrent Event와 Competing Risk 뿐만 아니라, Clustered Failure time과 Multi-state Model도 중요한 이슈인데, 이에 대한 분류가 빠져있는 것이 아쉽다. 등.

5 생존분석 관련 소프트웨어 패키지

Algorithm	Software	Language	Link
Kaplan-Meire Nelson-Aalen Life-Table	survival	R	https://cran.r-project.org/web/packages/survival/index.html
Basic Cox TD-Cox	survival	R	https://cran.r-project.org/web/packages/survival/index.html
Lasso-Cox Ridge-Cox EN-Cox	fastcox	R	https://cran.r-project.org/web/packages/fastcox/index.html
Oscar-Cox	RegCox	R	https://github.com/MLSurvival/RegCox
CoxBoost	CoxBoost	R	https://cran.r-project.org/web/packages/CoxBoost/index.html
Tobit	survival	R	https://cran.r-project.org/web/packages/survival/index.html
BJ	bujar	R	https://cran.r-project.org/web/packages/bujar/index.html
AFT	survival	R	https://cran.r-project.org/web/packages/survival/index.html
Bayesian Methods	BMA	R	https://cran.r-project.org/web/packages/BMA/index.html
RSF	randomForestSRC	R	https://cran.r-project.org/web/packages/randomForestSRC/index.html
BST	ipred	R	https://cran.r-project.org/web/packages/ipred/index.html
Boosting	mboost	R	https://cran.r-project.org/web/packages/mboost/index.html
Active Learning	RegCox	R	https://github.com/MLSurvival/RegCox
Transfer Learning	TransferCox	C++	https://github.com/MLSurvival/TransferCox
Multi-Task Learning	MTLSA	MATLAB	https://github.com/MLSurvival/MTLSA
Early Prediction Uncensoring	ESP	R	https://github.com/MLSurvival/ESP
Calibration	survutils	R	https://github.com/MLSurvival/survutils
Competing Risks	survival	R	https://cran.r-project.org/web/packages/survival/index.html
Recurrent Events	survrec	R	https://cran.r-project.org/web/packages/survrec

Table 2: Ping Wang et al.에 의해 소개된 생존분석 관련 소프트웨어 패키지

Ping Wang et al.[51]은, 생존 분석과 관련된 소프트웨어 패키지도 Table 2와 같이 정리하였으며, 이는 Figure 3와 비교하면서 보면 도움이 된다.

다만, Table 2의 Link는 Ping Wang et al.이 소개한 것에서 다소 달라진 부분이 있었으므로, 본 노트에는 2017년 8월 8일 기준으로 수정하였다.

대부분의 소프트웨어가 R을 이용하여 작성되었다는 점이 특이하다.

Part III

생존 함수

6 반응변수

관심있는 반응변수를 T 라고 표시하자. (여기서 $T > 0$) (일반적인 통계 모형에서는 Y 로 주로 표시한다.)

T 는 0보다 큰 확률변수이며, 확률밀도함수(PDF) $f(t)$ 와 누적분포함수 $F(t)$ 를 가진다.

$f(t)$ 는 수명 분포의 의미를 가진다. 즉, 수명분포는 비음(non-negative)의 값만을 취하는 확률변수 T 의 분포로, 이 때, $t \leq 0$ 에 대하여 $F(t) = 0$ 을 만족한다. 아마도 우리 연구에서는 이를 연속형 확률 변수로 간주할 것이지만, 수명이 '복사기의 사용 횟수'와 같이 정의되는 경우, 이산형 확률 변수로도 간주할 수 있다.

시점 $t = 0$ 에서는 모든 시스템이 정상 가동 상태임을 가정하며, 따라서 불량품이나 가동불능 시스템은 처음부터 고려대상에서 제외된다.

수명은 비음의 값만을 취하기 때문에, 이 경우 $F(t)$ 는 다음과 같은 조건을 만족한다.

- $F(t) = P(T \leq t)$ 는 시스템이 시점 t 이전에 고장이 나는 확률이며, $\lim_{t \rightarrow \infty} F(t) = 1$ 을 만족한다.
- 모든 $t \leq 0$ 에 대하여 $F(t) = 0$. 즉, 음의 수명을 가질 수 없다.
- $t_1 < t_2$ 이면, $F(t_1) \leq F(t_2)$ 이다.

일반적인 통계분석에서는 관측치의 평균, 분산 등 요약통계량을 구하는 것이 의미있을 수 있지만, 생존분석에서는 censored data로 인해 이런 기술통계량이 그다지 의미있게 사용되지 못할 수 있다. 대신 직접적으로 분포함수를 추정함으로써 자료를 요약할 수 있다. 특히 평균 수명(mean life length)나 백분위수(percentile) 등은 그 자체로 수명분포를 완전하게 묘사할 수는 없으나, 수명에 관한 전체적인 경향을 설명할 수 있는 척도로 사용되고 있으며, 수명에 대한 정량적인 대표값이다.

우리가 알고자 하는 대상인 사건 발생 시간 $\tilde{T}_1, \dots, \tilde{T}_n$ 은 서로 독립이고, 동일한 분포 F 를 가진다.

7 생존 함수(survival function), 신뢰도 함수(reliability function)

사건까지의 시간(time-to-event)에 대한 현상을 설명하기 위한 생존함수(survival function)를 다음과 같이 정의한다.

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx = 1 - P(T \leq t) = 1 - F(t)$$

앞서 각주에서 이야기했지만, 생존 분석에서는 $S(t)$, 신뢰성 공학에서는 $R(t)$ 로 표현한다. 하지만 이 둘의 차이는 없다. 이 노트에서는 가급적 $S(t)$ 로 표기할 것이다.⁷ 혼동하지 않도록 주의하자.

- 생존함수는 비증가함수(non-increasing function)이다. 즉, $t_1 < t_2$ 이면, $S(t_1) \geq S(t_2)$.
- 사건 발생시간 t 는 처음부터 발생하거나 영원히 발생하지 않을 수 있으므로, 범위는 $t \in [0, \infty)$ 이다.
- $S(0) = 1$, $S(\infty) = \lim_{t \rightarrow \infty} S(t) = 0$
- 생존함수는 확률로 표현되므로 $0 \leq S(t) \leq 1$ 을 만족한다.
- $S(t) = P(T > t)$ 는 ”적어도 t 시점까지 생존할 확률”이며, 사건 발생이 t 시점 이후에 일어날 가능성이다. 신뢰성 공학이라면 ”나이가 t 인 시스템이 아직도 가동되고 있을 확률”이라고 표현할 것이다.
- $P(T \leq t)$ 는 t 시점 이전에 사건이 발생할 확률이다.
- $S(t)$ 에 대한 1차 도함수의 음의 값이, 확률밀도함수 $f(t)$ 가 된다. $f(t) = -\frac{dS(t)}{dt}$
- 일반적으로 $F(t) = 1 - S(t)$ 를 누적분포함수라고 하되, 신뢰성 공학에서는 신뢰도 함수와 대응하여 ”불신뢰도 함수(unreliability function)”이라고 부르기도 한다. 따라서, 신뢰도 함수와 불신뢰도 함수의 합은, 모든 t 에 대해서 1이다.

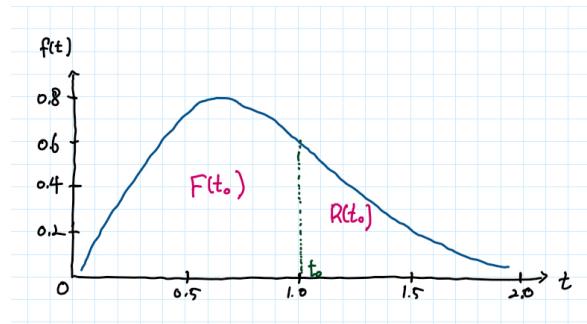


Figure 4: $F(t)$ 와 $R(t)$ 의 관계

Figure 5의 왼쪽은 연속변수에 대한 연속 생존함수를 보여주고, 오른쪽은 관측된 데이터에서 추정된 비모수적 계단함수 형태의 생존함수를 보여준다.

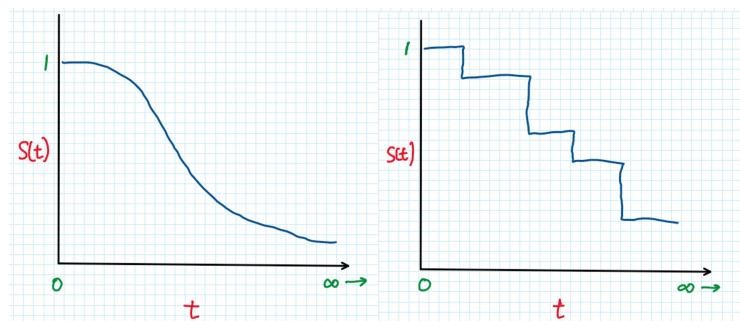


Figure 5: 이론적 생존함수(왼쪽)과 추정된 생존함수(오른쪽)

⁷ 다만 ”신뢰성 공학과 관련된 문맥에 따라” 혼용하여 사용할 수 있을 것이다.

8 위험 함수(hazard function), 위험률 함수(hazard rate function; 고장률 failure rate function)

위험률(hazard rate)은 ” t 시점에서 생존한 조건 하에서 사건이 발생할 확률”이다. 신뢰성 공학이라면, ”나이가 t 인 시스템이 가지는 고장 발생의 위험 정도”로 정의할 것이며, 이를 ”고장률(failure rate)”로 부르기도 한다.

Note 5. 위험률 함수 $h(t)$ 의 유도

시스템이 구간 $(t, \Delta t]$ 에 고장이 발생하는 확률은 다음과 같다.

$$\begin{aligned} P(t < T \leq t + \Delta t) &= \int_t^{t+\Delta t} f(u) du \\ &= F(t + \Delta t) - F(t) \end{aligned}$$

시스템이 시점 t 에 가동된다는 것이 주어지면,

$$\begin{aligned} P(t < T \leq t + \Delta t | T > t) &= \frac{P(t < T \leq t + \Delta t)}{P(T > t)} \\ &= \frac{F(t + \Delta t) - F(t)}{S(t)} \end{aligned}$$

이 값을 Δt 로 나누어서, 구간 $(t, \Delta t]$ 에서의 평균 조건부 확률을 구하면,

$$\frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)}$$

위험률 함수 $h(t)$ 는 $\Delta \rightarrow 0$ 을 위한 평균 조건부 확률이다.

$$\begin{aligned} h(t) &= \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)} \\ &= \frac{1}{S(t)} \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \frac{f(t)}{S(t)} \quad (t \geq 0) \end{aligned}$$

위험함수에 대한 해석을 생각해보자.

- 위험함수 값은 0 이상이다. $h(t) \geq 0, t \geq 0$
- $\int_0^\infty h(t) dt = \infty$
- 위험함수는 확률로 표현되므로, 일반적으로 관측 불가능한 양이다. 그러므로 데이터에 근거하여 추정해야 한다.
- 위험률은 표본 전체에 대한 특성이라기 보다는, 개인에 대한 특성으로 이해할 수 있다. 그러므로 개개인의 위험함수를 추정할 수 있다.
- 일반적인 위험함수의 세 가지 형태는 꾸준히 증가 / 꾸준히 감소 / 일정한 형태이다. (물론 다른 형태의 위험함수도 고려할 수 있다.)

9 누적 위험 함수(cumulative hazard function), 누적 고장률 함수(cumulative failure rate function)

누적 위험 함수는 시점 0부터 t 시점 까지의 누적된 위험률(누적된 고장률)로 정의한다. 수명 분포의 비모수적 분류에 유용하게 사용하는 척도이다.

$$\begin{aligned} H(t) &= \int_0^t h(u)du \\ &= \int_0^t \frac{\frac{d}{dt}[1 - S(t)]}{S(t)} \\ &= -\log S(t) \end{aligned}$$

Note 6. 누적함수를 이용하여 생존함수 구하기

생존함수를 누적함수를 이용해 다음과 같이 표현할 수 있다.

$$\begin{aligned} S(t) &= \exp[-H(t)] \\ &= \exp\left[\int_0^t h(u)du\right] \end{aligned}$$

$H(t)$ 는 다음과 같은 성질을 가진다.

- $H(0) = 0$
- $\lim_{t \rightarrow \infty} H(t) = \infty$
- $H(t)$ 는 t 의 비 감소함수
- $\frac{dH(t)}{dt} = h(t)$, 즉, 누적 위험 함수의 1차 도함수는 위험률 함수이다.

10 생존 함수간 관계

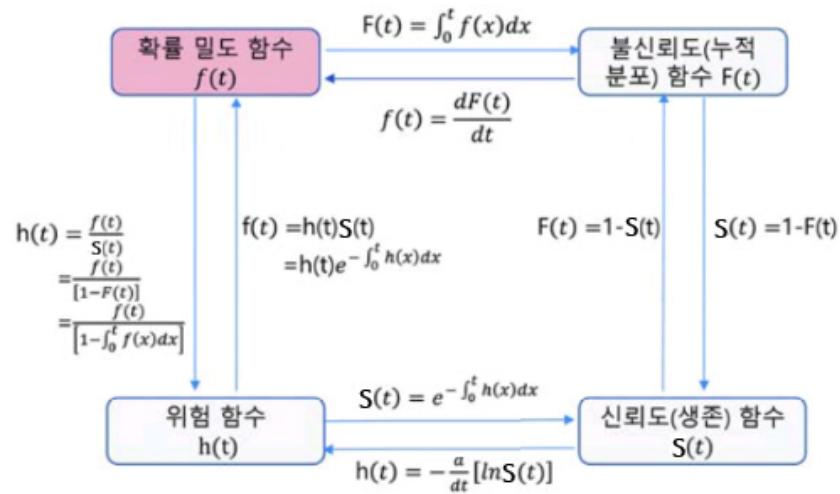


Figure 6: 생존 함수들 간 관계도

11 위험률 함수와 생존함수의 1:1 대응

위험률 함수(고장률 함수)가 중요한 척도 함수로 사용되고 있는 것은, 생존 함수(신퇴도 함수)와 1:1 대응 관계가 성립한다는 것이다. 이러한 성질은 위험률 함수에 대한 지식과 생존 함수에 대한 지식은 동치라는 것을 의미하며, 경우에 따라 생존 함수에 대한 직접적인 추정 대신, 위험률 함수에 대한 추정을 수행함으로써 수명 분포에 대한 분석을 할 수 있다는 의미가 된다.

위험률 함수가 주어지면, 생존 함수는 다음과 같이 계산된다.

$$\begin{aligned} h(t) &= \frac{1}{S(t)} \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \frac{f(t)}{S(t)} = \frac{\frac{d}{dt} F(t)}{S(t)} = \frac{\frac{d}{dt} [1 - S(t)]}{S(t)} \\ &= \frac{-S'(t)}{S(t)} \\ &= -d \ln[S(t)] \end{aligned}$$

양변을 0부터 t 까지 적분하면, $S(0) = 1$ 이므로,

$$\begin{aligned} \int_0^t h(u) du &= \int_0^t \frac{-S'(u)}{S(u)} du \\ &= -\ln S(u)|_0^t \\ &= -\ln S(t) \end{aligned}$$

이 관계로부터, 다음과 같은 공식을 얻는다. 이를 역 공식(inversion formula)라고 부르며, 위험률 함수와 생존 함수의 1:1 관계를 나타낸다.

$$S(t) = e^{-\int_0^t h(u) du}$$

누적 위험 함수 $H(t)$ 를 이용하면, 이 역 공식은 다음과 같이 표시된다.

$$S(t) = e^{-H(t)}$$

11.1 역공식을 이용한 위험함수와 생존함수의 관계

여러 가지 형태의 위험함수에 대한 생존함수를 역 공식을 이용하여 구할 수 있다.

- 위험 함수가 상수인 경우 ($h(t) = \lambda$, $\lambda > 0$)

$$\int_0^t h(u) du = \int_0^t \lambda du = \lambda t$$

따라서, $S(t) = e^{-\lambda t}$ 가 얻어진다. 이 분포는 지수분포로 알려져 있다.

* 지수 분포에 대한 자세한 내용은 Chapter 16.1.11를 참고하자.

- 위험 함수가 선형 증가인 경우 ($h(t) = \alpha t + \beta$, $\alpha > 0, \beta > 0$)

$$\int_0^t h(u) du = \int_0^t (\alpha u + \beta) du = \frac{1}{2} \alpha t^2 + \beta t$$

따라서, $S(t) = e^{-\frac{1}{2} \alpha t^2 + \beta t}$ 가 얻어진다. 이 분포는 선형증가 위험률 분포로 알려져 있다.

* 선형 증가 분포에 대한 자세한 내용은 Chapter 16.1.21를 참고하자.

- 위험 함수가 멱함수인 경우 ($h(t) = \alpha t^{\beta-1}$, $\alpha > 0, \beta > 0$)

$$\int_0^t h(u) du = \int_0^t (\alpha u^{\beta-1}) du = \frac{\alpha}{\beta} t^\beta$$

따라서, $S(t) = e^{-\frac{\alpha}{\beta} t^\beta}$ 가 얻어진다. 이 분포는 와이블 분포로 알려져 있다.

- 위험 함수가 지수적으로 증가하는 경우 ($h(t) = \alpha e^{\beta t}$, $\alpha > 0, \beta > 0$)

$f(t) = \alpha e^{\beta t} e^{-\frac{\alpha}{\beta}(e^{\beta t}-1)}$, $S(t) = e^{-\frac{\alpha}{\beta}(e^{\beta t}-1)}$ 가 얻어진다.

이 분포는 극한값 분포(extreme value distribution)인 Gompertz 분포로 알려져 있다.

* Gompertz 분포에 대한 자세한 내용은 Chapter 16.1.16를 참고하자.

- 욕조 모양(bathtub-shaped)의 위험 함수인 경우

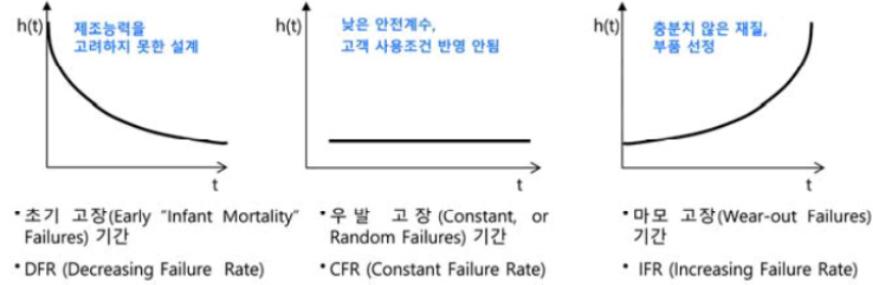


Figure 7: 일반적인 위험률 곡선의 유형

- **초기 고장기간(Infant Mortality Failures; DFR: Decreasing Failure Rate):**

일반적으로 대부분의 시스템에 대한 고장률의 함수는 초기의 짧은 기간동안 감소하게 된다.

부품의 결함이나 제조, 공정상의 실수, 설계상의 오류 등에 기인한 것이다. 의학에서는, 과거 의학이 발달하지 않은 시절에, 초기 신생아의 사망률은 높지만, 어느정도 나이가 지나면 사망률이 줄어드는 경우를 들 수 있다.

- **우발 고장기간(Constant Failures or Random Failures; CFR: Constant Failure Rate):**

초기 고장기간이 끝나게 되면, 일정한 고장률에 도달하게 된다.

이 일정기간 동안에는 고장이 우연히 발생하게 되며, 비교적 시스템의 나이에 영향을 받지 않는다.

- **마모 고장기간(Wear-out Failures; IFR: Increasing Failure Rate):**

시스템의 나이가 많아짐에 따라 고장률이 급속도로 증가하게 된다.

이 때는 부식, 산화, 마모, 피로도 누적, 균열, 수명이 짧은 부품의 사용, 충분치 않은 정비 등에 기인한다. 의학에서는, 나이가 들수록 사망할 확률이 증가하는 경우를 들 수 있다.

DFR, CFR, IFR을 모두 합친 뒤 매끄럽게 연결하면, 그림 8와 같은 욕조 곡선(Bathhub Curve)을 얻는다.

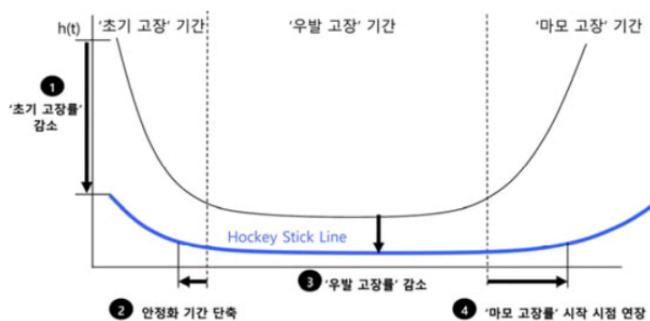


Figure 8: 욕조 곡선(bath-hub curve)

11.2 분포의 비모수적 분류

화률분포의 모수적 분류는 화률밀도함수나 누적분포함수의 형태에 의하여 이루어지는데 반해, 수명분포의 비모수적 분류는 수명분포를 특징지어주는 생존함수, 위험률함수, 평균잔여수명함수 등 여러 가지 척도함수들의 단조성(monotonicity)이나 기타 성질에 의하여 분류된다.

11.2.1 비모수적 군

Note 7. 평균 위험률(평균 고장률)

$$\bar{H}(t) = \frac{1}{t} H(t) = \frac{1}{t} \int_0^t h(u) du$$

- 위험률 함수 $h(t)$ 에 의한 분류:

만일 $h(t)$ 가 $t \geq 0$ 의 비감소함수이면, 수명분포 F 는 IFR(Increasing Failure Rate)에 속하고,

만일 $h(t)$ 가 $t \geq 0$ 의 비증가함수이면, 수명분포 F 는 DFR(Decreasing Failure Rate)에 속한다.

다만, 이 정의는 불신뢰도 함수 $F(t)$ 가 미분 불가능하면 적용되지 않는다. 따라서, 보다 일반적인 정의는 다음과 같다.

만일 $F(0) = 0$ 이고, 모든 $0 \leq s \leq \infty$ 와 $t > 0$ 에 대하여 $\frac{S(s+t)}{S(s)}$ 가 s 의 비증가(감소)함수이면, F 는 IFR(DFR)에 속한다.

- 누적 위험률 함수 $H(t)$ 에 의한 분류:

만일 $F(0) = 0$ 이고,

평균위험률 $\bar{H}(t)$ 이 $t \geq 0$ 의 비감소함수이면, 수명분포 F 는 IFRA(Increasing Failure Rate Average)에 속하고,

평균위험률 $\bar{H}(t)$ 이 $t \geq 0$ 의 비증가함수이면, 수명분포 F 는 DFRA(Decreasing Failure Rate Average)에 속한다.

다만, 이 정의는 불신뢰도 함수 $F(t)$ 가 미분 불가능하면 적용되지 않는다. 따라서, 보다 일반적인 정의는 다음과 같다.

만일 $F(0) = 0$ 이고, 모든 $-\frac{1}{t} \ln S(t)$ 가 비감소(증가)함수이면, F 는 IFRA(DFRA)에 속한다.

IFRA는 IFR보다 큰 군으로, 만일 수명분포가 IFRA에 속하는 독립적인 요소들로 구성된 대부분의 시스템의 수명도 IFRA에 속한다. 그러나 이러한 성질은 IFR 군에는 적용되지 않는다.

- 평균잔여수명함수 $m(t)$ 에 의한 분류:

- 만일 $m(t)$ 가 $t \geq 0$ 의 비증가함수이면, 수명분포 F 는 DMRL(Decreasing Mean Residual Life)에 속하고,
 $m(t)$ 가 $t \geq 0$ 의 비감소함수이면, 수명분포 F 는 IMRL(Increasing Mean Residual Life)에 속한다.

- $\mu = E(T)$ 가 존재한다고 하자. 만일 모든 $t \geq 0$ 에 대하여

$\int_t^\infty S(u)du \leq \mu S(t)$ 이면, F 는 NBUE(New Better than Used in Expectation)에 속하고,
 $\int_t^\infty S(u)du \geq \mu S(t)$ 이면, F 는 NWUE(New Worse than Used in Expectation)에 속한다.

만일 F 가 NBUE이면, 나이 t 의 중고시스템의 평균잔여수명이 새로운 시스템의 평균수명보다 적다는 것을 의미한다.

- $\mu = E(T)$ 가 존재한다고 하자. 만일 모든 $t \geq 0$ 에 대하여

$\int_t^\infty S(u)du \leq \mu e^{-\frac{t}{\mu}}$ 이면, F 는 HNBUE(Harmonic New Better than Used in Expectation)에 속하고,
 $\int_t^\infty S(u)du \geq \mu e^{-\frac{t}{\mu}}$ 이면, F 는 HNWUE(Harmonic New Worse than Used in Expectation)에 속한다.

- 생존 함수 $S(t)$ 에 의한 분류:

- 모든 $s \geq 0$ 과 $t \geq 0$ 에 대하여,

$S(s+t) \leq S(s)S(t)$ 이면, F 는 NBU(New Better than Used)에 속하며,
 $S(s+t) \geq S(s)S(t)$ 이면, F 는 NWU(New Worses than Used)에 속한다.

만일 F 가 NBU이면, 중고시스템의 신뢰도는, 나이에 관계 없이 새로운 시스템의 신뢰도에 비하여 적다.

- 고정된 $t_0 \geq 0$ 과 모든 $t \geq 0$ 에 대하여,

$S(t+t_0) \leq S(t)S(t_0)$ 이면, F 는 NBU- t_0 (New Better than Used of age t_0)에 속하며,
 $S(t+t_0) \geq S(t)S(t_0)$ 이면, F 는 NWU- t_0 (New Worse than Used of age t_0)에 속한다.

만일 F 가 NBU- t_0 이면, 나이 t_0 의 중고시스템의 신뢰도는, 새로운 시스템의 신뢰도에 비하여 적다는 것을 의미한다.

11.2.2 비모수적 군들의 상호관계

앞서 정의된 수명 분포들의 비모수적 군은, 상호 간에 다음과 같은 관계를 갖고 있다.

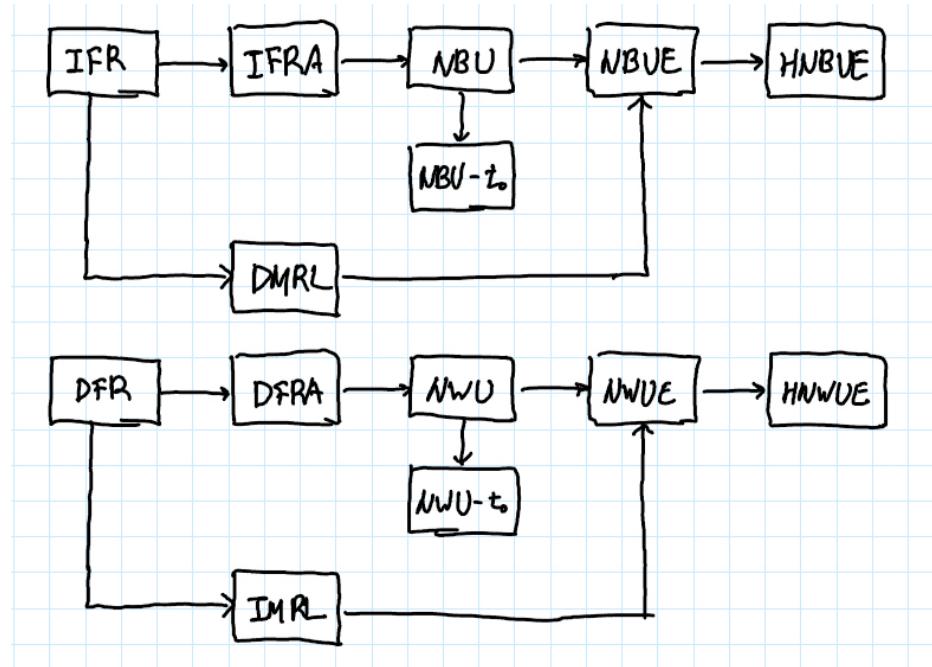


Figure 9: 비모수적 군들의 상호 관계

Part IV

척도 함수

생존률, 신뢰성, 위험률 등에 대한 정량적인 평가는 다음 2가지 견해가 있다.

- 재발(고장) 발생까지 시간 길이의 길고 짧음

재발(고장) 발생까지의 시간 길이에 착안해서, 그 시간이 길면 제품의 신뢰성이 높다고 생각하는 견해이다.

이 경우 생존자료 척도로서 재발(고장)까지의 시간을 확률변수 T 로 표시한다면, 생존함수(신뢰도 함수), 평균고장시간(MTTF; mean time to failure), 백분위수(percentile), Bearing Life 등을 정의한다.

- 일정시간 내 재발(고장) 발생 수의 많고 적음

재발(고장) 수의 많고 적음에 착안하여, 재발(고장)이 적을수록 생존률(신뢰성)이 높다고 하는 견해이다.

이에 대한 대표적인 척도로, 위험률(고장률)에 대해, 고장(재발) 난 경우 수리(치료)를 생각하는 수리계(repairable system)과 수리(치료)를 생각하지 않는 비수리계(non-repairable system)로, 이 둘의 정의가 조금 다르다.

12 '고장(재발) 발생까지 시간 길이의 길고 짧음'에 따른 척도 함수

12.1 생존 함수(Survival Function)

Chapter 7 참고

12.2 평균 수명(mean time to failure)

평균수명(mean time)은 시스템이 고장 날 때까지의 평균 시간으로, 수명 T 의 기대값으로 정의된다.⁸

시스템의 평균 수명은 시스템의 수명을 대표하는 가장 중요한 척도로서, 분포의 무게중심에 관한 정량적인 값을 나타내며, 항상 양의 값을 취한다.

앞서 언급했듯이, 평균수명은 평균잔여수명함수(mean residual life function)에서 $t = 0$ 인 경우로, $m(0)$ 이다.

일반적으로 평균 수명은 T 의 확률밀도함수를 이용하여 다음과 같이 계산된다.

$$E[T] = \int_0^\infty tf(t)dt$$

Note 8. $T \geq 0$ 인 경우 평균수명의 유도

시스템의 수명을 나타내는 확률변수 T 는 항상 비음(non-negative)의 값을 취하므로, 만약 $E(T)$ 가 존재한다면, 즉 $E(T) < \infty$ 이면, $E(T)$ 를 Note 8와 같이 보다 편리하게 유도할 수 있다.

$$\begin{aligned} E[T] &= \int_0^\infty tf(t)dt \\ &= \int_0^\infty td[-S(t)] \\ &= t[-S(t)]_0^\infty - \int_0^\infty [-S(t)] dt \\ &= -\lim_{y \rightarrow \infty} yS(y) + \int_0^\infty [S(t)] dt \end{aligned}$$

이제, $E(T) < \infty$ 이므로, $yS(t) = y \int_y^\infty f(t)dt < \int_0^\infty tf(t)dt$ 성립되며, $\lim_{y \rightarrow \infty} \int_0^\infty tf(t)dt = 0$ 이므로, $\lim_{y \rightarrow \infty} yS(t) = 0$ 된다.

따라서, 확률변수 $T \geq 0$ 인 경우, 다음과 같은 결론을 얻을 수 있다.

$$E(T) = \int_0^\infty S(t)dt$$

⁸ 신뢰성 공학에서는 이 평균수명에 대해, 비수리 시스템에서는 MTTF(Mean Time to Failure), 수리 시스템의 경우에는 수리에 의해 완전히 복구할 수 있다는 가정 하에 MTBF(Mean Time Between Failure)라고 부르기도 한다.

12.3 평균 고장 간격(mean time between failure)

식 (1)이 일정한 경우, $h(t) = \lambda$ 와 같이 표현할 수 있다.

이의 역수를 평균 고장 간격(MTBF; mean time between failure)⁹이라고 한다.

$$MTBF = \frac{1}{\lambda} = \frac{1}{고장률} = \frac{\text{일정 시간}}{\text{일정 시간 내의 고장 수}}$$

이는 수리 후 다음 고장이 발생할 때 까지의 시간(고장 간격)의 평균의 의미를 지닌다.

시스템이 비수리계(non-repairable system)인 경우의 평균 수명을 MTTF(mean time to failure)⁹,

시스템이 수리계(repairable system)인 경우의 평균 수명을 MTBF로 표시한다.

다만, 이 둘은 혼용되기도 한다.

위험률(고장률)이 시간당 λ 라면, 이 장비의 평균 수명을 다음과 같이 구할 수 있다.

$$MTBF = MTTF = E[T] = \int_0^\infty S(t)dt = \int_0^\infty e^{-\lambda t} dt = \frac{1}{\lambda}$$

12.4 백분위수

시스템의 수명 분포가 알려지게 되면, 적률(moment)이나 백분위수(percentile)에 대한 계산이 가능해진다. 이러한 척도는 수명분포 자체보다는 수명에 대해 적은 정보를 제공하지만, 수명분포의 전체적인 패턴을 요약하는 유익한 대표값 중 하나이다.

수명분포와 함께 수명분포의 중심을 나타내는 또 다른 정량적인 척도로 중앙값(median)과 최빈값(mode)이 있다.

이 중 최빈값(mode)은 확률밀도함수가 최대값을 취하는 시점을 나타내며, 중앙값(median)은 확률밀도함수의 면적을 정확하게 양분하는 시점을 나타낸다. 즉, $S(t) = 0.5$ 를 만족하는 t 의 분포의 중앙값이다.

확률밀도함수, 생존함수(신뢰도함수) 및 불신뢰도함수를 각각 $f(t)$, $S(t)$, $F(t)$ 로 각각 표시하면, 100p 백분위수(100pth percentile, pth fractile)은 다음의 두 식을 만족하는 t 값으로 정의되며, t_p 로 나타낸다. (단, $0 < p < 1$)

$$P(T \leq t) \geq p$$

$$P(T \geq t) \geq 1 - p$$

만일 T 가 연속확률변수이면, 위 두 식은 $p \leq F(t) \leq p + P(T = t) = p^o$ 되며,
따라서, $F(t_p) = p$ 또는 $S(t_p) = 1 - p$ 가 된다.

12.5 Bearing Life

생존률(신뢰성)의 평가는, 고객에 대한 제품의 보증을 중요하게 생각하는 입장에서 보면, 평균에 대한 개념 보다는, '제품을 언제까지 안심하고 사용할 수 있는가'와 같이 제품 하나하나에 대해 수명을 보증하는 "한계 보증치"가 중요하다.

Bearing Life는 산포의 크기를 고려한 한계치 보증의 입장에서 고려되고 있는 척도이다.

B_{10} 수명(Bearing 10 % life)은, 전체의 10 %가 고장(재발)날 때까지의 시간을 의미한다.

$$\int_0^{B_{10}} f(t)dt = 0.10$$

생존 함수와 B_{10} 수명과의 관계는 다음과 같다.

$$S(t = B_{10}) = \int_{B_{10}}^\infty f(t)dt = 0.90$$

⁹Chapter 12.2 참고

13 '고장(재발) 발생까지 시간 길이의 길고 짧음'에 따른 척도 함수

13.1 평균 잔여 수명 함수

평균잔여수명함수(mean residual life function)는, 시스템이 나이 t 까지 가동된다는 것이 주어져 있을 때, 앞으로의 잔여 수명에 대한 기대값으로 정의된다.

$$m(t) = E[T - t | T > t]$$

평균수명이 유한한 값을 가지는 경우, 평균잔여수명함수 $m(t)$ 는 모든 $t \geq 0$ 에 대해 다음과 같은 성질을 가진다.

- $m(t) \geq 0$
- $m'(t) \geq -1$
- $\int_0^\infty \frac{1}{m(t)} dt = \infty$
- 새로운 시스템의 평균수명¹⁰은 $t = 0$ 에서의 평균잔여수명이다.

Note 9. 평균잔여수명함수의 유도

먼저 $T > t$ 가 주어졌을 때, 다음과 같은 T 의 조건부 확률밀도함수가 필요하다.

$$f_{T|T>t}(u) = \frac{f(u)}{S(u)}, \quad u > t$$

이를 바탕으로, 조건부 기대값은 다음과 같이 계산된다.

$$\begin{aligned} m(t) &= E[T - t | T > t] \\ &= \int_t^\infty (u - t) f_{T|T>t}(u) du \\ &= \int_t^\infty (u - t) \frac{f(u)}{S(u)} du \\ &= \frac{1}{S(u)} \int_t^\infty (u - t) f(u) du - t \\ &= \frac{1}{S(u)} \int_0^\infty S(u + t) du \end{aligned}$$

Note 10. 평균잔여수명함수와 위험률 함수, 생존 함수간의 1:1 관계

$$\begin{aligned} m'(t) &= \frac{S(t) \frac{d}{dt} \int_t^\infty S(u) du - \int_t^\infty S(u) du \frac{d}{dt} S(t)}{[S(t)]^2} \\ &= \frac{-[S(t)]^2 + f(t) \int_t^\infty S(u) du}{[S(t)]^2} \\ &= -1 + \frac{f(t) \int_t^\infty S(u) du}{[S(t)]^2} \\ \therefore h(t) &= \frac{f(t)}{S(t)} = \frac{1 + m'(t)}{m(t)} \end{aligned}$$

따라서, $m(t)$ 과 $h(t)$ 는 1:1 관계이다. 또한, $S(t)$ 와 $h(t)$ 가 1:1 관계이기 때문에, $m(t)$ 과 $S(t)$ 도 1:1 관계이다.

¹⁰"평균수명(mean time)"과 "평균잔여수명(mean residual life)"은 구분되는 개념이다.

13.2 가용도 함수

수리가 가능한 시스템의 경우, 현실적으로 수리나 보전에 일정한 시간이 걸리는 경우가 대부분이므로, ”수리 시간”을 고장 시간과 별도로 모형화하는 것이 보다 합리적일 수 있다.¹¹ 따라서, 이러한 경우에는 시점 t 에서 시스템이 실제로 가동 상태에 있는지에 대하여 관심을 가지게 된다.

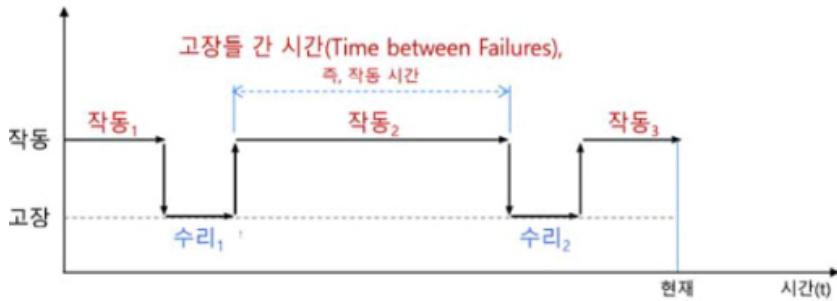


Figure 10: 수리가 가능한 시스템

시점 t 에서 수리 가능 시스템의 상태 함수를 다음과 같이 정의한다.

$$X(t) = \begin{cases} 0 & \text{시스템이 시점 } t \text{에서 고장 상태} \\ 1 & \text{시스템이 시점 } t \text{에서 가동 상태} \end{cases}$$

수리 가능 시스템의 노화에 의한 성능 평가를 위한 척도 중 하나가 **가용도**로, 주어진 시점 t 에서 시스템이 가동 상태에 있는 확률로 정의된다.

가용도를 측정하는 척도에는 다음과 같은 것이 있다. (단, $A(t)$ 는 시스템이 시점 t 에서 가동되고 있는 확률)

- 가용도 함수

$$A(t) = P[X(t) = 1] = E[X(t)] \quad (t > 0)$$

- 극한 가용도

$$A(t) = \lim_{t \rightarrow \infty} A(t)$$

만일, 시스템의 고장률(위험율)과 수리율이 각각 상수인 λ 와 μ 이면, 극한 가용도 $A = \frac{\mu}{\lambda + \mu}$

만일 어느 시스템에 대하여 $A = 0.96$ 이면, 장기적으로 그 시스템은 전체 기간의 96% 기간 동안 가동되고 있다는 것을 의미한다.

수리 불가능 시스템에서는 $A(t) = S(t)$ 가 된다.

¹¹ 의학에서는, 재발 가능한 사건(recurrence available event)에 있어, ”치료 기간”을 별도로 모델링하는 경우를 들 수 있다.

13.3 위험률(고장률)

13.3.1 비수리계(non-repairable system)의 위험률(고장률)

비수리계¹²에서의 위험률(고장률)이란, 임의의 시점 t 까지 동작을 계속해 온 시스템 또는 기기가 그 시점에서 계속해서 단위 시간당 고장이 나는 확률을 말한다.

이를 확률변수 T 의 조건부 확률을 이용하여 표시하면 다음과 같다.

$$P[t < T \leq t + dt | T > t]$$

이는 즉, 시점 t 에서 고장나지 않은 아이템이, 다음 작은 시간 $(t, t + dt)$ 사이에서 고장이 날 확률을 의미한다.

'단위 시간당 고장 확률'로 변환하면 다음과 같고, 이를 작은 시간 $(t, t + dt)$ 에 있어서의 평균 고장률(average failure rate)이라고 한다. 자세한 것은 Chapter 13.3.3을 참고하도록 하자.

$$P\left[\frac{t < T \leq t + dt | T > t}{dt}\right]$$

13.3.2 수리계(repairable system)의 위험률(고장률)

수리계¹³에 대해 시간 $(0, t)$ 의 고장 수(재발 수)를 $n(t)$ ¹⁴라 두면, 이것이 확률변수가 된다.¹⁵ 이 확률변수에 대해 단위 시간당 고장 발생 확률을 다음과 같이 생각할 수 있다.

$$\text{위험률(고장률)} = \lim_{dt \rightarrow 0} \frac{1}{dt} P[(n(t + dt) - n(t)) \geq 1] = \frac{\text{일정 시간 내의 고장 수}}{\text{일정 시간}} \quad (1)$$

13.3.3 평균 위험률(평균 고장률, average failure rate)

구간 $[0, t]$ 의 누적 위험률(누적 고장률)을 $H(t)$ 라 하면, $S(t) = e^{-\int_0^t h(t)dt}$ 로부터,

$$H(t) = \int_0^t h(t)dt = -\ln S(t)$$

구간 $[t_1, t_2]$ 의 구간 위험률(구간 고장률) $FR(t_1, t_2)$ 은

$$FR(t_1, t_2) = H(t_2) - H(t_1) = \ln S(t_1) - \ln S(t_2)$$

구간 $[t_1, t_2]$ 의 구간 평균 위험률(구간 평균 고장률) $AFR(t_1, t_2)$ 은

$$AFR(t_1, t_2) = \frac{\ln S(t_1) - \ln S(t_2)}{t_2 - t_1}$$

구간 $[0, t]$ 의 평균 위험률(평균 고장률) $AFR(t)$ 은

$$AFR(t) = -\frac{\ln S(t)}{t}$$

¹²고장이 났을 때 수리를 생각하지 않는 경우

¹³아이템의 고장 시, 수리에 의해 그 기능을 회복하는 경우

¹⁴수리 시간은 무시하였다. 수리 시간을 고려한다면 별도의 모델링이 필요하다.

¹⁵정확히는, 시간 t 를 파라미터로 하는 확률변수이므로, 확률 과정(random process)가 된다.

Part V

수명 자료(Life Data), 중도 절단 자료(Censored Data)

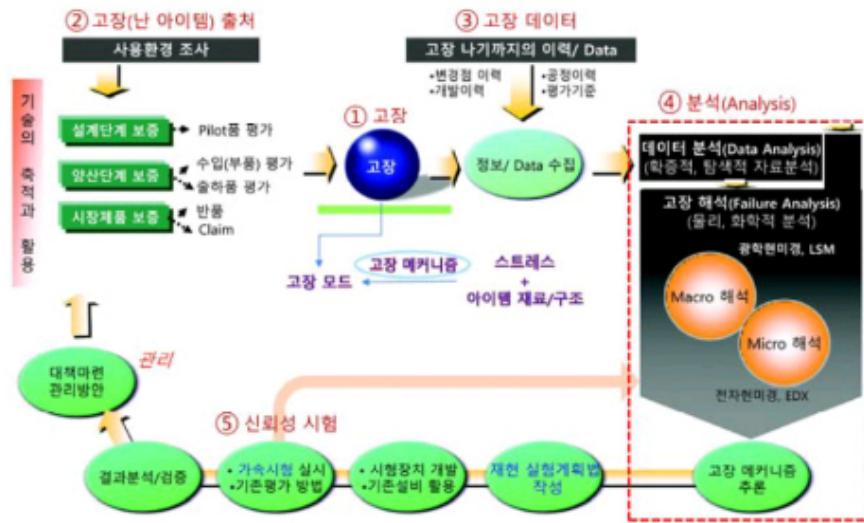


Figure 11: 신뢰성 보증 사이클

"고장이 일어나는 유형", 즉 "위험 유형"은 앞서 욕조 곡선(Bath-tub Curve)를 이용하여 소개한 바 있다. 도메인에 따라 "위험"이 일어나는 매커니즘은 달라질 것이다. 신뢰성 공학에서는 기계의 고장(failure)이 "설계 단계 보증", "시장 제품 보증"의 업무 결과 나타난다고 설명하고 있으며, 고장의 원인을 Figure 12[3]와 같이 정리하고 있다.

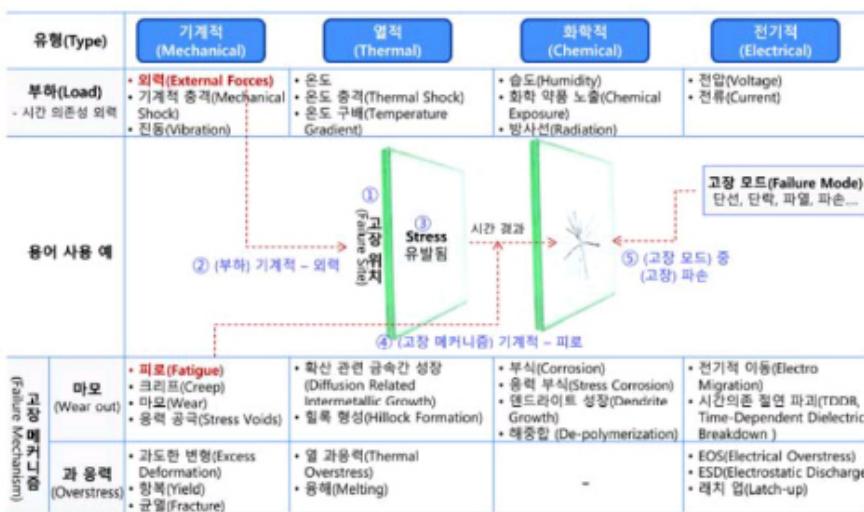


Figure 12: "고장"이 일어나는 유형 및 매커니즘

이러한 수명 데이터(Life Data)의 유형은 다음과 같다.

- 완전 자료(complete data): 각 아이템의 고장(재발) 발생 시간을 정확하게 알고 있는 데이터
- 중도 절단 자료(censored data)

- 우측 중도절단 자료(right-censored data): 수명 시험 동안 고장(재발)이 발생하지 않은 데이터
- 좌측 중도절단 자료(left-censored data): 특정 시간 전에 고장(재발)이 발생한 데이터
- 구간 중도절단 자료(interval-censored data): 특정 시간 사이에 고장(재발)이 발생한 데이터

중도절단의 가장 일반적인 형태는 **구간중도절단(interval censoring)**이다. 중도절단의 가장 일반적인 개념으로 우중도절단(right censoring)과 좌중도절단(left censoring)은 구간중도절단의 특별한 형태이다.

구간중도절단자료란 \tilde{T}_i 를 정확하게 관찰할 수 없으며, 다만 $(L_i, R_i]$ 라는 구간 안에 발생하였다는 것을 알 수 있다.

$$\tilde{T}_i \in (L_i, R_i] \quad (\text{단}, L_i \leq R_i)$$

여기에서, 만일 한쪽 방향의 중도절단만을 가정한다면 이것이 우중도절단(right censoring) 혹은 좌중도절단(left censoring)이 된다.

중도절단시간 C_1, \dots, C_n 이 서로 독립이고, 동일한 분포 F_c 를 가진다고 하자.
만약

- $\tilde{T}_i \leq C_i$ 이면 우중도절단(right censored)
- $\tilde{T}_i \geq C_i$ 이면 좌중도절단(left censored)

이 된다. 좌중도절단이 일어난 경우, 정확한 사건발생시간을 알 수 없다.

그리고 학교 수업시간이나 텍스트북 등으로 배우는 것은 대부분 우중도절단(right censored) 자료들이었을 것이다.

사실 구간중도절단자료의 발생빈도가 매우 높긴 하지만 계산의 복잡성 등으로 인해 관련 연구가 부족한 현실이다. 구간중도절단이 우중도절단의 generalized 형태라고 할 수 있기는 하지만, 중도절단 자료의 특성상 우중도절단 자료를 분석하는 방법으로 구간중도절단 자료를 분석하는 데에 사용하는 것은 곤란하다는 점만 알아두자.

중도절단의 이유는 다음 3가지로 분류할 수 있다.

- **연구 종료:** 연구가 종료될 때 까지 사건이 일어나지 않았다. 이를 행정상 중도절단(administrative censoring)이라고도 한다. 이 경우 중도절단은 관심있는 사건과 무관하다고 할 수 있다.
- **추적 실패(Loss to follow up):** 연구자 연락을 두절한 경우, 더 이상 추적 조사가 불가능하게 된다. 이 경우, 두절된 원인을 조사하는 것이 필요하다.
 - 임상연구에서 참여자가 너무 건강하여 치료법에 대한 필요를 느끼지 못해 연구에서 중도 탈락할 경우, 추정된 생존률은 실제보다 더 낮을 수 있다.
 - 약물의 부작용 등으로 연구에서 일찍 탈락한 겸응, 더 높은 생존률로 잘못 추정될 수 있다.
- **경쟁위험모형(competing risk):** 다른 종류의 사건 발생으로 인해 관심있는 사건이 중도절단된 경우에, 경쟁위험모형을 고려해야 한다.
- **측정 기구의 한계:** 가령 100kg까지의 무게를 챌 수 있는 경우, 120kg인 개인의 몸무게는 100+로 기록된다.

14 우중도절단(right censoring)

Figure 13를 살펴보자.

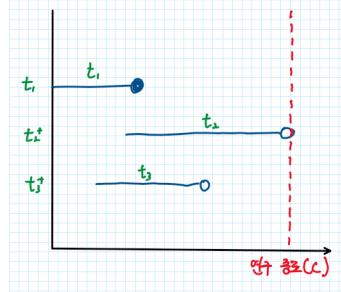


Figure 13: 우중도절단(right-censored)

- 첫 번째 관측 대상은, $t = 0$, 즉 연구 시작 시점부터 시작하여, t_i 시점에서 사건 발생을 겪었다.
- 두 번째 관측 대상은, 연구 시작 후 어느정도 시간이 지난 후에 연구에 참여하여 연구가 종료할 때 까지 사건을 경험하지 않았다.
- 세 번째 관측 대상은, 연구 시작 후 어느정도 시간이 지난 후에 연구에 참여하였으나, 어떠한 이유로 하여금 도중에 종도 절단하게 되었다.

사건 발생 시간 $\tilde{T}_1, \dots, \tilde{T}_n$ 은 서로 독립이고, 동일한 분포 F 를 가지며, 이는 우리가 알고자 하는 대상이다. 한편, 우중도절단시간 C_1, \dots, C_n 또한 서로 독립이고, 동일한 분포 F_C 를 가진다고 하자.
만약 $\tilde{T}_i \leq C_i$ 이면, T_i 를 관측할 수 있지만, 그렇지 못한 경우 C_i 가 관측된다.
이를 우중도절단(right censoring) 되었다고 한다.

앞으로 특별한 언급이 없으면, 이 노트에서 이야기하는 중도절단은 우중도절단(right censoring)을 말한다.

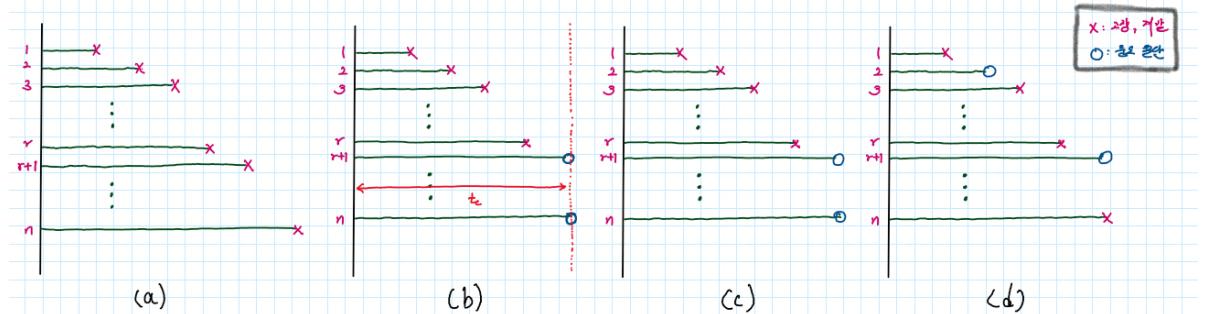


Figure 14: (a): 완전 자료(complete data), (b): Type 1 우중도절단, (c): Type 2 우중도절단, (d): 임의 우중도절단

일반적인 우중도절단은 다음의 3가지 유형으로 나눌 수 있다.

- Type 1 우중도절단(정시종단자료):** 모든 실험 object들은 미리 정해진 시점, 가령 C 까지 관측된다.
즉, 모든 실험 object의 우중도절단 시간이 동일하다. ($C_1 = C_2 = \dots = C_n = C$) 따라서, 중도절단시점 전에 사건이 발생되면, 그들의 T_i 를 관측할 수 있고, 그렇지 않은 경우 $C_i = C$ 가 사용된다.
- Type 2 우중도절단(정수종단자료):** 전체 실험 object 중 미리 정해놓은 사건 발생률을 가질 경우(예를 들어, 전체 실험 object 중 70%가 사건을 가질 경우) 관측을 중지한다고 하자. 따라서 언제 관측이 종료될지는 알 수 없다.
이 경우에도, (비록 알려져 있지 않지만) 모든 관측 object는 같은 중도절단시점 $C_i = C$ 를 가지게 된다.
- 임의(random) 우중도절단:** 위 2가지 유형은, 모든 object의 연구시작 시점이 동일하다는 가정이 있다.
하지만 실제 자료에서는 서로 다른 시점에서 연구에 참여할 수 있으며, 연구가 종료되지 않았음에도 불구하고 여러가지 이유로 더 이상 연구에 참여하지 못할 수도 있다.

우리가 사용할 수 있는 자료는, 중도절단되지 않은 실험 대상자의 정확한 사건발생시점(\tilde{T}_i)과, 중도절단된 대상자에 대한 관측종단시점(C_i)이 된다. 즉, 중도절단된 대상자에 대해선, 종단시점까지는 사건이 발생되지 않음을 알 수 있으므로, 중도절단 시점이 사용된다. 이러한 자료는 다음과 같이 표현한다.

$$\{(T_i, \delta_i), i = 1, \dots, n\} \quad (2)$$

$$\text{단, } T_i = \min(\tilde{T}_i, \delta_i) \text{ and } \delta_i = I(\tilde{T}_i < C_i)$$

14.1 완전 자료(complete data)

시험 대상이 되는 n 개의 표본을 가지고, 시점 $t = 0$ 에서 동시에 수명 시험을 시작한다고 하자. 만일 모든 표본의 고장(재발) 시간 t_1, t_2, \dots, t_n 이 관측되었다면 완전자료가 얻어지게 되며, 이 경우 수명 시험에 소요되는 시간은 $t_{(n)} = \max(t_1, t_2, \dots, t_n)$ 이다.

만일 수명분포에 대한 형태를 모르는 경우에는, 비모수적 방법에 의하여 수명분포의 생존함수 $S(t)$ 를 추정하여야 하는데, 완전자료를 이용하는 경우에 보편적으로 경험적 추정 방법을 사용한다.

이 방법에서는 각각의 고장시간 $t_i (i = 1, 2, \dots, n)$ 에서, 생존율(신뢰도)가 $\frac{1}{n}$ 만큼식 감소하는 계단함수의 형태로 생존함수를 추정한다. 즉, 시점 t 에서의 생존함수에 대한 비모수적 추정값은 다음과 같이 계산된다.

$$\hat{S}_n(t) = \frac{\text{시점 } t\text{에서 아직도 가동되고 있는 부품 수(아직 재발하지 않은 사람 수)}}{n} \quad t \geq 0$$

따라서, 경험적 추정방법에 의한 $F(t)$ 의 추정값은 $\hat{F}_n(t) = 1 - \hat{S}_n(t)$ 에 의해 얻어지며, $\hat{F}_n(t)$ 을 경험적 분포 함수(empirical distribution function)이라고 부른다. n 을 무한히 크게 하면, 모든 t 에 대하여 $\hat{S}_n(t)$ 와 $\hat{F}_n(t)$ 가, 참값 $S(t)$, $F(t)$ 에 각각 수렴한다.

경험적 분포함수를 구하기 위하여, 먼저 n 개의 완전자료 t_1, t_2, \dots, t_n 을 작은 것부터 큰 순으로 나열하여, $t_{(1)} < t_{(2)} < \dots < t_{(n)}$ (순서 통계량으로 정의)으로 표시하자. 그러면, $t_{(i)} \leq t \leq t_{(i-1)}$, ($i = 0, 1, \dots, n$ 에서), 다음과 같이 추정된다.

$$\hat{F}_n(t) = \frac{i}{n} = \frac{\text{시점 } t\text{까지의 고장 부품 수(재발 환자 수)}}{\text{총 표본 수}}$$

여기에서, $t_{(0)} = 0$ 과 $t_{(n+1)} = \infty$ 로 정의한다. 따라서,

$$\hat{S}_n(t) = 1 - \hat{F}_n(t) = \frac{n-i}{n} = \frac{\text{시점 } t\text{에서 아직도 가동되고 있는 부품 수(아직 재발하지 않은 환자 수)}}{\text{총 표본 수}}$$

14.2 Type 1 우중도절단 자료(정시중단자료)

완전자료에서와 마찬가지로 수명 시험은 시험 대상이 되는 n 개의 표본을 가지고 시점 $t = 0$ 에서 동시에 시작한다. 이 경우에는 n 개의 표본에 대한 고장(재발) 시간이 모두 관측될 때까지 기다리지 않고, 미리 정해진 시점 t_c 에서 시험을 중단한다.

이러한 시험 중단은 시간적인 제약이나 비용 등의 이유로, 모든 표본이 고장 날 때까지(재발 할 때까지) 기다리는 것이 불가능한 경우에 적용되는 것으로, 특히 높은 신뢰도를 가진 전자부품 등의 경우에는 모든 부품에 대한 고장 자료를 얻는 것이 매우 어렵다.

따라서, 표본에 대한 시험기간 $(0, t_c)$ 동안에는, 고장 날 일부 표본의 고장시간만이 관측되고, 나머지 표본에 대해서는 고장 시간이 t_c 를 초과한다는 정보만 얻은 채, 정확한 고장 시간을 관측하는 것이 불가능하게 된다. 이러한 경우에 얻어지는 자료가 **Type 1 우중도절단 자료(정시중단자료)**이며, 만일 모든 표본이 시험기간 동안에 고장나는 경우에는 완전자료가 얻어지게 된다.

정시 중단의 경우에는, 시험 중단 시점이 미리 정해지기 때문에, 시험기간 $(0, t_c)$ 동안에 발생하는 고장(재발) 수는 이산화률 변수¹⁶가 되며, 표본 수가 많아지면 고장(재발) 수 또한 증가하게 된다. 시험을 시작한 n 개의 표본에서 얻어지는 정시중단자료의 예는 다음과 같다.

$$t_1, \dots, t_r, t_c^+, \dots, t_n^+$$

여기에서 $r (\leq n)$ 개의 고장 시간은 t_1, \dots, t_r 로 표시되고, 나머지 $(n-r)$ 개의 관측되지 않은 고장시간은 t_c^+ 로 표시된다.

14.3 Type 2 우중도절단 자료(정수중단자료)

n 개의 표본을 가지고 시험을 시작하여, 미리 정해진 $r (\leq n)$ 번째 고장(재발)이 발생한 시점에서 시험을 중단하고 얻는 자료이다. 따라서 이 경우에는 r 개의 고장(재발)시간이 관측되고, 나머지 $(n-r)$ 개의 표본에 대해서는 고장(재발)시간이 r 번째 고장(재발)시간보다 크다는 사실만이 관측되고, 정확한 고장(재발)시간은 관측되지 않는다. 이러한 자료가 **Type 2 우중도절단 자료(정수중단자료)**이며, n 번째 고장(재발)시간에 시험이 중단되면 완전자료가 얻어진다. 정수중단의 경우에는, 시점 중단 시점이 연속확률변수로 간주되며, r 을 크게 잡으면 시험기간이 길어지게 된다. 일반적으로 규모가 작고 대량으로 생산되는 전자부품 등의 수명시험에는, 정해진 시간 대신에 미리 정해진 개수의 고장이 발생한 시점에서 시험을 중단하는 방법이 종종 사용되고 있다.

Type 1 우중도절단(정시중단)자료의 경우, 시험중단 시점이 미리 정해지기 때문에, 시험기간 동안 고장(재발) 자료가 전혀 관측되지 않을 위험이 있으나, Type 2 우중도절단(정수중단)자료의 경우, 미리 정해진 개수의 고장(재발)자료를 얻는 것이 보장되어 있다.

n 개의 표본을 가지고 시작된 시험에서 얻어지는 정수중단자료의 예는 다음과 같다.

$$t_{(1)}, \dots, t_{(r)}, t_{(r)}^+, \dots, t_{(n)}^+$$

여기에서 $t_{(1)} < t_{(2)} < \dots < t_{(n)}$ 은 관측된 r 개의 고장(재발) 시간을 순서대로 나열한 것이고, $(n-r)$ 개의 관측되지 않은 고장(재발) 시간은 $t_{(r)}^+$ 로 표시된다.

¹⁶예를 들면, 포아송 분포

14.4 임의 우중도절단 자료

불완전 자료 중에서 정시중단자료와 정수중단자료의 경우에는, 모든 표본에 대한 수명시험이 $t = 0$ 에서 동시에 시작되어, 일정기간이 경과된 시점, 또는 일정한 개수의 고장(재발)이 발생한 시점에서 시험이 중단된다. 따라서, 수명 시험이 중단되는 시점까지는 고장(재발)시간이 관측되고, 중단 시점이 아직도 가능되는 표본에 대해서는 모두 동일한 시험 중단시간(censoring time)이 관측된다. 그러나 경우에 따라서는, 고장(재발) 시간이 관측되기 이전에 각 표본의 사정에 따라 랜덤하게 시험이 중단되는 경우가 많이 발생한다.¹⁷

이러한 수명시험에는 시험중단시점이 각 표본에 따라서 서로 다르게 되며, 고장(재발)이 시험중단시점보다 먼저 발생하는 경우에만 고장(재발)시간이 관측되고, 그렇지 못한 표본에 대해서는 시험중단시간이 관측된다. 이러한 수명시험을 통하여 얻어진 고장(재발)자료는 **랜덤중단자료(randomly censored data)**가 되며, 이 경우에는 시험 대상이 되는 부품(환자)의 수명분포와 시험중단시간에 대한 분포를 동시에 고려함으로써, 고장(재발)시간에 대한 해석적인 분석이 가능하게 된다.

랜덤중단자료는 가장 일반적인 불완전자료로, 정시중단자료와 정수중단자료를 특별한 경우로 포함하고 있다. 랜덤중단자료의 경우에는 각 표본의 시험주단시간이 랜덤하게 발생하기 때문에, 각 고장(재발)의 발생시점에서의 생존함수가 $\frac{1}{n}$ 만큼 감소되는 것이 아니고, 시험 중단된 표본의 수에도 의존하게 된다. 따라서 **완전자료의 경험적 분포함수와 동일한 방법으로 표본의 수명분포를 추정할 수 없다.**

이 경우, 표본의 수명분포를 추정하기 위해 사용하는 방법으로는 piecewise exponential 방법 등도 있으나, 가장 보편적으로 많이 사용하는 추정 방법은 Kaplan-Meier에 의해 제안된 PL(Product Limit) 추정방법이 많이 사용된다.

14.4.1 Kaplan-Meier Estimation, Product Limit Estimation

n 개의 표본을 가지고 수명시험을 시행하는 경우 얻어지는 랜덤중단자료는 다음과 같은 형태로 얻어진다.

$$(t_1, \delta_1), (t_2, \delta_2), \dots, (t_n, \delta_n)$$

- $\delta_i = 0$ 이면 대응되는 t_i 는 표본 i 의 시험중단시점
- $\delta_i = 1$ 이면 대응되는 t_i 는 표본 i 의 관측된 고장(재발)시간

랜덤중단자료를 이용하여 생존함수에 대한 Kaplan-Meier 추정량을 구하기 위하여, 먼저 t_1, t_2, \dots, t_n 을 작은 것부터 큰 것으로 순차적으로 나열하여 순서 통계량을 $t_{(1)} < t_{(2)} < \dots < t_{(n)}$ 을 얻는다. 그러면 관측값과 대응되는 표본은 다음과 같이 나열된다.

$$(t_{(1)}, \delta_{(1)}), (t_{(2)}, \delta_{(2)}), \dots, (t_{(n)}, \delta_{(n)})$$

그러면 $t_{(i)} \leq t \leq t_{(i+1)}$, $i = 0, 1, \dots, n$ 에서의 생존함수 $S(t)$ 에 대한 Kaplan-Meier 추정 값은 다음과 같이 계산된다. 단, 여기에서 $t_{(0)} = 0$, $t_{(n+1)} = \infty$ 로 정의한다.

$$\bar{S}_n(t) = \prod_{\ell=1}^i \left(\frac{n-\ell}{n-\ell+1} \right)^{\delta_{(\ell)}}$$

Example 1. Kaplan-Meier 방법에 의한 생존함수 추정

$n = 5$ 인 랜덤중단자료가 다음과 같이 얻어졌다고 하자. 단, +가 표시된 것은 시험 중단 시간을 나타낸다.

$$100, 500, 240+, 350, 450+$$

즉, $(100, 1), (240, 0), (350, 1), (450, 0), (500, 1)$ 의 5개 관측값이며, Figure 15와 같이 표현할 수 있다.

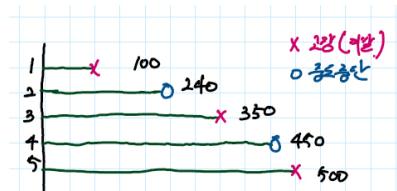


Figure 15: 랜덤 중단 자료

이 때 생존함수 $\bar{S}_n(t)$ 는 다음과 같이 구할 수 있다.

$0 \leq t < 100$ 일 때, $\bar{S}_5(t) = \frac{5}{5} = 1$ 이 된다. Kaplan-Meier 추정 값에 대한 공식을 이용하면, 구간별 생존 함수는 다음과 같이 추정된다.

$$\begin{aligned} 100 \leq t < 350 & \quad \bar{S}_5(t) = \left(\frac{4}{5} \right)^1 = 0.80 \\ 350 \leq t < 500 & \quad \bar{S}_5(t) = \left(\frac{4}{5} \right)^1 \left(\frac{3}{4} \right)^0 \left(\frac{2}{3} \right)^1 = 0.53 \\ 500 \leq t & \quad \bar{S}_5(t) = \left(\frac{4}{5} \right)^1 \left(\frac{3}{4} \right)^0 \left(\frac{2}{3} \right)^1 \left(\frac{1}{2} \right)^0 \left(\frac{0}{1} \right)^1 = 0.00 \end{aligned}$$

위 결과로부터, $\bar{R}_5(300) = 0.8$, $\bar{R}_5(400) = 0.53$ 이 된다는 사실도 알 수 있다.

¹⁷임상실험의 경우, 실험 중에 있는 환자가 다른 병원으로 옮긴다거나, 치료를 중도거부하는 등의 이유로 해서 실험에 빠지는 경우가 있다. 전자부품의 경우, 시험 중의 부품이 마모에 의한 고장이 발생하기 이전에 사용자의 과오나 외부 충격, 부품의 분실 등에 의해 시험이 중단되는 경우가 있다.

14.4.2 Life-table Method

Life-Table Method(생명표 방법)에서는, 측정 단위가 일정한 폭을 가진 구간¹⁸이며, 보험과 신뢰성 분석 자료에서 널리 사용되었다.

Example 2. Life-table Method: 심근경색 환자의 임상시험결과

Table 3는 심근경색 환자(myocardial infarction:MI)와 연관된 임상시험결과를 요약한 자료이다. 여기서 자료는 1년 단위로 생존 환자수, 사망 환자수, 그리고 중도절단된 환자수를 보여 준다.

Year(I_i)	생존 환자 수 (위험 그룹 수, n_i)	사망 환자 수 (d_i)	중도절단 환자 수 (m_i)
[0, 1)	146	27	3
[1, 2)	116	18	10
[2, 3)	88	21	10
[3, 4)	57	9	3
[4, 5)	45	1	3
[5, 6)	41	2	11
[6, 7)	28	3	5
[7, 8)	20	1	8
[8, 9)	11	2	1
[9, 10)	8	2	6

Table 3: 심근경색 환자 자료

위 표에서

- m_i : 구간 i 에서 중도절단된 환자 수
- d_i : 구간 i 에서 사망 환자 수
- n_i : 구간 i 에서 위험 환자 수, $n_i = n_1 - \sum_{j=1}^{i-1} (d_j + m_j)$

생존 함수를 구하기 위해, 다음의 조건부 확률을 고려한다.

$$q_i = P[T \in I_i | T \geq I_{i-1}] = \frac{d_i}{\binom{n_i - m_i}{2}}$$

여기서 q_i 는 $i-1$ 번째 구간까지 살아 있다가, i 번째 구간에서 사망할 확률을 의미한다.

이 때, i 번째 구간에서 m_i 명이 중도절단되었을 때, 그들 중 절반은 생존하였다고 가정한다.

따라서, i 번째 구간에서 위험에 노출된 사람 수는 $\frac{n_i - m_i}{2}$ 로, i 번째 구간에서 생존한 사람 수에서 그 구간에서 중도절단된 환자 수의 절반을 뺀 수로 정의한다.

따라서, i 번째 구간에서 생존할 확률은 $1 - q_i$ 가 된다.

i 번째 구간(I_i)에서 생존하기 위해서는, 처음부터 그 구간까지 계속 생존함을 의미하므로, 조건부 확률의 곱 공식(multiplicative formulae)에 의해 i 번째 구간에서 생존할 확률은 다음과 같다.

$$\begin{aligned} \hat{S}(I_i) &= (1 - q_1)(1 - q_2) \cdots (1 - q_i) \\ &= \prod_{j=1}^i (1 - q_j) \\ &= \prod_{j=1}^i \left[1 - \frac{d_j}{\binom{n_j - m_j}{2}} \right] \end{aligned}$$

이제 $\hat{S}(I_i)$ 의 분산을 고려해보자. $\hat{S}(I_i)$ 는 i 개의 확률($1 - q_j$, $j = 1, \dots, i$)들의 곱 형태로 표현되므로, $\hat{S}(I_i)$ 의 분산을 구하기 위해 먼저 로그 변환을 한다.

$$\log \hat{S}(I_i) = \log(1 - q_1) + \log(1 - q_2) + \cdots + \log(1 - q_i) = \hat{\theta}$$

따라서, $\{\log(1 - q_j), j = 1, \dots, i\}$ 의 분산을 이용하여, $\log S(I_i)$ 의 분산을 구할 수 있다.

여기서, $Var[\log(1 - q_j)] = \frac{d_j}{\binom{n_j - m_j}{2} - d_j}$ 으로, Delta Method를 이용하여 $\hat{S} = e^{\hat{\theta}}$ 의 분산을 유도한다.

$$\begin{aligned} \widehat{Var} [\hat{S}(I_i)] &= \hat{S}^2(I_i) [Var \{\log(1 - q_1)\} + Var \{\log(1 - q_2)\} + \cdots + Var \{\log(1 - q_i)\}] \\ &= \hat{S}^2(I_i) \sum_{j=1}^i \left[\frac{d_j}{\binom{n_j - m_j}{2} - d_j} \right] \end{aligned}$$

¹⁸예를 들어, 월, 분기, 년 등

작성중...

15 Truncation

생존분석 시 나오는 용어로 truncation이라는 것이 있다.

이 truncation은 censoring과는 다른 기전에 의해 발생되는데, 어떤 관찰된 특정한 시간의 전/후에 사건이 발생하는 개인에게만 생긴다.

Truncation이란, 어떤 시간 간격(T_L, T_R)에서만 생존할 것으로 예측되는 개인에게서만 발생되는 sampling bias이다.

여기서 $Y_R = \inf$ 이면 left truncation, $Y_L = 0$ 이면 right truncation이라고 한다.

예를 들면, 고령의 퇴직자들이 거주하는 community에 거주하는 거주자들이 사망하는 데까지 걸리는 시간을 연구하고자 할 때, 여기에 거주하는 대상자는 특정 연령이 되어야만 들어올 수 있으며, 또한 어떤 특정 연령이 되기 전에 사망한 사람들은 관찰 대상에서 제외되어야 하므로 left truncation이 발생한 것이다.

Right truncation의 예는, 1978년 4월 1일 이후에 HIV에 감염되었고, 1987년 3월 31일까지 AIDS가 발병된 258명을 대상으로 연구를 한다면, 이 기간에 HIV에 감염되었고 아직 AIDS가 발병하지 않은 대상자는 연구에서 제외되어야 한다. Event of interest가 HIV에 감염되고 AIDS가 발병하는 기간인 induction time이라고 하면, 이런 특정 기간에 감염된 대상자의 induction은 감염된 시간부터 1987년 3월 31일까지 만큼의 기간 right truncation된 것이다.

Part VI

모수적 방법을 이용한 생존함수의 추정

16 확률분포

일반적으로 통계 자료 분석에서 가장 많이 사용되는 확률분포는 정규분포로, 모든 고전적인 통계 이론과 응용 분야에서 중심적인 역할을 하고 있다. 특히, 표본의 크기가 비교적 큰 경우에는, 중심극한정리에 의하여 정규분포의 중요성이 특히 강조된다. 그러나 음수가 아닌 수명을 확률변수로 고려하는 생존분석과 수명시험에서는, 정규분포를 사용하는 것이 비현실적이다. 따라서 수명 모형 분야에서는 정규분포가 아닌 다른 분포들이 중심적인 역할을 한다.

생존함수의 추정은, 표본으로부터 수명 특성치를 구하는 것으로, 크게 점추정과 구간추정으로 나뉜다.

- 점추정: 모수의 추정값으로써 하나의 수치를 주어진 자료로부터 계산하여 구하고, 오차의 한계를 제공한다.
- 구간추정: 위 점추정에서의 오차의 한계 개념을 이용하여, 모수가 속하게 될 범위를 추정한다.

일반적으로 좋은 추정치는 다음과 같은 성질을 갖는다.

- 불편성(unbiasedness)

모수 θ 의 모든 참값에 대해 $E[\hat{\theta}] = \theta$ 이면, $\hat{\theta}$ 를 θ 의 불편추정량(unbiased estimator)이라고 한다.

- 일치성(consistency)

표본 크기와 관계되는 성질로, 표본 크기가 증가함에 따라 일치 추정량은 모수의 참값에 더욱 가까워지며,

$$\lim_{n \rightarrow \infty} P[|\hat{\theta}_n - \theta| < \epsilon] = 1$$

이 성립하면, $\hat{\theta}_n$ 은 일치성을 갖는다고 한다.

- 유효성(efficiency)

$Var[\hat{\theta}] = E[(\hat{\theta} - \theta)^2]$ 의 값이 최소가 되는 불편추정량 $\hat{\theta}$ 로, 표준오차가 적은 추정량일수록 추정의 정도가 높은 추정량이라 할 수 있다.

- 충분성(sufficiency)

$$\prod_{i=1}^n f(x_i; \theta) = f(\hat{\theta}, \theta) h(x_1, x_2, \dots, x_n)$$

즉, $f(x_i, \theta)$ 의 결합밀도함수가 $\hat{\theta}$, θ 의 함수 g 와, x_1, x_2, \dots, x_n 의 함수 h 로 인수분해될 수 있는 경우로, 표본이 갖고 있는 모수에 대한 정보 모두를 이용하는 추정치를 말한다.

참고로, 충분통계량 T 와 불편추정량 $\hat{\theta}$ 가 있을 때, 조건부 기대값 $E[\hat{\theta}|T]$ 을 취하면, 이것이 충분통계량의 함수인 불편추정량이 되고, 동시에 분산이 줄어든다. (Rao-Blackwell 정리: Note 11 참고.) 그런데, 어떤 불편추정량 $\hat{\theta}$ 으로 시작하던 $E(\hat{\theta}|T)$ 는 $\hat{\theta}$ 에 관계 없이 같은 것이 되고, 따라서 이렇게 구한 $E(\hat{\theta}|T)$ 는 분산이 최소가 된다. 즉, 최소분산 불편추정량(MVUE; Minimum Variance Unbiased Estimator)이 된다.

문제는 $E(\hat{\theta}|T)$ 를 구하는 과정이 때로는 복잡할 수 있고, 따라서 가장 순쉬운 방법은 충분통계량 T 의 함수들 가운데 불편추정량을 찾아보는 것이며, 이것마저 여의치 않다면, 번거롭더라도 $E(\hat{\theta}|T)$ 를 구할 수밖에 없다.

좀 더 엄밀한 의미에서, MVUE임을 보이려면 통계량이 최소 충분성(minimal sufficiency)을 갖고 있음을 보여야 할 필요가 있다. 작성중...

Note 11. 라오-블랙웰(Rao-Blackwell) 정리

$\hat{\theta}$ 이 θ 의 불편 추정량이고, T 가 충분통계량이라고 하자. $\theta^* = E[\hat{\theta}|T]$ 라 정의하면

- θ^* 는 θ 의 불편추정량이다.

Proof: (X_1, \dots, X_n) 이 확률표본이면, $\hat{\theta}$ 는 (X_1, \dots, X_n) 의 함수이고, T 가 충분통계량이므로 $T = t$ 가 주어졌을 때, (X_1, \dots, X_n) 의 분포는 θ 와 무관하다. 따라서 $E[\hat{\theta}|T]$ 는 T 의 함수로 θ 와 무관하다. 즉, θ^* 는 통계량이다.

그런데 다음에 따라, θ^* 는 θ 의 불편추정량이 된다.

$$E[\theta^*] = E[E(\hat{\theta}|T)] = E(\hat{\theta}) = \theta$$

- $Var(\theta^*) \leq Var(\hat{\theta})$ 이다.

Proof:

$$Var(\hat{\theta}) = Var[E(\theta|\hat{T})] + E[Var(\theta|\hat{T})] = Var(\theta^*) + E[Var(\hat{\theta}|T)]$$

여기서 $Var[E(\theta|\hat{T})]$ 는 분산이기 때문에, 모든 t 에 대해서 $Var[E(\theta|\hat{T})] \geq 0$ 이 성립하고, 따라서 $E[Var(\theta|\hat{T})]$ 이 되므로, $Var(\theta^*) \leq Var(\hat{\theta})$ 이 성립한다.

16.1 연속확률분포

여기에서 파라메터에 대한 notation을 다음과 같이 사용하였다.

- $a \in \mathbf{R}$: Location Parameter
- $b > 0$: Scale Parameter
- 그 외의 라틴 문자, 혹은 그리스 문자: Shape Parameter

또한, 특별한 함수에 대해서는 다음과 같이 정의하였다.

- Complete beta function

$$B(c, d) = \frac{\Gamma(c)\Gamma(d)}{\Gamma(c+d)} = \int_0^1 t^{c-1}(1-t)^{d-1} dt$$

- Incomplete beta function

$$B(P, c, d) = \frac{\Gamma(c)\Gamma(d)}{\Gamma(c+d)} = \int_0^P t^{c-1}(1-t)^{d-1} dt$$

- Complete gamma function

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$$

- Incomplete gamma function

$$\Gamma(a, u) = \int_0^u t^{a-1} e^{-t} dt$$

- Complementary incomplete gamma function

$$\gamma(a, u) = \int_u^\infty t^{a-1} e^{-t} dt$$

- CDF of the standardized normal distribution

$$\Phi(u) = \int_{-\infty}^u \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} du$$

- PDF of the standardized normal distribution

$$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

- Error function

$$\text{erf}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} du$$

16.1.1 Alpha Distribution(알파 분포)

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{be^{-\frac{1}{2}(\alpha - \frac{b}{t})^2}}{\sqrt{2\pi}\Phi(\alpha)t^2}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	$\frac{\Phi(\alpha - \frac{b}{t})}{\Phi(\alpha)}$	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - \frac{\Phi(\alpha - \frac{b}{t})}{\Phi(\alpha)}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{be^{-\frac{1}{2}(\alpha - \frac{b}{t})^2}}{\sqrt{2\pi}[\Phi(\alpha) - \Phi(\alpha - \frac{b}{t})]t^2}$	IDHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$\mu(t) = E[T] = \int_0^\infty tf(t)dt$ $= \int_0^\infty S(t)dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u)du}{S(t)}$	존재하지 않음	
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수	$t \geq 0$		
파라메터	$\alpha \in \mathbf{R}, b > 0$		

Table 4: Alpha 분포함수에 기반한 척도 함수

확률변수 $Y \sim N(y; \mu, \sigma)$ 라고 하자. 다만, $y = 0$ 의 왼쪽으로 절삭되어있다.

그리면, $X = \frac{1}{Y} \sim \text{alpha}(x; \alpha, b)$ 이며, 이 때 $\alpha = \frac{\mu}{\sigma}$, $b = \frac{1}{\sigma}$ 이다.

이 분포는 Salvia(1985)[58]가 가속 수명 모형(accelerated life testing)에 적용한 바 있다.

16.1.2 Arcsine Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\left[b\pi \sqrt{1 - \left(\frac{t-a}{b}\right)^2} \right]^{-1}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\frac{1}{2} - \frac{\arcsin\left(\frac{t-a}{b}\right)}{\pi}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\left[b\sqrt{\left(1 - \left(\frac{t-a}{b}\right)^2\right)} \arccos\left(\frac{t-a}{b}\right) \right]^{-1}$	DIHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	$(a - t) + \frac{\sqrt{(a+b-t)(t+b-a)}}{\arccos\left(\frac{t-a}{b}\right)}$	IDMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$a - b \leq t \leq a + b$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 5: Arcsine 분포함수에 기반한 척도 함수

이 분포는 베타 분포의 두 파라메터가 $c = d = 0.5^2$ 일 때에 해당하는 특별한 경우이다. 또한 베타 분포의 두 파라메터가 $c + d = 1, c \neq 0.5$ 일 경우를 generalized arcsine distribution이라고 한다.

Arcsine distribution은 location-scale distribution의 한 종류이다.

작성중...

16.1.3 Beta Distribution(베타 분포)

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{B(c,d)} \cdot \frac{(t-a)^{c-1}(a+b-t)^{d-1}}{b^{c+d-1}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - \frac{1}{B(c,d)} \int_0^{\frac{t-a}{b}} u^{c-1} (1-u)^{d-1} du$ $= 1 - I_{\left(\frac{t-a}{b}\right)}(c, d)$ $= I_{\left(1-\frac{t-a}{b}\right)}(c, d)$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	No closed form	DIHR for $0 < c \lesssim 0.8$ and d arbitrarily IHR for all other combinations of c and d
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty tf(t)dt$ $= \int_0^\infty S(t)dt$	$\frac{c}{c+d}$	
평균잔여수명함수	$m(t) = E[T-t T > t]$ $= \frac{\int_t^\infty S(u)du}{S(t)}$	No closed form	IDMRL for $0 < c \lesssim 0.8$ and d arbitrarily DMRL for all other combinations of c and d
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$a \leq t \leq a+b$	
파라미터		$a \in \mathbf{R}, b > 0, c > 0, d > 0$	

Table 6: Beta 분포함수에 기반한 척도 함수

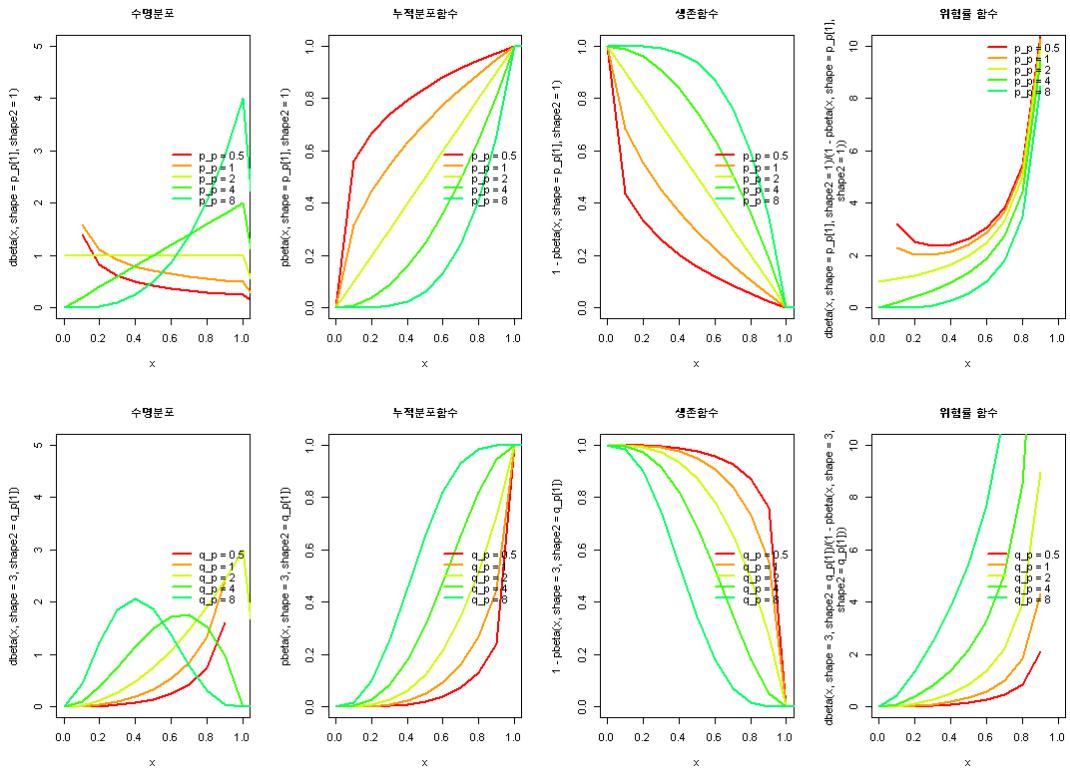


Figure 16: 베타분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 1. 베타분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```
1 ##### Beta Distribution
2 par(mfrow = c(2, 4))
3
4 ### parameter: p_p
5 p_p = c(0.25, 0.5, 1, 2, 4) # shape1
6
7 ### Input Variable
8 x <- seq(0, 10, length.out = 101)
9
10 color = rainbow(10)
11
12 ### Life Distribution
13 plot(x, dbeta(x, shape=p_p[1], shape2=1), xlim=c(0, 1), ylim=c(0, 5), col=color[1], lwd=2, type = 'l', main="Life Distribution")
14 for (i in 2:5) { lines(x, dbeta(x, shape=p_p[i], shape2=1), col=color[i], lwd=2); }
15 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('p_p = 0.5', 'p_p = 1', 'p_p = 2', 'p_p = 4', 'p_p = 8'))
16
17 ### Cumulative Distribution
18 plot(x, pbeta(x, shape=p_p[1], shape2=1), xlim=c(0, 1), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative
19 Distribution")
20 for (i in 2:5) { lines(x, pbeta(x, shape=p_p[i], shape2=1), col=color[i], lwd=2); }
21 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('p_p = 0.5', 'p_p = 1', 'p_p = 2', 'p_p = 4', 'p_p = 8'))
22
23 ### Survival Function
24 plot(x, 1-pbeta(x, shape=p_p[1], shape2=1), xlim=c(0, 1), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
25 for (i in 2:5) { lines(x, 1-pbeta(x, shape=p_p[i], shape2=1), col=color[i], lwd=2); }
26 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('p_p = 0.5', 'p_p = 1', 'p_p = 2', 'p_p = 4', 'p_p = 8'))
27
28 ### Hazard Function
29 plot(x, dbeta(x, shape=p_p[1], shape2=1)/(1-pbeta(x, shape=p_p[1], shape2=1)), xlim=c(0, 1), ylim=c(0, 10), col=color[1], lwd=2,
30 type = 'l', main="Hazard Function")
31 for (i in 2:5) { lines(x, dbeta(x, shape=p_p[i], shape2=1)/(1-pbeta(x, shape=p_p[i], shape2=1)), col=color[i], lwd=2); }
32 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('p_p = 0.5', 'p_p = 1', 'p_p = 2', 'p_p = 4', 'p_p = 8'))
33
34 ### parameter: q_p
35 q_p = c(0.25, 0.5, 1, 2, 4) # shape2
36
37 ### Input Variable
38 x <- seq(0, 10, length.out = 101)
39
40 color = rainbow(10)
41
42 ### Life Distribution
43 plot(x, dbeta(x, shape=3, shape2=q_p[1]), xlim=c(0, 1), ylim=c(0, 5), col=color[1], lwd=2, type = 'l', main="Life Distribution")
44 for (i in 2:5) { lines(x, dbeta(x, shape=3, shape2=q_p[i]), col=color[i], lwd=2); }
45 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('q_p = 0.5', 'q_p = 1', 'q_p = 2', 'q_p = 4', 'q_p = 8'))
46
47 ### Cumulative Distribution
48 plot(x, pbeta(x, shape=3, shape2=q_p[1]), xlim=c(0, 1), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative
49 Distribution")
50 for (i in 2:5) { lines(x, pbeta(x, shape=3, shape2=q_p[i]), col=color[i], lwd=2); }
51 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('q_p = 0.5', 'q_p = 1', 'q_p = 2', 'q_p = 4', 'q_p = 8'))
52
53 ### Survival Function
54 plot(x, 1-pbeta(x, shape=3, shape2=q_p[1]), xlim=c(0, 1), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
55 for (i in 2:5) { lines(x, 1-pbeta(x, shape=3, shape2=q_p[i]), col=color[i], lwd=2); }
56 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('q_p = 0.5', 'q_p = 1', 'q_p = 2', 'q_p = 4', 'q_p = 8'))
57
58 ### Hazard Function
59 plot(x, dbeta(x, shape=3, shape2=q_p[1])/(1-pbeta(x, shape=3, shape2=q_p[1])), xlim=c(0, 1), ylim=c(0, 10), col=color[1], lwd=2,
60 type = 'l', main="Hazard Function")
61 for (i in 2:5) { lines(x, dbeta(x, shape=3, shape2=q_p[i])/(1-pbeta(x, shape=3, shape2=q_p[i])), col=color[i], lwd=2); }
62 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('q_p = 0.5', 'q_p = 1', 'q_p = 2', 'q_p = 4', 'q_p = 8'))
```

베타분포의 특징은 다음과 같다.

1. 베타 분포의 확률밀도함수는 0과 1 사이의 값에 대하여 정의된다.
2. 베타분포의 두 개의 파라메터에 따라 다음과 같은 관계를 갖게 된다.
 - $c = d = 1$: Uniform Distribution
 - $c = 2, d = 1$: The right-angled negatively skewed triangular distribution
 - $c = 1, d = 2$: The right-angled positively skewed triangular distribution
 - $c = d = 0.5$: Arcsine Distribution
 - $c > 0, d = 1$: Power Function Distribution
3. 베타분포는 유사대칭이다. 즉, 베타분포는 두 개의 모수 c 와 d 가 서로 바뀌면, 그 확률밀도함수는 원래 확률밀도함수의 mirror image가 된다.
4. 베타분포의 두 개의 파라메터에 따라 다음과 같은 모양을 하게 된다.
 - $c = d$: Symmetric
 - $c > 1, d > 1$: Unimodal
 - $c < 1, d < 1$: U-shaped
 - $d \leq 1 \leq c, c \neq d$: J-shaped
 - $c \leq 1 \leq d, c \neq d$: Inversely (or reflected) J-shaped

16.1.4 Birnbaum-Saunders Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{\sqrt{\frac{t}{b}} + \sqrt{\frac{b}{t}}}{2ct\sqrt{2\pi}} e^{-\frac{1}{2c^2}(\sqrt{\frac{t}{b}} - \sqrt{\frac{b}{t}})^2}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\Phi\left[-\frac{1}{c}\left(\sqrt{\frac{t}{b}} - \sqrt{\frac{b}{t}}\right)\right]$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	No closed form	
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty tf(t)dt$ $= \int_0^\infty S(t)dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u)du}{S(t)}$	No closed form	
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq 0$	
파라메터		$b > 0, c > 0$	

Table 7: Birnbaum-Saunders 분포함수에 기반한 척도 함수

이 분포는 Birnbaum and Saunders(1968, 1969)[15][16]에 의해 제안되었다.

$Y = \frac{\sqrt{\frac{t}{b}} - \sqrt{\frac{b}{t}}}{c} \sim N(0, 1)$ 이 된다. 생존함수 $S(t)$ 는 이에 착안하여 작성된 것이다.
작성중...

16.1.5 Burr Distribution of Type XII

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{cd}{b} \left(\frac{t-a}{b}\right)^{d-1} \left[1 + \left(\frac{t-a}{b}\right)^d\right]^{-(c+1)}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\left[1 + \left(\frac{t-a}{b}\right)^d\right]^{-c}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{cd}{b} \left(\frac{t-a}{b}\right)^{d-1} \left[1 + \left(\frac{t-a}{b}\right)^d\right]^{-1}$	DHR for $0 < d \leq 1$ IDHR for $d > 1$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T>t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	IMRL for $0 < d \leq 1$ DIMRL for $d > 1$ 존재하지 않음 for $cd \leq 1$
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0, c > 0, d > 0$	

Table 8: Burr 분포함수에 기반한 척도 함수

작성중...

16.1.6 Cauchy Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\left[\pi b \left(1 + \left(\frac{t-a}{b} \right)^2 \right) \right]^{-1}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \leq t) \\ &= 1 - F(t) \\ &= e^{-H(t)} \end{aligned}$	$\frac{1}{2} - \frac{1}{\pi} \arctan \left(\frac{t-a}{b} \right)$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\left[b \left(1 + \left(\frac{t-a}{b} \right)^2 \right) \left(\frac{\pi}{2} - \arctan \left(\frac{t-a}{b} \right) \right) \right]^{-1}$	IDHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$\begin{aligned} E[T] &= \int_0^\infty t f(t) dt \\ &= \int_0^\infty S(t) dt \end{aligned}$	작성중	
평균잔여수명함수	$m(t) = E[T-t T>t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	존재하지 않음	
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \in \mathbf{R}$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 9: Cauchy 분포함수에 기반한 척도 함수

16.1.6.1 Half-Cauchy Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$2 \left[b\pi \left(1 + \left(\frac{t-a}{b} \right)^2 \right) \right]^{-1}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - \frac{2}{\pi} \arctan \left(\frac{t-a}{b} \right)$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$2 \left[b \left(1 + \left(\frac{t-a}{b} \right)^2 \right) (\pi - 2 \arctan \left(\frac{t-a}{b} \right)) \right]^{-1}$	IDHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T>t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	존재하지 않음	
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 10: Half-Cauchy 분포함수에 기반한 척도 함수

16.1.7 Chi(χ) Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$t^{\nu-1}e^{-\frac{t^2}{2}} [2^{\frac{\nu}{2}-1}\Gamma(\frac{\nu}{2})]^{-1}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - \frac{\Gamma(\frac{\nu}{2}, \frac{t^2}{2})}{\Gamma(\frac{\nu}{2})}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{t^{\nu-1}e^{-\frac{t^2}{2}}}{2^{\frac{\nu}{2}-1} [\Gamma(\frac{\nu}{2}) - \Gamma(\frac{\nu}{2}, \frac{t^2}{2})]}$	DIHR for $0 < \nu < 1$ IHR for $\nu \geq 1$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty tf(t)dt$ $= \int_0^\infty S(t)dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u)du}{S(t)}$	No closed form	IDMRL for $0 < \nu < 1$ DMRL for $\nu \geq 1$
k 차 적률	$E(T^k)$	작성중	
$100p$ 백분위수	t_p	작성중	
변수		$t \geq 0$	
파라메터		$\nu > 0$	

Table 11: χ 분포함수에 기반한 척도 함수

파라메터 $\nu = 2$ 이면, χ 분포가 Rayleigh 분포와 같아진다.

16.1.8 Chi-square(χ^2) Distribution(카이제곱 분포)

Table 12: 카이제곱분포함수에 기반한 척도 함수

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$t^{(\frac{\nu}{2}-1)} e^{-\frac{t}{2}} [2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})]^{-1}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	$1 - \frac{\Gamma(\frac{\nu}{2}, \frac{t}{2})}{\Gamma(\frac{\nu}{2})}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{t^{\frac{\nu}{2}-1} e^{-\frac{t}{2}}}{2^{\frac{\nu}{2}} [\Gamma(\frac{\nu}{2}) - \Gamma(\frac{\nu}{2}, \frac{t}{2})]}$	DHR for $0 < \nu < 2$ 0.5 for $\nu = 2$ IHR for $\nu > 2$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$	ν	
평균잔여수명함수	$m(t) = E[T - t T > t] = \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	IMRL for $0 < \nu < 2$ 2 for $\nu = 2$ DMRL for $\nu > 2$
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq 0$	
파라메터		$\nu > 0$	

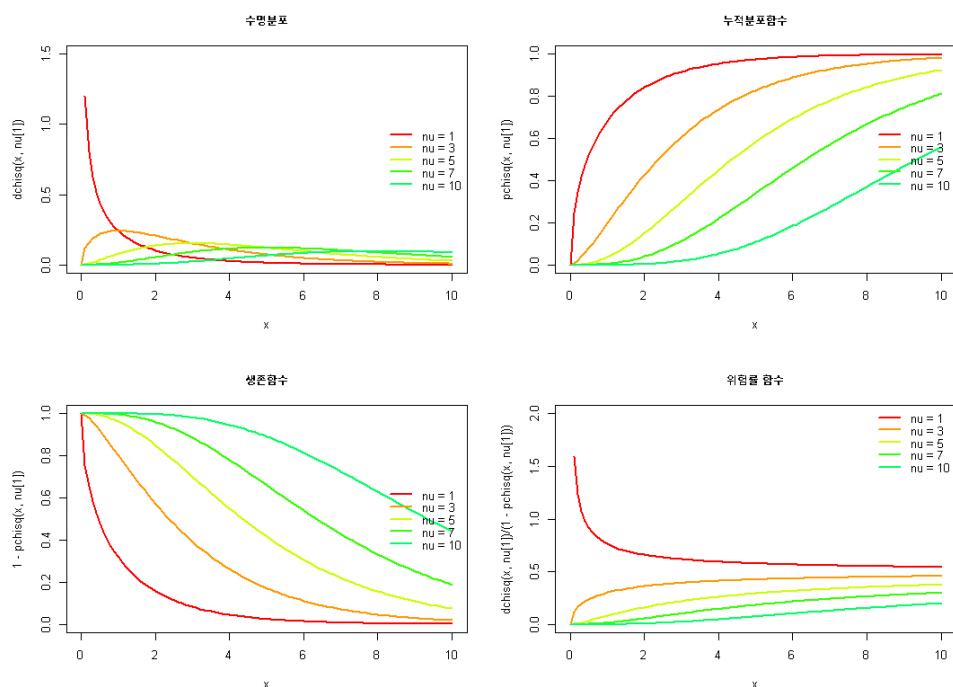


Figure 17: 카이제곱분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 2. 카이제곱분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```
1 ##### Chi-square Distribution
2 ### parameter
3 nu = c(1, 3, 5, 7, 10) # nu
4
5 ### Input Variable
6 x <- seq(0, 10, length.out = 101)
7
8
9 color = rainbow(10)
10 par(mfrow = c(2, 2))
11
12 ### Life Distribution
13 plot(x, dchisq(x, nu[1]), xlim=c(0, 10), ylim=c(0, 1.5), col=color[1], lwd=2, type = 'l', main="Life Distribution")
14 for (i in 2:5) { lines(x, dchisq(x, nu[i]), col=color[i], lwd=2); }
15 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('nu = 1', 'nu = 3', 'nu = 5', 'nu = 7', 'nu = 10'))
16
17 ### Cumulative Distribution
18 plot(x, pchisq(x, nu[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative Distribution")
19 for (i in 2:5) { lines(x, pchisq(x, nu[i]), col=color[i], lwd=2); }
20 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('nu = 1', 'nu = 3', 'nu = 5', 'nu = 7', 'nu = 10'))
21
22 ### Survival Function
23 plot(x, 1-pchisq(x, nu[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
24 for (i in 2:5) { lines(x, 1-pchisq(x, nu[i]), col=color[i], lwd=2); }
25 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('nu = 1', 'nu = 3', 'nu = 5', 'nu = 7', 'nu = 10'))
26
27 ### Hazard Function
28 plot(x, dchisq(x, nu[1])/(1-pchisq(x, nu[1])), xlim=c(0, 10), ylim=c(0, 2), col=color[1], lwd=2, type = 'l', main="Hazard Function")
29 for (i in 2:5) { lines(x, dchisq(x, nu[i])/(1-pchisq(x, nu[i])), col=color[i], lwd=2); }
30 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('nu = 1', 'nu = 3', 'nu = 5', 'nu = 7', 'nu = 10'))
```

Y_1, Y_2, \dots, Y_ν 를 표준정규분포를 따르는 서로 독립인 확률변수라고 하자. 그러면 $X = Y_1^2 + Y_2^2 + \dots + Y_\nu^2$ 는 자유도가 ν 인 카이제곱분포를 따른다. 또한, 카이제곱분포는 감마 분포의 특별한 경우로, 척도 모수 $\theta = 2$ 이고 형상 모수가 β 인 카이제곱분포가 된다.

카이제곱분포의 성질은 다음과 같다.

- 카이제곱분포를 따르는 2개 또는 그 이상의 확률변수들의 합은 카이제곱분포를 따른다. 그 때 자유도는 각각 카이제곱분포의 자유도의 합과 같다.

- $T_1 \sim \chi^2(\nu_1)$, $T_2 \sim \chi^2(\nu_2)$ 이며, 서로 독립인 확률변수라고 하면, $\frac{T_1}{\nu_1} \sim F(\nu_1, \nu_2)$ 가 된다.

- T_1, T_2 가 각각 카이제곱분포를 따르며, 서로 독립이라고 하자. 그러면 $\frac{T_1}{T_1+T_2}$ 는 베타분포를 따른다.

- 자유도 ν 의 값이 커지면 커질수록, 카이제곱분포는 평균이 ν , 분산이 2ν 인 정규분포에 근접한다.

16.1.9 Cosine Distribution

16.1.9.1 Ordinary Cosine Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{2b} \cos\left(\frac{t-a}{b}\right)$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	$\frac{1}{2} [1 - \sin\left(\frac{t-a}{b}\right)]$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$[b \left(\sec\left(\frac{t-a}{b}\right) - \tan\left(\frac{t-a}{b}\right) \right)]^{-1}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T>t] = \frac{\int_t^\infty S(u) du}{S(t)}$	$\frac{a + \frac{b\pi}{2} - t - b \cos\left(\frac{t-a}{b}\right)}{1 + \sin\left(\frac{a-t}{b}\right)}$	DMRL
k 차 적률	$E(T^k)$	작성중	
100 p 백분위수	t_p	작성중	
변수		$a - \frac{2\pi}{2} \leq t \leq a + \frac{b\pi}{2}$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 13: Ordinary Cosine 분포함수에 기반한 척도 함수

16.1.9.2 Raised Cosine Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{2b} [1 + \cos(\pi \frac{t-a}{b})]$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\frac{1}{2} [1 - \frac{t-a}{b} - \frac{1}{\pi} \sin(\pi \frac{t-a}{b})]$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{1+\cos(\pi \frac{t-a}{b})}{b[1-\frac{t-a}{b}-\frac{1}{\pi}\sin(\pi \frac{t-a}{b})]}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	$b^{0.199339 - \frac{t-a}{2b} + (\frac{t-a}{b})^2 - 0.0506606 \cos(\pi \frac{t-a}{b})}$ $\frac{1}{2} [1 - \frac{t-a}{b} - \frac{1}{\pi} \sin(\pi \frac{t-a}{b})]$	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$a - b \leq t \leq a + b$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 14: Raised Cosine 분포함수에 기반한 척도 함수

16.1.10 Dhillon's Distribution(딜론 분포)

Dhillon distribution이라는 명칭이 통용되는 것은 아니지만, Dhillon(1979, 1981)[10]이 만든 분포에 이름이 붙여지지 않았기 때문에 딜론 분포(Dhillon's distribution)라고 부르기로 한다.

$$h(t) = k\lambda ct^{c-1} + (1-k)\beta t^{\beta-1}be^{bt^\beta}$$

○ 때, $0 \leq k \leq 1, \lambda > 0, b > 0, c > 0, \beta > 0$

이 수명분포는 5개의 모수를 가지고 있고, 모수들의 값에 따라 증가, 감소 및 육조 모양의 위험률 함수(고장률 함수)를 가지는 분포이다.

그리고 파라메터의 값에 따라 다음과 같은 관계를 갖기도 한다.

- $c = \beta = 1$: Gompertz-Makeham 분포
- $k = 0, \beta = 1$: 극치 분포(extreme value distribution)
- $k = 1$: Weibull 분포
- $\beta = 0.5$: 육조 모양의 위험률 함수(고장률 함수)

그러나 이 형태가 너무 복잡했기 때문에, Dhillon(1981)[11]을 통해 Dhillon I 분포¹⁹와 Dhillon II 분포²⁰로 조금 더 간소화, 구체화하였다.

¹⁹Chapter 16.1.10.1 참고
²⁰Chapter 16.1.10.2 참고

16.1.10.1 Dhillon's I Distribuition

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{c}{b} \left(\frac{t-a}{b}\right)^{c-1} \exp\left[1 - e^{\left(\frac{t-a}{b}\right)^c} + \left(\frac{t-a}{b}\right)^c\right]$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\exp\left[1 - e^{\left(\frac{t-a}{b}\right)^c}\right]$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{c}{b} \left(\frac{t-a}{b}\right)^{c-1} e^{\left(\frac{t-a}{b}\right)^c}$	DIHR for $0 < c < 1$ IHR for $c \geq 1$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	IDMRL for $0 < c < 1$ DMRL for $c \geq 1$
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라미터		$a \in \mathbf{R}, b > 0, c > 0$	

Table 15: Dhillon's I 분포함수에 기반한 척도 함수

16.1.10.2 Dhillon's II Distribuition

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{c+1}{t-a+b} \left[\ln \left(\frac{t-a}{b} + 1 \right) \right]^c e^{-\left[\ln \left(\frac{t-a}{b} + 1 \right) \right]^{c+1}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{-\left[\ln \left(\frac{t-a}{b} + 1 \right) \right]^{c+1}}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{c+1}{t-a+b} \left[\ln \left(\frac{t-a}{b} + 1 \right) \right]^c$	DHR for $c = 0$ IDHR for $c > 0$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	IMRL for $c = 0$ DIMRL for $0 < c \lesssim 2$ DMRL for $c \gtrsim 2$
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b \geq 0, c \geq 0$	

Table 16: Dhillon's II 분포함수에 기반한 척도 함수

16.1.11 Exponential Distribution(지수 분포)

적도 함수	기호	내용
수명분포함수	$f(t)$	$\lambda e^{-\lambda t}$
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	$1 - e^{-\lambda t}$
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{-\lambda t}$
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	λ (t 에 무관하게 상수임)
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	λt
평균수명	$E[T] = \int_0^{\infty} tf(t)dt$ $= \int_0^{\infty} S(t)dt$	$\frac{1}{\lambda}$
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^{\infty} S(u)du}{S(t)}$	$\frac{1}{\lambda}$
k 차 적률	$E(T^k)$	$\frac{\Gamma(k+1)}{\lambda^k}$
100p 백분위수	t_p	$-\frac{1}{\lambda} \ln(1 - p)$
변수	$t \geq 0$	
파라메터		$\lambda = \frac{1}{\beta} > 0$

Table 17: 지수분포함수에 기반한 적도 함수

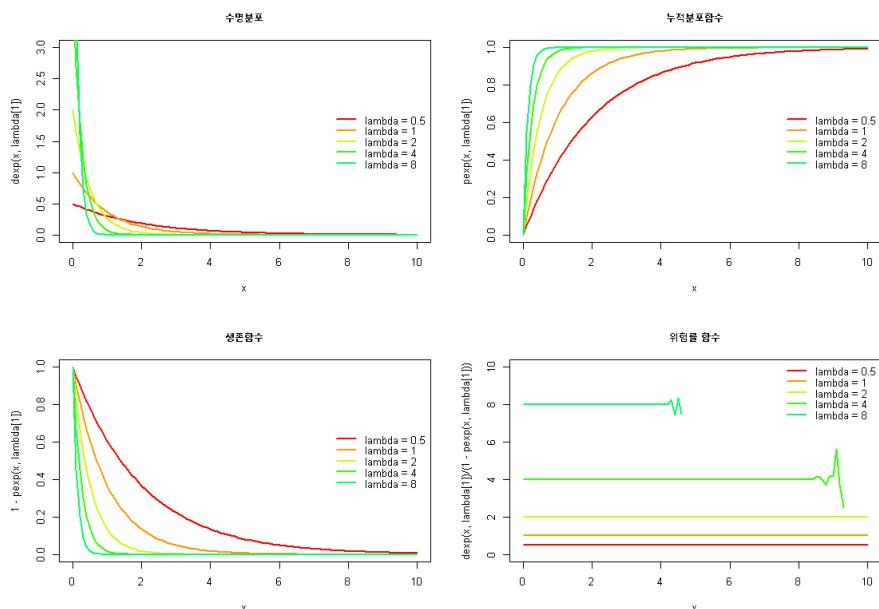


Figure 18: 지수분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 3. 지수분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```
1 ##### Exponential Distribution
2 ### parameter
3 lambda = c(0.5, 1, 2, 4, 8) # lambda
4
5 ### Input Variable
6 x <- seq(0, 10, length.out = 101)
7
8
9 color = rainbow(10)
10 par(mfrow = c(2, 2))
11
12 ### Life Distribution
13 plot(x, dexp(x, lambda[1]), xlim=c(0, 10), ylim=c(0, 3), col=color[1], lwd=2, type = 'l', main="Life Distribution")
14 for (i in 2:5) { lines(x, dexp(x, lambda[i]), col=color[i], lwd=2); }
15 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('lambda = 0.5', 'lambda = 1', 'lambda = 2', 'lambda = 4', 'lambda = 8'))
16
17 ### Cumulative Distribution
18 plot(x, pexp(x, lambda[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative Distribution")
19 for (i in 2:5) { lines(x, pexp(x, lambda[i]), col=color[i], lwd=2); }
20 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('lambda = 0.5', 'lambda = 1', 'lambda = 2', 'lambda = 4', 'lambda = 8'))
21
22 ### Survival Function
23 plot(x, 1-pexp(x, lambda[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
24 for (i in 2:5) { lines(x, 1-pexp(x, lambda[i]), col=color[i], lwd=2); }
25 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('lambda = 0.5', 'lambda = 1', 'lambda = 2', 'lambda = 4', 'lambda = 8'))
26
27 ### Hazard Function
28 plot(x, dexp(x, lambda[1])/(1-pexp(x, lambda[1])), xlim=c(0, 10), ylim=c(0, 10), col=color[1], lwd=2, type = 'l', main="Hazard Function")
29 for (i in 2:5) { lines(x, dexp(x, lambda[i])/(1-pexp(x, lambda[i])), col=color[i], lwd=2); }
30 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('lambda = 0.5', 'lambda = 1', 'lambda = 2', 'lambda = 4', 'lambda = 8'))
```

어느 부품의 수명에 대한 위험 함수(고장률 함수)가 상수인 경우를 고려하자. 즉, $h(t) = \lambda$, ($t \geq 0$)이면, 역공식에 의하여 다음과 같은 생존함수 $S(t)$ 를 가진다.

$$S(t) = e^{-\lambda t}$$

위험 함수가 상수인 수명분포는 지수분포가 유일하다. 이러한 이유로 하여금, 지수분포는 비모수 검정에서도 매우 중요한 역할을 한다. 특히 지수분포는 비교적 간단한 형태의 확률밀도함수를 가지고 있으므로, 모수들의 통계적 추론에 필요한 이론적인 결과들이 많이 증명되어 있어, 다른 수명분포에 비하여 광범위한 응용 범위를 가지고 있다. 또한 포아송 과정의 논의에도 필수적인 분포이다.

Note 12. 지수분포에 기반한 k 차 적률

$$\begin{aligned} E(T^k) &= \int_0^\infty t^k f(t) dt \\ &= \int_0^\infty t^k \lambda e^{-\lambda t} dt \\ &= \frac{1}{\lambda^k} \int_0^\infty x^k e^{-x} dx \quad (\text{substituting } x = \lambda t) \\ &= \frac{1}{\lambda^k} \Gamma(k+1) \end{aligned}$$

지수분포의 기본적인 성질은 다음과 같다.

- 만일 $T \sim \exp(\lambda)$ 이면, $\lambda T \sim \exp(1)$ 이 된다.

Proof: λT 의 생존함수를 $S(t)$ 라고 하면,

$$S(t) = P(\lambda T > t) = P\left(T > \frac{t}{\lambda}\right) = e^{-\lambda \frac{t}{\lambda}} = e^{-t}$$

- 만일 확률변수 $T \sim \text{Unif}(0, 1)$ 이라면, $-\ln T \sim \exp(1)$, $-\ln(1 - T) \sim \exp(1)$ 가 된다.

Proof: $-\ln T$ 의 생존함수를 $S(t)$ 라고 하면,

$$S(t) = P(-\ln T > t) = P(T < e^{-t}) = e^{-t}$$

- 만일 T 가 비음(non-negative)의 연속확률변수이면, $H(T) \sim \exp(1)$ 이 된다.

Proof: $H(T) = \int_0^T h(u)du = \int_0^T \frac{f(u)}{S(u)}du = \int_0^T -\ln S(u)du = -\ln S(T)$ 가 된다.

T 가 비음(non-negative)의 연속확률변수이면, $S(t) \sim \text{Unif}(0, 1)$ 이므로, 위 2번 성질에 의해 $H(T) \sim \exp(1)$ 이 성립 한다.

- (무기억성) 지수분포는 무기억성을 가지는 유일한 수명분포이다.

$$P(T > s + t | T > s) = \frac{P(T > s + t)}{P(T > s)} = \frac{S(s + t)}{S(s)} = \frac{e^{-\lambda(s+t)}}{e^{\lambda s}} = e^{\lambda t} = P(T > t)$$

- 만일 T_1, T_2, \dots, T_n 서로 독립이고, 각각 모수가 $\lambda_1, \lambda_2, \dots, \lambda_n$ 인 지수분포를 가진다면,

$T = \min(T_1, \dots, T_n) \sim \exp(\sum_{i=1}^n \lambda_i)$ 이다.

Proof: T 의 생존함수를 $S(t)$ 라고 하면

$$\begin{aligned} S(t) &= P[\min(T_1, T_2, \dots, T_n) > t] \\ &= P[T_1 > t, T_2 > t, \dots, T_n > t] \\ &= P(T_1 > t)P(T_2 > t) \cdots P(T_n > t) \\ &= e^{-\lambda_1 t}e^{-\lambda_2 t} \cdots e^{-\lambda_n t} \\ &= e^{-\sum_{i=1}^n \lambda_i t} \end{aligned}$$

- $T_i \sim \exp(\lambda)$ ($i = 1, 2, \dots, n$)이고, 서로 독립이라고 하자. $T_{(1)}, T_{(2)}, \dots, T_{(n)}$ 을 T_1, T_2, \dots, T_n 의 순서통계량이라고 하면, $D_k = T_{(k)} - T_{(k-1)}$, ($k = 1, 2, \dots, n$), $T_{(0)} = 0$ 는 k 번째 간격(spacing)으로 정의된다. 그러면

(a) D_1, D_2, \dots, D_n 은 서로 독립이다.

(b) $P(D_k \leq k) = 1 - e^{-(n-k+1)\lambda t}$, ($k = 1, 2, \dots, n$)

(c) $nD_1, (n-1)D_2, \dots, D_n$ 은 표준화 간격(normalized spacing)이라고 부르며, 서로 독립이고, 각각 $\exp(\lambda)$ 인 동일한 분포를 가진다.

- $T_i \sim \exp(\lambda)$ ($i = 1, 2, \dots, n$)이고, 서로 독립이라고 하자. $T_{(r)}$ 을 r 번째 순서통계량이라고 하면,

$$E(T_{(r)}) = \sum_{k=1}^r \frac{1}{(n-k+1)\lambda}$$

$$Var(T_{(r)}) = \sum_{k=1}^r \frac{1}{[(n-k+1)\lambda]^2}$$

Proof: $T_i \sim D_1 + D_2 + \cdots + D_r$ 이거나, 위 5번 성질로부터 $E(D_k) = \frac{1}{n-k+1}$, $Var(D_k) = \frac{1}{(n-k+1)^2}$ 이다.

- $T_i \sim \exp(\lambda)$ ($i = 1, 2, \dots, n$)이고, 서로 독립이라고 하자. 그러면 $2\lambda \sum_{i=1}^n T_i$ 는 자유도가 $2n$ 인 카이제곱 분포를 따른다.

9. Poisson Process와의 관계

X 가 주어진 사건이 단위시간 동안, 또는 단위 면적에서 발생하는 사건의 개수를 나타내는 확률변수라고 하고, 평균 λ 개의 사건이 발생한다고 하자. X 의 확률밀도함수가 다음 $f(x)$ 와 같으면, X 는 모수가 λ 인 포아송 분포를 따른다고 한다.

$$f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, \dots$$

(a) $X \sim Poi(\lambda)$ 이면, $E(X) = Var(X) = \lambda$ 이다.

(b) 주어진 사건이 단위시간당 평균 λ 개 발생하는 포아송 분포를 따른다고 하고, 구간 $(0, t]$ ($t > 0$) 동안에 발생하는 사건의 개수를 Y 라고 하자. 그러면 $Y \sim Poi(\lambda t)$ 가 성립한다.

$k = 1, 2, \dots$ 에 대하여, D_k 를 $(k-1)$ 번째와 k 번째 사건의 간격(spacing)을 나타낸다고 하면, D_1, D_2, \dots 는 서로 독립이고, 각각 위험률(고장률)이 λ 인 지수분포를 따른다. $D_k \sim exp(\lambda)$

Proof: 먼저 첫 번째 사건이 발생할 때까지 걸리는 시간인 D_1 의 확률분포를 구해 보자.

D_1 의 생존함수를 $S_1(t)$ 라고 하면, 다음과 같다.

$$\begin{aligned} S_1(t) &= P(D_1 > t) \\ &= P(\text{구간 } (0, t] \text{ 동안에 사건이 0번 발생}) \\ &= \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t} \sim exp(\lambda) \end{aligned}$$

D_2 의 생존함수를 $S_2(t)$ 로 표시하면, 다음과 같다.

$$\begin{aligned} S_2(t) &= P(D_2 > t | D_1 = d_1) \\ &= P(\text{구간 } (d_1, d_1 + t] \text{ 동안에 사건이 0번 발생} | D_1 = d_1) \\ &= P(\text{구간 } (0, t] \text{ 동안에 사건이 0번 발생}) \\ &= e^{-\lambda t} \sim exp(\lambda) \end{aligned}$$

마찬가지 방법으로, 모든 k 에 대하여 $D_k \sim exp(\lambda)$ 이 성립되고, 독립성은 포아송분포의 가정으로부터 성립한다.

(c) T_K 가 $(k-1)$ 번째와 k 번째 사건의 간격(spacing)을 나타내고, T_1, T_2, \dots 가 서로 독립이며, 각각 위험률(고장률)이 λ 인 지수분포를 따른다고 하자. 그러면 구간 $(0, t]$ 동안에 발생하는 사건의 개수는, 모수 λt 를 가지는 포아송 분포를 따른다.

Proof: $T = T_1 + T_2 + \dots + T_n$ ²¹은, n 번째 사건이 발생할 때까지 걸리는 시간을 나타내며, T 가 다음과 같은 확률밀도함수 $f_T(t)$ 와 생존함수 $S_T(t)$ 를 가진다.

$$f_T(t) = \frac{(\lambda t)^{n-1}}{(n-1)!} \lambda e^{-\lambda t} \quad t \geq 0$$

$$S_T(t) = \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad t \geq 0$$

구간 $(0, T]$ 동안에 발생하는 사건의 개수를 확률변수 N 으로 표시하면,

$$\begin{aligned} P(N = n) &= P[T_1 + T_2 + \dots + T_n \leq t < T_1 + T_2 + \dots + T_n + T_{n+1}] \\ &= P[T_1 + T_2 + \dots + T_n + T_{n+1} > t] - P[T_1 + T_2 + \dots + T_n > t] \\ &= \sum_{k=0}^n \frac{(\lambda t)^k}{k!} e^{-\lambda t} - \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t} \\ &= \frac{(\lambda t)^n}{n!} e^{-\lambda t} \sim Poi(\lambda t) \quad n = 0, 1, 2, \dots \end{aligned}$$

²¹ $\circ| T$ 는 $T \sim Gamma(n, \lambda)$ 를 따른다.

Note 13. 지수분포를 따르는 완전자료에 대한 모수 추정

지수분포는 한 개의 모수를 가지는 수명분포로, 위험률(고장률) λ 또는 평균 수명 $E(T) = \theta = \frac{1}{\lambda}$ 로 표시되는 모수의 값이 주어지면, 지수분포에 대한 모든 특성이 결정된다. 그러나 주어진 수명이 지수분포를 따른다는 것이 알려진 때에도 모수의 값을 모르는 경우가 대부분이며, 따라서 관측된 자료를 이용하여 모수의 값을 추정하여야 한다.

지수분포의 확률밀도함수는

$$f(t|\lambda) = \lambda e^{-\lambda t}, \quad (t \geq 0)$$

또는

$$f(t|\theta) = \frac{1}{\theta} e^{-\frac{t}{\theta}}, \quad (t \geq 0)$$

따라서, 평균수명을 중시할 때에는, θ 를 모수로 하고, 위험률(고장률)을 중시할 때에는 λ 를 모수로 하면 된다.

$T \sim exp(\lambda)$ 를 가정하고, n 개의 부품을 가지고 시점 $t = 0$ 에서 시험을 시작하여, 부품들의 고장 시간이 모두 관측되었다고 가정하자. 이 값들을 t_1, t_2, \dots, t_n 으로 표시하자. 이 경우 얻어진 고장자료는 완전 자료이다.

- 점추정

λ 의 최대 가능도 함수는

$$L(\lambda) = \lambda e^{-t_1} \lambda e^{-t_2} \cdots \lambda e^{-t_n}$$

양변에 로그를 취하면

$$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^n t_i$$

양변을 λ 에 미분하고 0으로 놓으면,

$$\frac{\partial L(\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n t_i = 0$$

이를 λ 에 대해서 풀면, λ 의 최대가능도추정량(MLE)은 다음과 같다.

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n t_i}$$

이 $\hat{\lambda}$ 는 λ 에 대한 unbiased estimator이며, MLE의 성질에 의해, θ 의 MLE 또한 $\hat{\theta} = \frac{1}{\hat{\lambda}} = \frac{\sum_{i=1}^n t_i}{n}$ 이 된다.

- 구간 추정

λ 에 대한 $100(1 - \alpha)\%$ 신뢰구간을 유도하기 위해서는, 먼저 점 추정량 $\hat{\lambda}$ 의 표본 분포를 구하여야 한다.

$\hat{\lambda}$ 를 확률변수로 간주하기 위하여, $\hat{\lambda} = \frac{n}{\sum_{i=1}^n T_i}$ 를 고려하자. 여기에서 $T_i \sim exp(\lambda)$ ($i = 1, 2, \dots, n$)이고, T_1, T_2, \dots, T_n 은 서로 독립이다. 그러면 **지수분포의 8번 성질**에 의해

$$2\lambda \sum_{i=1}^n T_i = \frac{2n\lambda}{\hat{\lambda}} \sim \chi^2(2n)$$

따라서, $\chi^2(k; p)$ 는, 자유도가 k 인 카이제곱분포의 $100(1 - p)$ 백분위수, 즉, 오른쪽 면적이 p 가 되는 점을 나타낸다고 하면, 다음과 같은 관계가 성립된다.

$$\begin{aligned} 1 - \alpha &= P \left[\chi^2 \left(2n; 1 - \frac{\alpha}{2} \right) < \frac{2n\lambda}{\hat{\lambda}} < \chi^2 \left(2n; \frac{\alpha}{2} \right) \right] \\ &= P \left[\frac{\hat{\lambda}}{2n} \chi^2 \left(2n; 1 - \frac{\alpha}{2} \right) < \lambda < \frac{\hat{\lambda}}{2n} \chi^2 \left(2n; \frac{\alpha}{2} \right) \right] \\ &= P \left[\frac{1}{2 \sum_{i=1}^n T_i} \chi^2 \left(2n; 1 - \frac{\alpha}{2} \right) < \lambda < \frac{1}{2 \sum_{i=1}^n T_i} \chi^2 \left(2n; \frac{\alpha}{2} \right) \right] \end{aligned}$$

$\lambda = \frac{1}{\theta}$ 의 관계를 이용하면, 평균수명 θ 에 대한 $100(1 - \alpha)\%$ 신뢰구간은 다음과 같다.

$$P \left[\frac{2n\hat{\theta}}{\chi^2(2n; 1 - \frac{\alpha}{2})} < \theta < \frac{2n\hat{\theta}}{\chi^2(2n; \frac{\alpha}{2})} \right]$$

대표본²²인 경우에는, 중심극한정리에 의해, θ 에 대한 대표본 $100(1 - \alpha)\%$ 신뢰구간은 다음과 같이 근사적으로 계산된다.

$$\left[\hat{\theta} - \frac{\hat{\theta}}{\sqrt{n}} z_{\frac{\alpha}{2}}, \hat{\theta} + \frac{\hat{\theta}}{\sqrt{n}} z_{\frac{\alpha}{2}} \right]$$

이는 $E[\hat{\theta}] = \theta$, $Var[\hat{\theta}] = \frac{\theta^2}{n}$ 을 이용한다.

²²대략 $n \geq 30$

Note 14. 지수분포를 따르는 Type II 우중도절단자료(정수중단자료)에 대한 모수 추정

Type II 우중도절단 자료(정수중단자료)는, n 개의 부품을 가지고 $t = 0$ 에서 수명시험을 시작하여, 고장(재발)시간이 순차적으로 기록되고, 미리 정해진 $r (\leq n)$ 개의 부품이 고장나는(환자의 병이 재발하는) 시점에서 시험을 중단하였을 때 얻어지는 자료이다. 이 때 고장(재발)이 관측된 r 개의 부품에 대한 고장(재발) 시간을 순서대로 나열하면, $t_{(1)} \leq t_{(2)} \leq \dots \leq t_{(r)}$ 이 되고, 나머지 $(n - r)$ 개의 부품(환자)는 수명이 $t_{(r)}$ 보다 크다는 것이 기록된다. 이 경우 r 의 값은 고정된 알려진 상수이나, 시험중단시점인 $t_{(r)}$ 의 값은, 부품(환자)의 수명 분포에 의존하는 확률변수이다. Type II 우중도절단 자료(정수중단자료)에 대한 평균수명 θ 에 대한 가능도함수는 다음과 같이 표시된다.

$$\begin{aligned} L(\theta) &= \frac{n!}{(n-r)!} \prod_{i=1}^r \frac{1}{\theta} e^{-\frac{t_{(i)}}{\theta}} e^{-\frac{t_{(i)}}{\theta(n-r)}} \\ &= \frac{n!}{(n-r)!} \frac{1}{\theta^r} e^{-\frac{1}{\theta} [\sum_{i=1}^r t_{(i)} + (n-r)t_{(r)}]} \end{aligned}$$

이 가능성도 함수로부터 구해진 θ 의 평균잔여수명함수는 다음과 같이 얻어진다.

$$\hat{\theta} = \frac{\sum_{i=1}^r t_{(i)} + (n-r)t_{(r)}}{r} = \frac{T_c}{r}$$

여기서, $T_c = \sum_{i=1}^r t_{(i)} + (n-r)t_{(r)}$: 수명시험에 소요된 전체 시험시간

Type II 우중도절단 자료(정수중단자료)의 경우, 평균수명 θ 에 대한 $100(1 - \alpha)\%$ 신뢰구간을 구하기 위해서는, 우선 전체 시험시간 T_c 의 표본 분포를 구해야 한다.

$$\begin{aligned} T_c &= \sum_{i=1}^r t_{(i)} + (n-r)t_{(r)} \\ &= \sum_{i=1}^{r-1} t_{(i)} + (n-r+1)t_{(r)} \\ &= \sum_{i=1}^r (n-i+1)(t_{(i)} - t_{(i-1)}) \\ &= \sum_{i=1}^r (n-i+1)D_i \end{aligned}$$

지수분포의 6번-(c) 성질에 의해 $(n-i+1)D_i \sim \exp(\theta)$ ($i = 1, 2, \dots, r$)이므로, 지수분포의 8번 성질에 의해 ($\lambda = \frac{1}{\theta}$), $2\lambda T_c = \frac{2}{\theta} T_c \sim \chi^2(2r)$ 된다. 따라서, θ 에 대한 $100(1 - \alpha)\%$ 신뢰구간은 다음과 같다.

$$\left[\frac{2T_c}{\chi^2(2r; \frac{\alpha}{2})} < \theta < \frac{2T_c}{\chi^2(2r; 1 - \frac{\alpha}{2})} \right]$$

위험률(고장률) λ 에 대한 $100(1 - \alpha)\%$ 신뢰구간은 $\lambda = \frac{1}{\theta}$ 의 관계를 이용하면, 다음과 같다.

$$\left[\frac{\chi^2(2r; \frac{\alpha}{2})}{2T_c} < \lambda < \frac{\chi^2(2r; 1 - \frac{\alpha}{2})}{2T_c} \right]$$

만일, 대표본²³이면, $\hat{\theta}$ 은 근사적으로 평균이 θ 이고, 다음과 같은 분산을 가진 정규분포를 따른다고 알려져 있다.

$$\widehat{Var}(\hat{\theta}) = \frac{\hat{\theta}^2}{\sum_{i=1}^n \left(1 - e^{-\frac{T_i}{\hat{\theta}}}\right)}$$

T_i 는 i 번째 부품(환자)에 대한 고장(재발)관측시간²⁴이다. 따라서, Type II 우중도절단 자료(정수중단자료)를 이용한 θ 의 대표본 $100(1 - \alpha)\%$ 신뢰구간은 다음과 같다.

$$\left[\hat{\theta} - z_{\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\theta})} < \theta < \hat{\theta} + z_{\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\theta})} \right]$$

²³대략 $n \geq 30$

²⁴만일, 시험중단시점까지 고장이 나지 않은 경우에는, 시험중단시간이다.

Note 15. 지수분포를 따르는 Type I 우중도절단자료(정시중단자료)에 대한 모수 추정

Type I 우중도절단 자료(정시중단자료)는 n 개의 부품(환자)을 가지고 시점 $t = 0$ 에서 시작한 수명시험을 미리 정해진 시점 t_c 에서 중단하였을 때 얻어지는 고장(재발)자료로서, 만일 $0 \leq r \leq n$ 개의 고장(재발)이 관측되었다고 가정하고 r 의 고장(재발)시간을 순서대로 나열하면 $t_{(1)} \leq \dots \leq t_{(r)}$ 과 나머지 $(n - r)$ 개의 부품(환자)은 수명이 t_c 보다 크다는 것이 기록된다.

이 경우, 시험중단시점인 t_c 의 값은 고정된 알려진 상수이나, r 은 부품(환자)의 수명 분포에 의존하는 확률변수이다. R 이 시험기간 $(0, t_c]$ 동안에 발생한 고장(재발) 개수를 나타내는 확률변수라고 하면, 정시중단자료에 대한 평균수명 θ 에 대한 가능도함수는 다음과 같이 표시된다.

$$0 < t_{(1)} \leq \dots \leq t_{(r)} < t_c \text{에 대하여}$$

$$\begin{aligned} L(\theta) &= f(t_{(1)}, \dots, t_{(r)} | R = r) f_R(r) \\ &= \left[r! \prod_{i=1}^r \frac{f(t_{(i)})}{F(t_c)} \right] \left[\frac{n!}{r!(n-r)!} \{F(t_c)\}^r \{1 - F(t_c)\}^{n-r} \right] \\ &= \frac{n!}{(n-r)!} \{1 - F(t_c)\}^{n-r} \prod_{i=1}^r f(t_{(i)}) \\ &= \frac{n!}{(n-r)!} \frac{1}{\theta^r} e^{-\frac{1}{\theta} [\sum_{i=1}^r t_{(i)} + (n-r)t_c]} \end{aligned}$$

F 는 부품(환자)의 누적분포함수(불신뢰도함수)로서, $F(t_c) = 1 - e^{-\frac{t_c}{\theta}}$ 이다. 가능도함수를 최대로 하는 θ 의 MLE는 다음과 같이 얻어진다.

$$\hat{\theta} = \frac{\sum_{i=1}^r t_{(i)} + (n-r)t_c}{r} = \frac{T^*}{r}$$

여기서, $T^* = \sum_{i=1}^r t_{(i)} = (n-r)t_c$. 전체 시험시간

$$\frac{2r\hat{\theta}}{\theta} = \frac{2T^*}{\theta} \sim \chi^2(2r)$$

따라서, Type 1 우중도절단 자료(정시중단자료)를 이용한 θ 의 $100(1 - \alpha)\%$ 신뢰구간은 다음과 같다.

$$\left[\frac{2T^*}{\chi^2(2r; \frac{\alpha}{2})} < \theta < \frac{2T^*}{\chi^2(2r; 1 - \frac{\alpha}{2})} \right]$$

16.1.11.1 Exponential Distribution with Location Parameter

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{b}e^{-\frac{t-a}{b}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{-\frac{t-a}{b}}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{1}{b}$	IHR and DHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^{\infty} tf(t)dt$ $= \int_0^{\infty} S(t)dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^{\infty} S(u)du}{S(t)}$	b	IMRL and DMRL
k 차 적률	$E(T^k)$	작성중	
100 p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 18: Exponential 분포함수에 기반한 척도 함수

16.1.11.2 Exponentiated Exponential Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{c}{b} e^{-\frac{t-a}{b}} \left[1 - e^{-\frac{t-a}{b}}\right]^{c-1}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - \left[1 - e^{-\frac{t-a}{b}}\right]^c$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	No closed form	DHR for $0 < c \leq 1$ IHR for $c \geq 1$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	IMRL for $0 < c \leq 1$ DMRL for $c \geq 1$
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 19: Exponentiated exponential 분포함수에 기반한 척도 함수

16.1.11.3 Reflected Exponential Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{b} e^{\frac{t-a}{b}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - e^{\frac{t-a}{b}}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\left[b \left(e^{\frac{t-a}{b}} - 1 \right) \right]^{-1}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq 0$	
파라메터		작성중	

Table 20: Reflected Exponential 분포함수에 기반한 척도 함수

16.1.12 Extreme Value Distribution(극치 분포)

16.1.12.1 최대극치분포(The Maximum Extreme Value Distribution)와 최소극치분포(The Minimum Extreme Value Distribution)

극치 분포는 어떤 분포로부터 무작위로 관찰된 많은 값들의 최소값이나 최대값에 대한 접근 분포(limiting distribution)이다. 수명 자료를 다루는 과정에서 종종 확률변수들의 극단적인 값(최대값이나 최소값)에 관심을 갖게 된다.

- 여러 개의 아이템이 직렬로 구성된 시스템의 수명을 평가할 때, 우리는 모든 아이템의 수명보다는 수명이 가장 짧은 아이템에 더 많은 관심을 갖게 된다. 왜냐하면 이와 같은 **직렬 구조 시스템의 수명은, 수명이 가장 짧은 아이템의 수명과 일치**하기 때문이다.
- 여러 개의 아이템이 병렬로 구성된 시스템의 수명을 평가할 때, 우리는 모든 아이템의 수명보다는 수명이 가장 긴 아이템에 더 많은 관심을 갖게 된다. 왜냐하면 이와 같은 **병렬 구조 시스템의 수명은, 수명이 가장 긴 아이템의 수명과 일치**하기 때문이다.

Note 16. 최대 극치 분포(The Maximum Extreme Value Distribution)

X_1, X_2, \dots, X_n 을 분포함수가 각각 F_1, F_2, \dots, F_n 인 확률변수라고 하자. 만일 Y 가 이 n 개의 확률변수의 최대값이라고 하면, 이것의 분포함수 $F_Y(y)$ 를 최대극치분포라고 부른다. 만일 X_i 들이 동일한 분포인 $F(y)$ 를 따르고 서로 독립이라면, 최대값에 대한 분포함수는 다음과 같이 얻을 수 있다.

$$\begin{aligned}F_Y(y) &= P[(X_1 \leq y)(X_2 \leq y) \cdots (X_n \leq y)] \\&= \prod_{i=1}^n P(X_i \leq y) \\&= [F(y)]^n\end{aligned}$$

이 분포함수의 확률밀도함수는 다음과 같이 구해진다.

$$\begin{aligned}f_Y(y) &= \frac{d}{dy} F_Y(y) \\&= n [F(y)]^{n-1} f(y)\end{aligned}$$

이 경우, 각각의 공통분포인 $F(y)$ 를, 최대극치분포인 $F_Y(y)$ 의 모분포(parent distribution)라고 부른다.

Note 17. 최소 극치 분포(The Minimum Extreme Value Distribution)

X_1, X_2, \dots, X_n 을 분포함수가 각각 F_1, F_2, \dots, F_n 인 확률변수라고 하자. 만일 Z 가 이 n 개의 확률변수의 최소값이라고 하면, 이것의 분포함수 $F_Z(z)$ 를 최소극치분포라고 부른다. 만일 X_i 들이 동일한 분포인 $F(z)$ 를 따르고 서로 독립이라면, 최소값에 대한 분포함수는 다음과 같이 얻을 수 있다.

$$\begin{aligned}F_Z(z) &= P[\min(X_1, X_2, \dots, X_n) \leq z] \\&= 1 - P[\min(X_1, X_2, \dots, X_n) > z] \\&= 1 - P(Z > z) \\&= 1 - \prod_{i=1}^n [1 - F_i(z)] \\&= 1 - \prod_{i=1}^n [1 - F(z)] \\&= 1 - [1 - F(z)]^n\end{aligned}$$

$$\text{여기서, } P(Z > z) = \prod_{i=1}^n P(X_i > z): Z \text{의 생존 함수}$$

이 분포함수의 확률밀도함수는 다음과 같이 구해진다.

$$\begin{aligned}f_Z(z) &= \frac{d}{dz} F_Z(z) \\&= n [1 - F(z)]^{n-1} f(z)\end{aligned}$$

16.1.12.2 Type I 최소값 극치분포(Gumbel 최소값 분포)

Table 21: Type I 극치분포(Gumbel 최소값 분포)함수에 기반한 척도 함수

척도 함수	기호	내용
수명분포함수	$f(t)$	$\frac{1}{\beta} e^{\frac{t-\mu}{\beta}} e^{-e^{\frac{t-\mu}{\beta}}}$
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{-e^{\frac{t-\mu}{\beta}}}$
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{1}{\beta} e^{\frac{t-\mu}{\beta}}$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	$\mu - \gamma\beta$
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	작성중
k 차 적률	$E(T^k)$	작성중
100p 백분위수	t_p	작성중
변수		$t \geq 0$
파라메터	μ (위치 모수; location parameter), β (척도 모수; scale parameter)	
상수		$\gamma = 0.5772 \dots$ (오일러 상수)

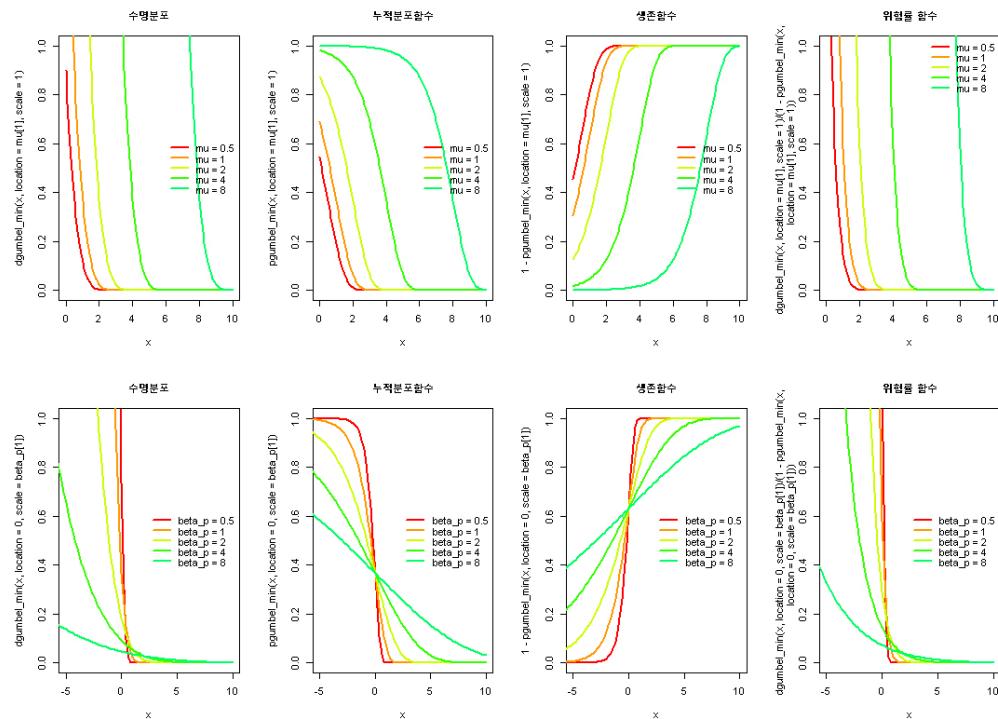


Figure 19: Type I 극치분포(Gumbel 최소값 분포)에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 4. Type I 극치분포(Gumbel 최소값 분포)에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```

1 ##### 극치 분포: Gumbel 최소값 분포
2 dgumbel_min = function (x, scale = 1, location = 0, log = FALSE)
3 {
4   fx <- 1/scale * exp(-(x - location)/scale) * exp(-exp((x - location)/scale))
5   if (log)
6     return(log(fx))
7   else return(fx)
8 }
9
10 pgumbel_min = function (q, scale = 1, location = 0, lower.tail = TRUE, log.p = FALSE)
11 {
12   Fx <- exp(-exp((q - location)/scale))
13   if (!lower.tail)
14     Fx <- 1 - Fx
15   if (log.p)
16     Fx <- log(Fx)
17   return(Fx)
18 }
19 qgumbel_min = function (p, scale = 1, location = 0, lower.tail = TRUE, log.p = FALSE)
20 {
21   if (log.p)
22     p <- exp(p)
23   if (!lower.tail)
24     p <- 1 - p
25   xF <- location - scale * log(-log(p))
26   return(xF)
27 }
28
29 rgumbel_min = function (n, scale = 1, location = 0)
30 {
31   qgumbel(runif(n), scale, location)
32 }
33
34
35 par(mfrow = c(2, 4))
36
37 ### parameter: mu
38 mu = c(0.5, 1, 2, 4, 8) # shape
39
40 ### Input Variable
41 x <- seq(0, 10, length.out = 101)
42
43 color = rainbow(10)
44
45 ### Life Distribution
46 plot(x, dgumbel_min(x, location=mu[1], scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Life
47   Distribution")
48 for (i in 2:5) { lines(x, dgumbel_min(x, location=mu[i], scale=1), col=color[i], lwd=2); }
49 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0.5', 'mu = 1', 'mu = 2', 'mu = 4', 'mu = 8'))
50
51 ### Cumulative Distribution
52 plot(x, pgumbel_min(x, location=mu[1], scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative
53   Distribution")
54 for (i in 2:5) { lines(x, pgumbel_min(x, location=mu[i], scale=1), col=color[i], lwd=2); }
55 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0.5', 'mu = 1', 'mu = 2', 'mu = 4', 'mu = 8'))
56
57 ### Survival Function
58 plot(x, 1-pgumbel_min(x, location=mu[1], scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival
59   Function")
60 for (i in 2:5) { lines(x, 1-pgumbel_min(x, location=mu[i], scale=1), col=color[i], lwd=2); }
61 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0.5', 'mu = 1', 'mu = 2', 'mu = 4', 'mu = 8'))
62
63 ### Hazard Function
64 plot(x, dgumbel_min(x, location=mu[1], scale=1)/(1-pgumbel_min(x, location=mu[1], scale=1)), xlim=c(0, 10), ylim=c(0, 1), col=
65   color[1], lwd=2, type = 'l', main="Hazard Function")
66 for (i in 2:5) { lines(x, dgumbel_min(x, location=mu[i], scale=1)/(1-pgumbel_min(x, location=mu[i], scale=1)), col=color[i], lwd
67   =2); }
68 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0.5', 'mu = 1', 'mu = 2', 'mu = 4', 'mu = 8'))
69
70
71
72
73
74
75
76
77
78
79

```

```

80  ## Cumulative Distribution
81  plot(x, pgumbel_min(x, location=0, scale=beta_p[1]), xlim=c(-5, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative Distribution")
82  for (i in 2:5) { lines(x, pgumbel_min(x, location=0, scale=beta_p[i]), col=color[i], lwd=2); }
83  legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8'))
84
85  ### Survival Function
86  plot(x, 1-pgumbel_min(x, location=0, scale=beta_p[1]), xlim=c(-5, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
87  for (i in 2:5) { lines(x, 1-pgumbel_min(x, location=0, scale=beta_p[i]), col=color[i], lwd=2); }
88  legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8'))
89
90  ### Hazard Function
91  plot(x, dgumbel_min(x, location=0, scale=beta_p[1])/(1-pgumbel_min(x, location=0, scale=beta_p[1])), xlim=c(-5, 10), ylim=c(0, 1),
92  col=color[1], lwd=2, type = 'l', main="Hazard Function")
93  for (i in 2:5) { lines(x, dgumbel_min(x, location=0, scale=beta_p[i])/(1-pgumbel_min(x, location=0, scale=beta_p[i])), col=color[i],
94  lwd=2); }
93  legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8'))

```

Type 1 극치분포는 와이블분포의 대수변환분포로, 위치-척도 모수를 갖는 분포족에 속해서 많이 사용한다.
이 분포는 최소값에 근거한 분포와 최대값에 근거한 분포 두 종류가 있으며, Gumbel 분포로 알려져 있다.

- 최소값에 근거한 점근분포는, 미사일에서 부식성의 화학 물질을 보내는 관의 누출시간을 모형화하거나, 기계 고장 문제를 모형화하는 데 적용된다.
- 최대값에 대한 점근분포는 최대 방류량이나 연중 최고 조류 등을 모형화하는 데 적합하다.

16.1.12.3 Type I 최대값 극치분포(Gumbel 최대값 분포)

Table 22: Type I 극치분포(Gumbel 최대값 분포)함수에 기반한 척도 함수

척도 함수	기호	내용
수명분포함수	$f(t)$	$\frac{1}{\beta} e^{-\frac{t-\mu}{\beta}} e^{-e^{-\frac{t-\mu}{\beta}}}$
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - e^{-e^{-\frac{t-\mu}{\beta}}}$
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{\frac{1}{\beta} e^{-\frac{t-\mu}{\beta}}}{e^{e^{-\frac{t-\mu}{\beta}}} - 1}$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중
평균수명	$E[T] = \int_0^{\infty} tf(t)dt$ $= \int_0^{\infty} S(t)dt$	$\mu + \gamma\beta$
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^{\infty} S(u)du}{S(t)}$	작성중
k 차 적률	$E(T^k)$	작성중
100p 백분위수 변수	t_p	작성중
파라메터	μ (위치 모수; location parameter), β (척도 모수; scale parameter)	
상수	$\gamma = 0.5772 \dots$ (오일러 상수)	

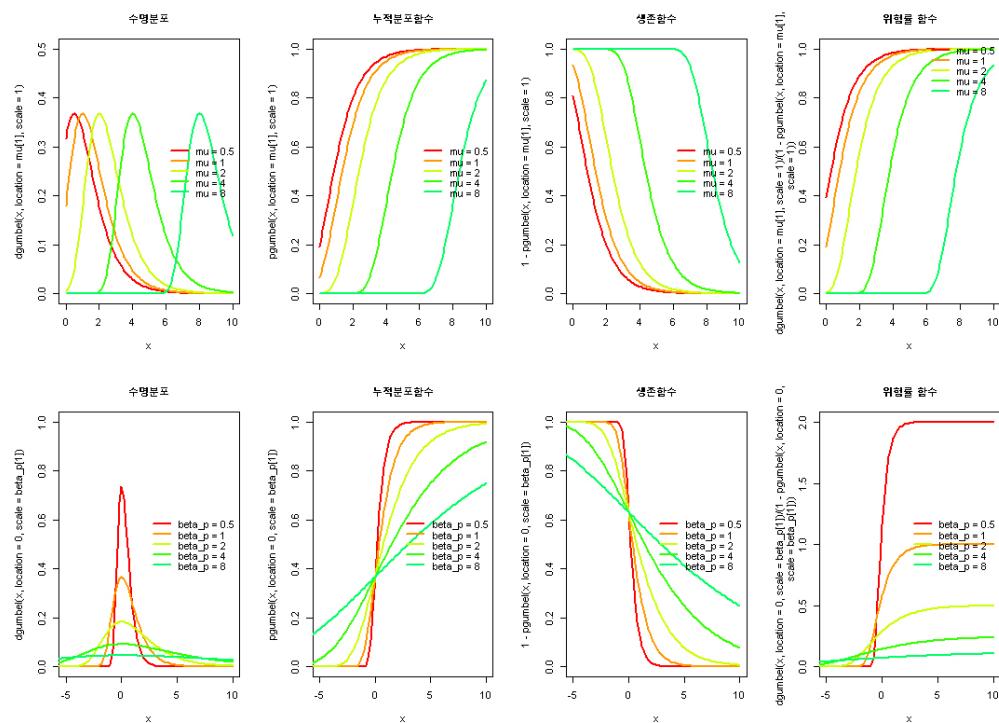


Figure 20: Type I 극치분포(Gumbel 최대값 분포)에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 5. Type I 극치분포(Gumbel 최대값 분포)에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```
1 ##### 극치 분포: Gumbel 최대값 분포
2 dgumbel_max = function (x, scale = 1, location = 0, log = FALSE)
3 {
4   fx <- 1/scale * exp(-(x - location)/scale - exp(-(x - location)/scale))
5   if (log)
6     return(log(fx))
7   else return(fx)
8 }
9
10 pgumbel_max = function (q, scale = 1, location = 0, lower.tail = TRUE, log.p = FALSE)
11 {
12   Fx <- exp(-exp(-(q - location)/scale))
13   if (!lower.tail)
14     Fx <- 1 - Fx
15   if (log.p)
16     Fx <- log(Fx)
17   return(Fx)
18 }
19 qgumbel_max = function (p, scale = 1, location = 0, lower.tail = TRUE, log.p = FALSE)
20 {
21   if (log.p)
22     p <- exp(p)
23   if (!lower.tail)
24     p <- 1 - p
25   xF <- location - scale * log(-log(p))
26   return(xF)
27 }
28
29 rgumbel_max = function (n, scale = 1, location = 0)
30 {
31   qgumbel(runif(n), scale, location)
32 }
33
34
35 par(mfrow = c(2, 4))
36
37 ### parameter: mu
38 mu = c(0.5, 1, 2, 4, 8) # shape
39
40 ### Input Variable
41 x <- seq(0, 10, length.out = 101)
42
43 color = rainbow(10)
44
45 ### Life Distribution
46 plot(x, dgumbel_max(x, location=mu[1], scale=1), xlim=c(0, 10), ylim=c(0, 0.5), col=color[1], lwd=2, type = 'l', main="Life
47   Distribution")
48 for (i in 2:5) { lines(x, dgumbel_max(x, location=mu[i], scale=1), col=color[i], lwd=2); }
49 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0.5', 'mu = 1', 'mu = 2', 'mu = 4', 'mu = 8'))
50
51 ### Cumulative Distribution
52 plot(x, pgumbel_max(x, location=mu[1], scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative
53   Distribution")
54 for (i in 2:5) { lines(x, pgumbel_max(x, location=mu[i], scale=1), col=color[i], lwd=2); }
55 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0.5', 'mu = 1', 'mu = 2', 'mu = 4', 'mu = 8'))
56
57 ### Survival Function
58 plot(x, 1-pgumbel_max(x, location=mu[1], scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival
59   Function")
60 for (i in 2:5) { lines(x, 1-pgumbel_max(x, location=mu[i], scale=1), col=color[i], lwd=2); }
61 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0.5', 'mu = 1', 'mu = 2', 'mu = 4', 'mu = 8'))
62
63 ### Hazard Function
64 plot(x, dgumbel_max(x, location=mu[1], scale=1)/(1-pgumbel_max(x, location=mu[1], scale=1)), xlim=c(0, 10), ylim=c(0, 1), col=
65   color[1], lwd=2, type = 'l', main="Hazard Function")
66 for (i in 2:5) { lines(x, dgumbel_max(x, location=mu[i], scale=1)/(1-pgumbel_max(x, location=mu[i], scale=1)), col=color[i], lwd
67   =2); }
68 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0.5', 'mu = 1', 'mu = 2', 'mu = 4', 'mu = 8'))
69
70
71
72
73
74
75
76
77
78
79
```

```

80  ## Cumulative Distribution
81  plot(x, pgumbel_max(x, location=0, scale=beta_p[1]), xlim=c(-5, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative Distribution")
82  for (i in 2:5) { lines(x, pgumbel_max(x, location=0, scale=beta_p[i]), col=color[i], lwd=2); }
83  legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8'))
84
85  ### Survival Function
86  plot(x, 1-pgumbel_max(x, location=0, scale=beta_p[1]), xlim=c(-5, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
87  for (i in 2:5) { lines(x, 1-pgumbel_max(x, location=0, scale=beta_p[i]), col=color[i], lwd=2); }
88  legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8'))
89
90  ### Hazard Function
91  plot(x, dgumbel_max(x, location=0, scale=beta_p[1])/(1-pgumbel_max(x, location=0, scale=beta_p[1])), xlim=c(-5, 10), ylim=c(0, 2),
92  col=color[1], lwd=2, type = 'l', main="Hazard Function")
93  for (i in 2:5) { lines(x, dgumbel_max(x, location=0, scale=beta_p[i])/(1-pgumbel_max(x, location=0, scale=beta_p[i])), col=color[i],
94  lwd=2); }
95  legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8'))

```

16.1.12.4 Type II 극치분포(Frechet 분포)

Table 23: Type II 극치분포(Frechet 분포)함수에 기반한 척도 함수

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{m}{\beta} \left(\frac{t-\mu}{\beta} \right)^{-(m+1)} e^{-\left(\frac{t-\mu}{\beta} \right)^{-m}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	$1 - e^{-\left(\frac{t-\mu}{\beta} \right)^{-(m+1)}}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{\frac{m}{\beta} \left(\frac{t-\mu}{\beta} \right)^{-(m+1)}}{e^{\left(\frac{t-\mu}{\beta} \right)^{-m}} - 1}$	
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t] = \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	
k 차 적률	$E(T^k)$	작성중	
100 p 백분위수	t_p	작성중	
변수		$t \geq 0$	
파라미터		$t \geq \mu, -\infty < \mu < \infty$ (위치 모수; location parameter) $m > 0$ (형상 모수; shape parameter), $\beta > 0$ (척도 모수; scale parameter)	

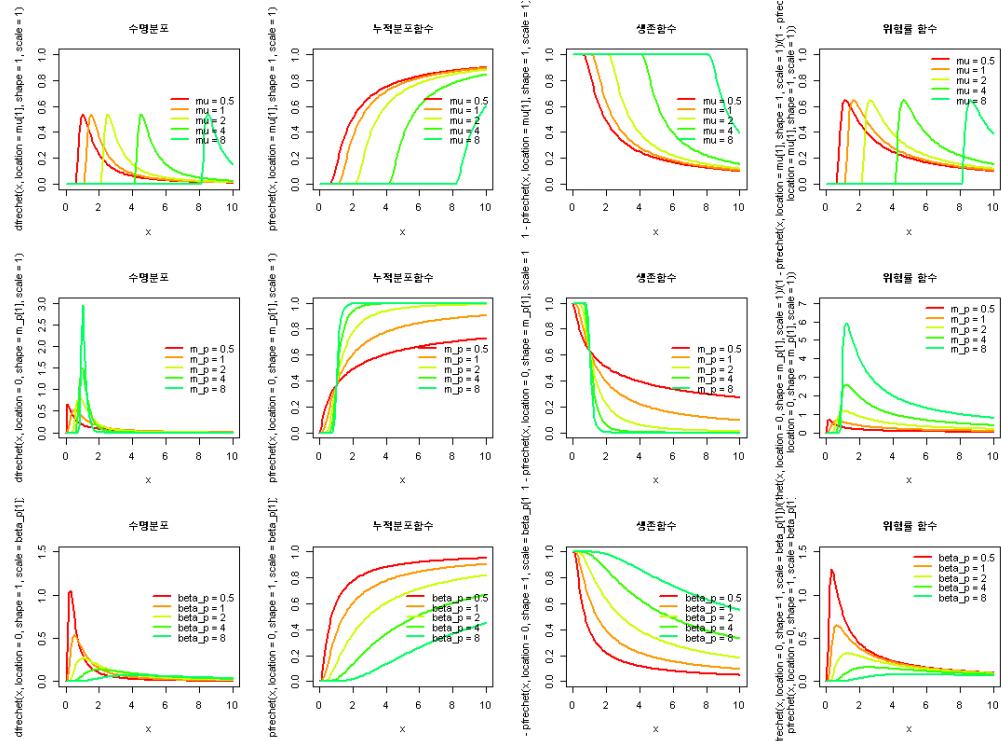


Figure 21: Type II 극치분포(Frechet 분포)에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 6. Type II 극치분포(Frechet 최대값 분포)에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```
1 ##### frechet 분포
2 require(VGAM)
3 par(mfrow = c(3, 4))
4
5 ### parameter: mu
6 mu = c(0.5, 1, 2, 4, 8) # location
7
8 ### Input Variable
9 x <- seq(0, 10, length.out = 101)
10
11 color = rainbow(10)
12
13 ### Life Distribution
14 plot(x, dfrechet(x, location=mu[1], shape=1, scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Life
   Distribution")
15 for (i in 2:5) { lines(x, dfrechet(x, location=mu[i], shape=1, scale=1), col=color[i], lwd=2); }
16 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0.5', 'mu = 1', 'mu = 2', 'mu = 4', 'mu = 8'))
17
18 ### Cumulative Distribution
19 plot(x, pfrechet(x, location=mu[1], shape=1, scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative
   Distribution")
20 for (i in 2:5) { lines(x, pfrechet(x, location=mu[i], shape=1, scale=1), col=color[i], lwd=2); }
21 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0.5', 'mu = 1', 'mu = 2', 'mu = 4', 'mu = 8'))
22
23 ### Survival Function
24 plot(x, 1-pfrechet(x, location=mu[1], shape=1, scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival
   Function")
25 for (i in 2:5) { lines(x, 1-pfrechet(x, location=mu[i], shape=1, scale=1), col=color[i], lwd=2); }
26 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0.5', 'mu = 1', 'mu = 2', 'mu = 4', 'mu = 8'))
27
28 ### Hazard Function
29 plot(x, dfrechet(x, location=mu[1], shape=1, scale=1)/(1-pfrechet(x, location=mu[1], shape=1, scale=1)), xlim=c(0, 10), ylim=c(0,
   1), col=color[1], lwd=2, type = 'l', main="Hazard Function")
30 for (i in 2:5) { lines(x, dfrechet(x, location=mu[i], shape=1, scale=1)/(1-pfrechet(x, location=mu[i], shape=1, scale=1)), col=
   color[i], lwd=2); }
31 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0.5', 'mu = 1', 'mu = 2', 'mu = 4', 'mu = 8'))
32
33
34
35
36 ### parameter: m_p
37 m_p = c(0.5, 1, 2, 4, 8) # shape
38
39 ### Input Variable
40 x <- seq(0, 10, length.out = 101)
41
42 color = rainbow(10)
43
44 ### Life Distribution
45 plot(x, dfrechet(x, location=0, shape=m_p[1], scale=1), xlim=c(0, 10), ylim=c(0, 3), col=color[1], lwd=2, type = 'l', main="Life
   Distribution")
46 for (i in 2:5) { lines(x, dfrechet(x, location=0, shape=m_p[i], scale=1), col=color[i], lwd=2); }
47 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('m_p = 0.5', 'm_p = 1', 'm_p = 2', 'm_p = 4', 'm_p = 8'))
48
49 ### Cumulative Distribution
50 plot(x, pfrechet(x, location=0, shape=m_p[1], scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative
   Distribution")
51 for (i in 2:5) { lines(x, pfrechet(x, location=0, shape=m_p[i], scale=1), col=color[i], lwd=2); }
52 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('m_p = 0.5', 'm_p = 1', 'm_p = 2', 'm_p = 4', 'm_p = 8'))
53
54 ### Survival Function
55 plot(x, 1-pfrechet(x, location=0, shape=m_p[1], scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival
   Function")
56 for (i in 2:5) { lines(x, 1-pfrechet(x, location=0, shape=m_p[i], scale=1), col=color[i], lwd=2); }
57 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('m_p = 0.5', 'm_p = 1', 'm_p = 2', 'm_p = 4', 'm_p = 8'))
58
59 ### Hazard Function
60 plot(x, dfrechet(x, location=0, shape=m_p[1], scale=1)/(1-pfrechet(x, location=0, shape=m_p[1], scale=1)), xlim=c(0, 10), ylim=c(
   0, 7), col=color[1], lwd=2, type = 'l', main="Hazard Function")
61 for (i in 2:5) { lines(x, dfrechet(x, location=0, shape=m_p[i], scale=1)/(1-pfrechet(x, location=0, shape=m_p[i], scale=1)), col=
   color[i], lwd=2); }
62 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('m_p = 0.5', 'm_p = 1', 'm_p = 2', 'm_p = 4', 'm_p = 8'))
63
64
65
66 ### parameter: beta_p
67 beta_p = c(0.5, 1, 2, 4, 8) #scale
68
69 ### Input Variable
70 x <- seq(0, 10, length.out = 101)
71
72 color = rainbow(10)
73
74 ### Life Distribution
75 plot(x, dfrechet(x, location=0, shape=1, scale=beta_p[1]), xlim=c(0, 10), ylim=c(0, 1.5), col=color[1], lwd=2, type = 'l', main="Life
   Distribution")
```

```

76 for (i in 2:5) { lines(x, dfrechet(x, location=0, shape=1, scale=beta_p[i]), col=color[i], lwd=2); }
77 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8
    '))
78
79 ### Cumulative Distribution
80 plot(x, pfrechet(x, location=0, shape=1, scale=beta_p[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main=
    "Cumulative Distribution")
81 for (i in 2:5) { lines(x, pfrechet(x, location=0, shape=1, scale=beta_p[i]), col=color[i], lwd=2); }
82 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8
    '))
83
84 ### Survival Function
85 plot(x, 1-pfrechet(x, location=0, shape=1, scale=beta_p[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main=
    "Survival Function")
86 for (i in 2:5) { lines(x, 1-pfrechet(x, location=0, shape=1, scale=beta_p[i]), col=color[i], lwd=2); }
87 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8
    '))
88
89 ### Hazard Function
90 plot(x, dfrechet(x, location=0, shape=1, scale=beta_p[1])/(1-pfrechet(x, location=0, shape=1, scale=beta_p[1])), xlim=c(0, 10),
    ylim=c(0, 1.5), col=color[1], lwd=2, type = 'l', main="Hazard Function")
91 for (i in 2:5) { lines(x, dfrechet(x, location=0, shape=1, scale=beta_p[i])/(1-pfrechet(x, location=0, shape=1, scale=beta_p[i])),
    col=color[i], lwd=2); }
92 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p
    = 8'))

```

Type II 극치분포는 Frechet 분포로도 알려져 있다.

독일의 주가지수에서 극단적인 경우가 발생할 확률의 추정, 태양 양성자 죠고 유속의 특성을 예측하는 데 응용되었다.

16.1.12.5 Type II 극치분포(Frechet 분포) with Location Parameter

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{c}{b} \left(\frac{a-t}{b}\right)^{-(c+1)} e^{-\left(\frac{a-t}{b}\right)^{-c}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{-\left(\frac{a-t}{b}\right)^{-c}}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{c}{b} \left(\frac{a-t}{b}\right)^{-(c+1)}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \leq a$	
파라메터		$a \leq 0, b > 0, c > 0$	

Table 24: Type II 극치분포(Frechet 분포) with Location Parameter 함수에 기반한 척도 함수

16.1.13 F Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \frac{t^{\frac{\nu_1-2}{2}}}{\left(1+\frac{\nu_1}{\nu_2}t\right)^{\frac{\nu_1+\nu_2}{2}}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \leq t) \\ &= 1 - F(t) \\ &= e^{-H(t)} \end{aligned}$	No closed form	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	No closed form	
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$\begin{aligned} E[T] &= \int_0^\infty tf(t)dt \\ &= \int_0^\infty S(t)dt \end{aligned}$	작성중	
평균잔여수명함수	$\begin{aligned} m(t) &= E[T - t T > t] \\ &= \frac{\int_t^\infty S(u)du}{S(t)} \end{aligned}$	No closed form	
k 차 적률	$E(T^k)$	작성중	
$100p$ 백분위수	t_p	작성중	
변수		$t \geq 0$	
파라메터		$\nu_1 > 0, \nu_2 > 0$	

Table 25: F 분포함수에 기반한 척도 함수

16.1.14 Gamma Distribution(감마 분포)

감마분포는 지수분포를 보다 일반화시킨 수명분포로, 모수가 특정한 값이 될 때 지수분포, 카이제곱분포, Erlang 분포 등으로 변환되는 특징이 있다.

감마분포는 위험률(고장률) 형태가 복잡하여 와이블분포보다 덜 사용되고 있으나, 여러 가지 수명 분포에 잘 적용되는 모형이라는 것이 입증되어, 자료를 분석할 때 유용한 분포이다.

16.1.14.1 Gamma Distribution with 2 Parameters(2-모수 감마 분포)

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{\theta\Gamma(\beta)} \left(\frac{t}{\theta}\right)^{\beta-1} e^{-\frac{t}{\theta}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	$\int_0^t \frac{1}{\theta\Gamma(\beta)} \left(\frac{u}{\theta}\right)^{\beta-1} e^{-\frac{u}{\theta}} du$	
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - F(t) = e^{-H(t)}$	$\int_t^\infty \frac{1}{\theta\Gamma(\beta)} \left(\frac{u}{\theta}\right)^{\beta-1} e^{-\frac{u}{\theta}} du$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{\frac{1}{\theta\Gamma(\beta)} \left(\frac{t}{\theta}\right)^{\beta-1} e^{-\frac{t}{\theta}}}{\int_t^\infty \frac{1}{\theta\Gamma(\beta)} \left(\frac{u}{\theta}\right)^{\beta-1} e^{-\frac{u}{\theta}} du}$	
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	$-\ln \int_t^\infty \frac{1}{\theta\Gamma(\beta)} \left(\frac{u}{\theta}\right)^{\beta-1} e^{-\frac{u}{\theta}} du$	
평균수명	$E[T] = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt$	$\theta\beta$	
평균잔여수명함수	$m(t) = E[T-t T>t] = \frac{\int_t^\infty S(u)du}{S(t)}$	$\frac{\int_t^\infty \int_u^\infty \frac{1}{\theta\Gamma(\beta)} \left(\frac{x}{\theta}\right)^{\beta-1} e^{-\frac{x}{\theta}} dx du}{\int_t^\infty \frac{1}{\theta\Gamma(\beta)} \left(\frac{u}{\theta}\right)^{\beta-1} e^{-\frac{u}{\theta}} du}$	
k 차 적률	$E(T^k)$	$\frac{\theta^k \Gamma(k+\beta)}{\Gamma(\beta)}$	
100p 백분위수	t_p	$\int_t^\infty \frac{1}{\theta\Gamma(\beta)} \left(\frac{u}{\theta}\right)^{\beta-1} e^{-\frac{u}{\theta}} du = 1-p$ 를 만족하는 t 값	
변수		$t \geq 0$	
파라미터	$\beta > 0$ (형상 모수; shape parameter), $\theta > 0$ (척도 모수; scale parameter)		

Table 26: 2모수 감마 분포함수에 기반한 척도 함수

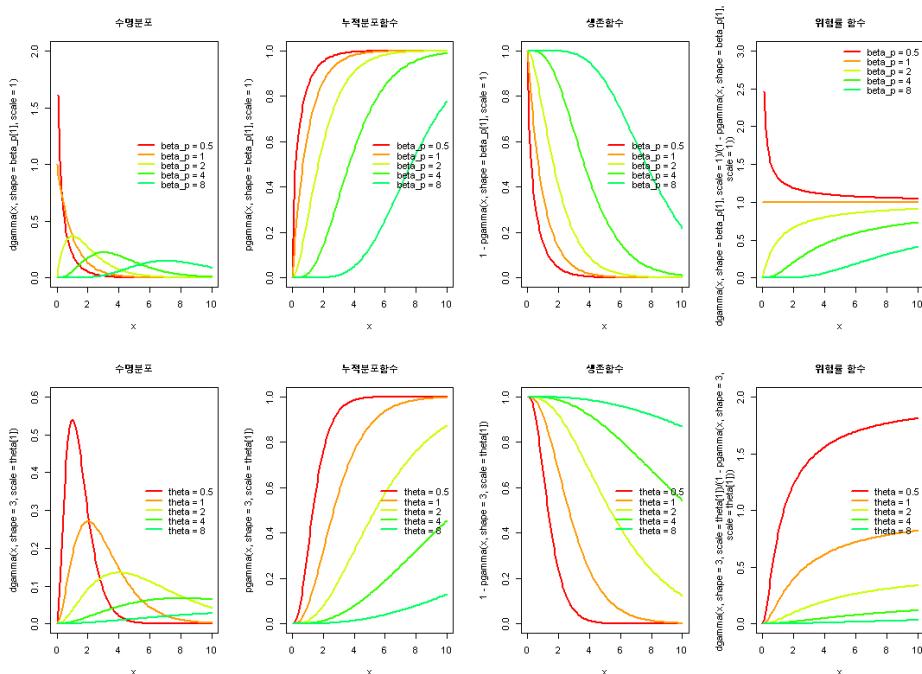


Figure 22: 감마분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 7. 감마분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```
1 ##### Gamma Distribution with 2 Parameters
2 par(mfrow = c(2, 4))
3
4 ### parameter: beta_p
5 beta_p = c(0.5, 1, 2, 4, 8) # shape
6
7 ### Input Variable
8 x <- seq(0, 10, length.out = 101)
9
10 color = rainbow(10)
11
12 ### Life Distribution
13 plot(x, dgamma(x, shape=beta_p[1], scale=1), xlim=c(0, 10), ylim=c(0, 2), col=color[1], lwd=2, type = 'l', main="Life Distribution")
14 for (i in 2:5) { lines(x, dgamma(x, shape=beta_p[i], scale=1), col=color[i], lwd=2); }
15 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8'))
16
17 ### Cumulative Distribution
18 plot(x, pgamma(x, shape=beta_p[1], scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative Distribution")
19 for (i in 2:5) { lines(x, pgamma(x, shape=beta_p[i], scale=1), col=color[i], lwd=2); }
20 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8'))
21
22 ### Survival Function
23 plot(x, 1-pgamma(x, shape=beta_p[1], scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
24 for (i in 2:5) { lines(x, 1-pgamma(x, shape=beta_p[i], scale=1), col=color[i], lwd=2); }
25 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8'))
26
27 ### Hazard Function
28 plot(x, dgamma(x, shape=beta_p[1], scale=1)/(1-pgamma(x, shape=beta_p[1], scale=1)), xlim=c(0, 10), ylim=c(0, 3), col=color[1], lwd=2, type = 'l', main="Hazard Function")
29 for (i in 2:5) { lines(x, dgamma(x, shape=beta_p[i], scale=1)/(1-pgamma(x, shape=beta_p[i], scale=1)), col=color[i], lwd=2); }
30 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8'))
31
32
33
34 ### parameter: theta
35 theta = c(0.5, 1, 2, 4, 8) #scale
36
37 ### Input Variable
38 x <- seq(0, 10, length.out = 101)
39
40 color = rainbow(10)
41
42 ### Life Distribution
43 plot(x, dgamma(x, shape=3, scale=theta[1]), xlim=c(0, 10), ylim=c(0, 0.6), col=color[1], lwd=2, type = 'l', main="Life Distribution")
44 for (i in 2:5) { lines(x, dgamma(x, shape=3, scale=theta[i]), col=color[i], lwd=2); }
45 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
46
47 ### Cumulative Distribution
48 plot(x, pgamma(x, shape=3, scale=theta[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative Distribution")
49 for (i in 2:5) { lines(x, pgamma(x, shape=3, scale=theta[i]), col=color[i], lwd=2); }
50 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
51
52 ### Survival Function
53 plot(x, 1-pgamma(x, shape=3, scale=theta[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
54 for (i in 2:5) { lines(x, 1-pgamma(x, shape=3, scale=theta[i]), col=color[i], lwd=2); }
55 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
56
57 ### Hazard Function
58 plot(x, dgamma(x, shape=3, scale=theta[1])/(1-pgamma(x, shape=3, scale=theta[1])), xlim=c(0, 10), ylim=c(0, 2), col=color[1], lwd =2, type = 'l', main="Hazard Function")
59 for (i in 2:5) { lines(x, dgamma(x, shape=3, scale=theta[i])/(1-pgamma(x, shape=3, scale=theta[i])), col=color[i], lwd=2); }
60 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
```

감마분포는 형상 모수 β 의 값에 따라서, 위험률 함수(고장률 함수)에 의한 비모수적 수명 분포 군(class of life distribution)이 달라진다. 즉, β 의 값이 위험률 함수(고장률 함수) $h(t)$ 의 단조성(monotonicity)에 영향을 미친다.

- $0 < \beta < 1$: $t = 0$ 에서 $h(t)$ 는 ∞ 의 값을 취하며, 그 이후 t 가 증가함에 따라 $h(t)$ 는 단조감소함수이며, $\lim_{t \rightarrow \infty} h(t) = \frac{1}{\theta}$ 을 만족하는 convex 함수이다.
따라서, 이 경우의 감마분포는 DFR 군에 속하게 되며, 초기 고장현상을 보이고, 시간이 경과함에 따라 위험률(고장률)이 감소하는 수명을 나타내는 데 적합해진다.
 - $\beta = 1$: $h(t) = \frac{1}{\theta}$ 로서, 나이에 종속되지 않는 상수 위험률(고장률)을 가지는 CFR 군에 속하게 된다. 이 때는 우연한 고장이 발생하는 경우의 수명을 모형화하는 데 적합하다.
 - $\beta > 1$: $t = 0$ 에서 $h(t) = 0$ 이 되며, 그 이후 t 가 증가함에 따라 $h(t)$ 는 단조증가함수로서, IFR 군에 속하며, $\lim_{t \rightarrow \infty} h(t) = \frac{1}{\theta}$
-

Note 18. 감마 함수

감마함수 $\Gamma(a)$ 는 다음과 같이 정의된다.

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

$\Gamma(a)$ 는 위 식의 우변 항에 있는 적분을 나타내는 기호이다.

감마 함수의 성질은 다음과 같다.

- 부분적분에 의하여, 다음과 같은 성질이 있다.

$$\Gamma(a+1) = a\Gamma(a)$$

- 만일, $a = n$, 즉 a 가 정수이면, 다음과 같은 식이 성립한다.

- $\Gamma(n+1) = n!$
- $\Gamma(\frac{1}{2}) = 2\Gamma(\frac{3}{2}) = \sqrt{\pi}$
- $\Gamma(n + \frac{1}{2}) = \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2^n} \sqrt{\pi}$

Note 19. 감마분포에 기반한 생존함수(신뢰도함수)

감마분포에 기반한 생존함수는, 감마함수를 적분할 수 없기 때문에 적분을 포함한 식으로 표현된다.

$$S(t) = \int_t^\infty \frac{1}{\theta\Gamma(\beta)} \left(\frac{u}{\theta}\right)^{\beta-1} e^{-\frac{u}{\theta}} du$$

만일 β 가 정수이면, 생존함수(신뢰도함수)에 대해 부분적분을 반복하여 다음과 같은 식을 얻을 수 있다.

$$S(t) = \sum_{k=0}^{\beta-1} \frac{1}{k!} \left(\frac{t}{\theta}\right)^k e^{-\frac{t}{\theta}} \quad (3)$$

식 (3)의 오른쪽 항은, 평균이 $\frac{t}{\theta}$ 인 포아송 분포에서, $\beta - 1$ 까지의 누적분포함수와 같다.

Note 20. 감마분포에 기반한 위험률함수(고장률함수)

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{\theta\Gamma(\beta)} \left(\frac{t}{\theta}\right)^{\beta-1} e^{-\frac{t}{\theta}}}{\int_t^\infty \frac{1}{\theta\Gamma(\beta)} \left(\frac{u}{\theta}\right)^{\beta-1} e^{-\frac{u}{\theta}} du}$$

감마분포의 위험률(고장률)은, β 에 따라 다음과 같은 성질이 있다.

- $\beta > 1$ 일 때: 위험률(고장률)이 0에서 출발하여 단조증가하여, $t \rightarrow \infty$ 이면 $\frac{1}{\theta}$ 로 수렴한다.
 - $0 < \beta < 1$ 일 때: 위험률(고장률)이 ∞ 에서 출발하여 단조감소하여, $t \rightarrow \infty$ 이면 $\frac{1}{\theta}$ 로 수렴한다.
 - $\beta = 1$ 일 때: 위험률(고장률)이 $\frac{1}{\theta}$ 로 일정한 값을 가진다.
-

Note 21. 감마분포에 기반한 k차 적률

$$\begin{aligned} E[T^k] &= \int_0^\infty t^k f(t) dt \\ &= \int_0^\infty t^k \left(\frac{t}{\theta}\right)^{\beta-1} e^{-\frac{t}{\theta}} dt \\ &= \frac{\theta^k \Gamma(k+\beta)}{\Gamma(\beta)} \int_0^\infty \frac{1}{\theta\Gamma(k+\beta)} \left(\frac{t}{\theta}\right)^{k+\beta-1} e^{-\frac{t}{\theta}} dt \\ &= \frac{\theta^k \Gamma(k+\beta)}{\Gamma(\beta)} \end{aligned}$$

감마분포는 다음과 같은 성질이 있다.

- Figure 22에서 알 수 있듯이, 감마분포는 오른쪽으로 치우쳐 있다.

– $\beta < 1$ 일 때: t 값이 0에 가까워질수록 확률밀도함수의 값은 무한대가 된다.

– $\beta = 1$ 일 때: 감마분포는 지수분포와 동일하게 $t = 0$ 에서 확률밀도함수의 값은 $\frac{1}{\theta}$ 이며, t 가 커짐에 따라 감소한다.

– $\beta > 1$ 일 때: 감마분포는 비대칭 종모양을 하는데, $t = 0$ 일 때 $f(t) = 0$ 이며, 증가하다가 감소하는 형태를 보여준다.

이러한 특징은 와이블 분포와 유사하다.

- 감마분포를 따르는 제품의 평균수명은 $\theta\beta$ 표준편차는 $\theta\beta^{\frac{1}{2}}$ 이다.
- 감마분포의 중앙값(t_m)은 다음과 같다.

$$0.5 = \int_0^{t_m} \frac{1}{\theta\Gamma(\beta)} \left(\frac{t}{\theta}\right)^{\beta-1} e^{-\frac{t}{\theta}} dt$$

주어진 θ 와 β 값에 대해 불완전 감마함수 표를 이용하여 중앙값을 구할 수 있으나, $t_m \approx \theta(\beta - 0.3)$ 식을 통해 근사값을 구하기도 한다.

- 감마분포에서 $\beta = 1$ 인 특별한 경우에, 감마분포는 지수분포가 된다.

반대로, 만일 X_1, X_2, \dots, X_r 이 평균이 $\frac{1}{\lambda}$ 인 지수분포를 따르는 서로 독립적인 확률변수라면,

이 확률변수들의 합인 $\sum_{i=1}^r X_i$ 은 형상모수(shape parameter) β 가 r 인 감마분포를 따른다.

- 위험률(고장률)이 λ (상수)인 지수분포를 따르는 서로 독립인 β (정수) 개의 확률변수 T_1, T_2, \dots, T_β 들에 대하여

$T = \sum_{i=1}^\beta T_i \sim Gamma(\beta, \theta = \frac{1}{\lambda})$ 가 된다.²⁵

- 감마분포에서 $\beta = \frac{\nu}{2}$, $\theta = 2$ 인 감마분포는, 자유도가 ν 인 카이제곱분포가 된다.

- β 가 정수로 제한되는 경우, 감마분포는 Erlang분포가 된다.²⁶

- 감마분포의 확률밀도함수는, β 가 충분히 커지면, 평균이 $\theta\beta$ 이고 표준편차가 $\theta\beta^{\frac{1}{2}}$ 인 정규분포에 근접한다.

²⁵이는, 적률생성함수를 이용하여 증명할 수 있다.

²⁶Erlang 분포는 대기 이론이나 traffic 분석에서 자주 이용된다. Erlang 분포의 확률밀도함수는 다음과 같다.

$$f(t; \beta, \theta) = \frac{1}{\theta(\beta-1)!} \left(\frac{t}{\theta}\right)^{\beta-1} e^{-\frac{t}{\theta}}$$

Note 22. 감마분포를 따르는 완전자료에 대한 모수 추정

T_1, T_2, \dots, T_n 을 시험 중인 n 개의 부품(환자)에 대한 고장(재발) 시간을 나타내는 확률변수라 하고, t_1, t_2, \dots, t_n 을 이들의 실제 관측값이라고 하자. T_1, T_2, \dots, T_n 은 감마 분포를 따른다고 하자. 완전 자료를 이용하면, 감마분포의 형상모수 β 와 척도모수 θ 의 추정량은 다음과 같이 계산된다.

- 적률 추정량(moment estimator)

적률 함수를 이용하여 감마분포의 모수를 추정하는 방법은 다음과 같다.

$m_1 = E(X) = \theta\beta$ 이고, $m_2 = E(X^2) = \theta^2\beta(\beta + 1)$ 이므로, 표본적률값을 이용하여 다음과 같은 방정식을 세울 수 있다.

$$\theta\beta = \frac{1}{n} \sum_{i=1}^n t_i = \bar{t}$$

$$\theta^2\beta(\beta + 1) = \sum_{i=1}^n \frac{t_i^2}{n}$$

위의 방정식을 풀면, θ 와 β 에 대해 다음과 같은 적률추정량을 구할 수 있다.

$$\hat{\theta} = \frac{\frac{\bar{t}^2}{n} - \bar{t}^2}{\bar{t}}$$

$$\hat{\beta} = \frac{\bar{t}}{\frac{\bar{t}^2}{n} - \bar{t}^2}$$

감마분포의 경우, 표본의 크기가 크지 않은 경우, 적률추정량은 최대가능도추정량에 비하여 유효성에 차이가 없으므로, 최대가능도추정량 대신 사용할 수 있다.s

- 최대 가능도 추정량(Maximum Likelihood Estimator)

감마함수의 가능도함수는 다음과 같다.

$$\begin{aligned} L(t_i; \beta, \theta) &= \prod_{i=1}^n \frac{1}{\theta \Gamma(\beta)} \left(\frac{t_i}{\theta} \right)^{\beta-1} e^{-\frac{t_i}{\theta}} \\ &= \left(\frac{1}{\theta^\beta \Gamma(\beta)} \right)^n \prod_{i=1}^n t_i^{\beta-1} e^{-\frac{\sum_{i=1}^n t_i}{\theta}} \end{aligned}$$

양변에 자연로그를 취하면,

$$\ln L(t_i; \beta, \theta) = -n\beta \ln \theta - n \ln \Gamma(\beta) + (\beta - 1) \ln \prod_{i=1}^n t_i - \left(\frac{n\bar{t}}{\theta} \right)$$

위 식을 θ 와 β 에 대한 1차도함수를 구하고 0으로 놓으면 얻어지는 가능도 방정식을 풀면, 다음과 같은 식을 얻고, 이 식으로부터 최대가능도추정량(MLE)을 구할 수 있다.

$$\ln \beta - \psi(\beta) = \ln \bar{t} - \frac{1}{b} \ln \prod_{i=1}^n t_i \quad (4)$$

$$\theta = \frac{\bar{t}}{\beta}$$

$$\text{여기서, } \psi(\beta) = \frac{\Gamma''(\beta)}{\Gamma(\beta)}$$

실제 데이터를 이용하여 $d = \ln \bar{t} - \frac{1}{n} \ln \prod_{i=1}^n t_i$ 가 주어졌을 때, 식 (4)을 만족하는 $\hat{\beta}$ 를 구하는 방법들이 제안되었는데, 그 중 2가지만 소개하면 다음과 같다.

- Greenwood and Duran(1960)[27]

$$\hat{\beta} = \begin{cases} \frac{0.5000876 + 0.1648852d - 0.0544274d^2}{d} & 0 < d \leq 0.5772 \\ \frac{8.898919 + 9.059950d + 0.9775373d^2}{d(17.79728 + 11.968477d + d^2)} & 0.5772 < d \leq 17 \end{cases}$$

이 식은 무언가 rough하게 정해진 듯 하면서도, 추정오차가 0.1% 미만으로 매우 정확한 것으로 알려져 있다.

- Bowman and Shenton(1983)[17]

$$\hat{\beta} = \begin{cases} \frac{1}{d + \ln d} & n \circ \text{작을 때} \\ \frac{1}{2d} + \frac{1}{6} - \frac{d}{18} - \frac{4d^2}{135} + \frac{47d^2}{810} & n \circ \text{클 때} \end{cases}$$

- 신뢰구간

감마분포의 모수에 대한 신뢰구간은 Bain(1978)이 제안한 방법과, 표본의 크기가 큰 경우 근사적으로 구하는 방법이 있다.

- Bain(1978)의 방법[9]

표본의 크기가 작아도 $2n\beta d$ 는 근사적으로 자유도가 $(n - 1)$ 인 카이제곱분포를 따른다. 따라서, 만일 $\beta \geq 2$ 이고, $n \geq 10$ 이라면, 다음과 같은 신뢰구간의 상한과 하한을 구할 수 있다.

$$\left[\frac{1}{2nd} \chi^2 \left(n - 1; \frac{1 - \alpha}{2} \right) < \beta < \frac{1}{2nd} \chi^2 \left(n - 1; \frac{\alpha}{2} \right) \right]$$

$$\left[\frac{2n\bar{t}}{\chi^2 \left(2n [\eta_U]; \frac{\alpha}{4} \right)} < \theta < \frac{2n\bar{t}}{\chi^2 \left(2n [\eta_L]; \frac{1-\alpha}{4} \right)} \right]$$

여기서, $[\eta_U], [\eta_L]$ 은 β 의 신뢰상한과 하한의 정수부분을 의미함

- 근사적 신뢰구간

표본의 크기가 충분히 큰 경우, $\hat{\theta}$ 와 $\hat{\beta}$ 의 근사적 분포는 다음과 같다.

$$\hat{\theta} \sim AN \left[\theta, \frac{\theta^2}{n} \left(\frac{\psi'(\beta)}{\beta\psi'(\beta) - 1} \right) \right]$$

$$\hat{\beta} \sim AN \left[\beta, \frac{1}{n} \left(\frac{\beta}{\beta\psi'(\beta) - 1} \right) \right]$$

여기서, $\psi(\beta) = \frac{d \ln \Gamma(\beta)}{d\beta}$: Digamma 함수

이에 따른 근사적 신뢰구간은 다음과 같다.

$$\left[\hat{\beta} - z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n}} \left(\frac{\hat{\beta}}{\hat{\beta}\psi'(\hat{\beta}) - 1} \right)^{\frac{1}{2}} < \beta < \hat{\beta} + z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n}} \left(\frac{\hat{\beta}}{\hat{\beta}\psi'(\hat{\beta}) - 1} \right)^{\frac{1}{2}} \right]$$

$$\left[\hat{\theta} - z_{\frac{\alpha}{2}} \frac{\hat{\theta}}{\sqrt{n}} \left(\frac{\psi'(\hat{\beta})}{\hat{\beta}\psi'(\hat{\beta}) - 1} \right)^{\frac{1}{2}} < \theta < \hat{\theta} + z_{\frac{\alpha}{2}} \frac{\hat{\theta}}{\sqrt{n}} \left(\frac{\psi'(\hat{\beta})}{\hat{\beta}\psi'(\hat{\beta}) - 1} \right)^{\frac{1}{2}} \right]$$

Note 23. Digamma 함수

$$\text{Digamma 함수 : } \psi(\beta) = \frac{d \ln \Gamma(\beta)}{d\beta}$$

$\psi(\beta)$ 는 다음과 같은 성질을 가지고 있다.

- $\beta > 0$ 에 대해, $\psi(\beta + 1) = \psi(\beta) + \frac{1}{\beta} = -\gamma + \sum_{n=1}^{\infty} \frac{\beta}{n(n+\beta)}$
- $\beta > 0$ 에 대해, $\psi'(\beta) = \sum_{k=0}^{\infty} (\beta + k)^{-2}$
- $n \geq 2, 0 < z < 1$ 에 대해, $\psi'(n + z) = \psi(1 + z) - \sum_{j=1}^{n-1} (j + z)^{-2}$

Note 24. 감마분포를 따르는 Type II 우중도절단자료(정수중단자료)에 대한 모수 추정

2모수 감마분포를 따르는 n 개의 부품(환자)을 가지고 수명시험을 수행하여 $r(\leq n)$ 개의 고장(재발)이 발생한 시점에서 시험을 중단하였을 때, 일어진 고장(재발)자료를 크기 순서대로 나열하여 $t_{(1)} \leq t_{(2)} \leq \dots \leq t_{(r)}$ 으로 나타낸다. 나머지 $(n - r)$ 개의 부품(환자)에 대한 고장(재발)자료는 시점 $t_{(r)}$ 에서 절단된다.

Type II 우중도절단자료(정수중단자료)가 주어진 경우의 β 와 θ 에 대한 가능도함수는 다음과 같다.

$$\begin{aligned} L(\beta, \theta) &= \prod_{i=1}^r f(t_{(i)}; \beta, \theta) (P(T > t_{(r)}))^{n-r} \\ &= \prod_{i=1}^r \frac{1}{\theta \Gamma(\beta)} \left(\frac{t_{(i)}}{\theta}\right)^{\beta-1} e^{-\frac{t_{(i)}}{\theta}} \left(1 - IG\left(\beta, \frac{t_{(r)}}{\theta}\right)\right)^{n-r} \\ &= \left(\frac{1}{\Gamma(\beta)}\right)^r \left(\frac{1}{\theta}\right)^{r\beta} \prod_{i=1}^r t_{(i)}^{\beta-1} e^{-\frac{\sum_{i=1}^r t_{(i)}}{\theta}} \left(1 - IG\left(\beta, \frac{t_{(r)}}{\theta}\right)\right)^{n-r} \end{aligned}$$

여기서, $IG(a, b) = \frac{1}{\Gamma(a)} \int_0^b u^{a-1} e^{-u} du$: 불완전 감마 함수

양변에 자연로그를 취하면,

$$\ln L(\beta, \theta) = -r\beta \ln \theta - r \ln \Gamma(\beta) + (\beta - 1) \ln \tilde{t} - \frac{r\bar{t}}{\theta} + (n - r) \ln \left[1 - IG\left(\beta, \frac{t_{(r)}}{\theta}\right)\right] \quad (5)$$

$$\text{여기서, } \tilde{t} = \prod_{i=1}^r t_{(i)} \quad \bar{t} = \frac{\prod_{i=1}^r t_{(i)}}{r}$$

β 와 θ 의 최대가능도추정량을 구하기 위해, 식(5)의 1차도함수를 구하여 각각 0으로 놓으면, 다음과 같은 2개의 방정식을 얻게 된다.

$$\frac{d \ln L(\beta, \theta)}{d\beta} = -r \ln \theta - r\psi(\beta) + \ln \tilde{t} - \frac{n - r}{1 - IG\left(\beta, \frac{t_{(r)}}{\theta}\right)} \cdot \frac{dIG\left(\beta, \frac{t_{(r)}}{\theta}\right)}{d\beta} \quad (6)$$

$$\frac{d \ln L(\beta, \theta)}{d\theta} = -\frac{r\beta}{\theta} + \frac{r\bar{t}}{\theta^2} - \frac{n - r}{1 - IG\left(\beta, \frac{t_{(r)}}{\theta}\right)} \cdot \frac{dIG\left(\beta, \frac{t_{(r)}}{\theta}\right)}{d\theta} \quad (7)$$

$$\begin{aligned} \text{여기서, } \psi(\beta) &= \frac{\Gamma'(\beta)}{\Gamma(\beta)} \\ IG\left(\beta, \frac{t_{(r)}}{\theta}\right) &= \frac{1}{\Gamma(\beta)} \int_0^{\frac{t_{(r)}}{\theta}} u^{\beta-1} e^{-u} du \\ \frac{dIG\left(\beta, \frac{t_{(r)}}{\theta}\right)}{d\beta} &= \frac{1}{\Gamma(\beta)} \int_0^{\frac{t_{(r)}}{\theta}} u^{\beta-1} (\ln u) e^{-u} du - \psi(\beta)IG\left(\beta, \frac{t_{(r)}}{\theta}\right) \\ \frac{dIG\left(\beta, \frac{t_{(r)}}{\theta}\right)}{d\theta} &= \frac{-1}{\theta \Gamma(\beta)} \left(\frac{t_{(r)}}{\theta}\right)^\beta e^{-\frac{t_{(r)}}{\theta}} \end{aligned}$$

감마분포의 경우, 최대가능도추정량은 가능도방정식을 풀어서 해를 구하는 것이 복잡하므로,

- 보통 로그가능도함수 식 (5)을 최대화하는 β , θ 를 직접 찾는 방법을 사용하거나,
 - '가능한 β ' 값의 범위에서 여러 개의 β 값을 선택하고,
주어진 β 에 대해서 식 (6), (7)을 풀어 $\hat{\theta}(\beta)$ 를 구한 후, 이를 대수가능도함수인 식 (5)에 대입하고,
이 식 (5)를 최대로 하는 β 를 구하여 최대가능도추정량 $\hat{\beta}$ 으로 하고, θ 의 최대가능도추정량 $\hat{\theta}$ 는 $\hat{\theta}(\hat{\beta})$ 로 구하는 방법을 이용한다.
-

16.1.14.2 Gamma Distribution with Location Parameter

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$(t - a)^{c-1} e^{-\frac{t-a}{b}} [b^c \Gamma(c)]^{-1}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\frac{\gamma[c, \frac{t-a}{b}]}{\Gamma(c)}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$(t - a)^{c-1} e^{-\frac{t-a}{b}} [b^c \Gamma(c)]^{-1}$	
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	
k 차 적률	$E(T^k)$	작성중	
$100p$ 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0, c > 0$	

Table 27: 3모수 감마 분포함수에 기반한 척도 함수

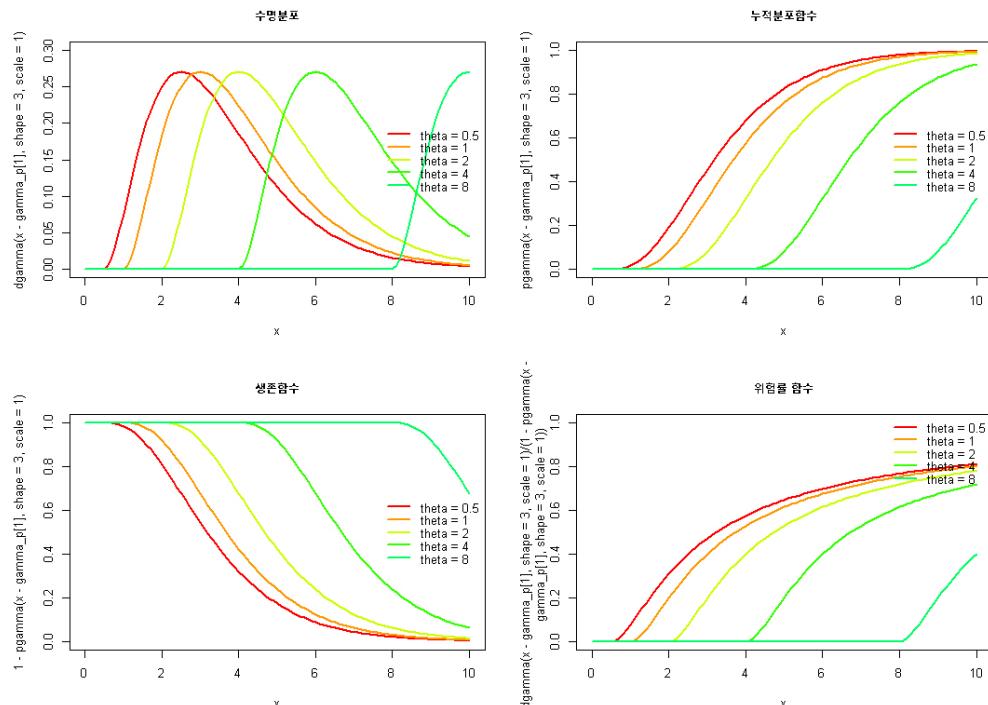


Figure 23: 3모수 감마분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 8. 3모수 감마분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```
1 ##### Gamma Distribution with 3 Parameters
2 ### parameter: alpha
3 gamma_p = c(0.5, 1, 2, 4, 8) # location
4
5 ### Input Variable
6 x <- seq(0, 10, length.out = 101)
7
8 color = rainbow(10)
9 par(mfrow = c(2, 2))
10
11
12 ### Life Distribution
13 plot(x, dgamma(x-gamma_p[1], shape=3, scale=1), xlim=c(0, 10), ylim=c(0, 0.3), col=color[1], lwd=2, type = 'l', main="Life
   Distribution")
14 for (i in 2:5) { lines(x, dgamma(x-gamma_p[i], shape=3, scale=1), col=color[i], lwd=2); }
15 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
16
17 ### Cumulative Distribution
18 plot(x, pgamma(x-gamma_p[1], shape=3, scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative
   Distribution")
19 for (i in 2:5) { lines(x, pgamma(x-gamma_p[i], shape=3, scale=1), col=color[i], lwd=2); }
20 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
21
22 ### Survival Function
23 plot(x, 1-pgamma(x-gamma_p[1], shape=3, scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival
   Function")
24 for (i in 2:5) { lines(x, 1-pgamma(x-gamma_p[i], shape=3, scale=1), col=color[i], lwd=2); }
25 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
26
27 ### Hazard Function
28 plot(x, dgamma(x-gamma_p[1], shape=3, scale=1)/(1-pgamma(x-gamma_p[1], shape=3, scale=1)), xlim=c(0, 10), ylim=c(0, 1), col=color
   [1], lwd=2, type = 'l', main="Hazard Function")
29 for (i in 2:5) { lines(x, dgamma(x-gamma_p[i], shape=3, scale=1)/(1-pgamma(x-gamma_p[i], shape=3, scale=1)), col=color[i], lwd=2);
  }
30 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
```

시스템의 수명이 일정기간 동안 보장되는 경우에는, 3모수 감마분포를 적용하는 것이 보다 적절하다.

척도 모수(scale parameter)는 종종 $\lambda = \frac{1}{\theta}$ 로 표시되는 경우도 있다.
위치모수 $\gamma = 0$ 이면 2모수 감마 분포가 된다.

16.1.14.3 Log-gamma Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{e^{c\frac{t-a}{b}} - e^{\frac{t-a}{b}}}{b\Gamma(c)}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\frac{\gamma \left[c, e^{\frac{t-a}{b}} \right]}{\Gamma(c)}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	No closed form	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty tf(t)dt$ $= \int_0^\infty S(t)dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T > t]$ $= \frac{\int_t^\infty S(u)du}{S(t)}$	No closed form	DMRL
k 차 적률	$E(T^k)$	작성중	
100 p 백분위수	t_p	작성중	
변수		$t \in \mathbf{R}$	
파라메터		$a \in \mathbf{R}, b > 0, c > 0$	

Table 28: Log-gamma 분포함수에 기반한 척도 함수

16.1.14.4 Generalized Gamma Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{d(t-a)^{cd-1}}{b^{cd}\Gamma(c)} e^{-\left(\frac{t-a}{b}\right)^d}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	$\frac{\gamma\left[c, \left(\frac{t-a}{b}\right)^d\right]}{\Gamma(c)}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{d(t-a)^{cd-1} e^{-\left(\frac{t-a}{b}\right)^d}}{b^{cd} \gamma\left[c, \left(\frac{t-a}{b}\right)^d\right]}$	
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T>t] = \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0, c > 0, d > 0$	

Table 29: Generalized Gamma 분포함수에 기반한 척도 함수

Stacy(1962)[59]는 파라메터가 3개인 Gamma 분포에서, 파라메터가 4개인 위의 generalized gamma 분포를 유도하였다.

16.1.15 Generalized Exponential Geometric Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{c(1-p)e^{-\frac{t-a}{b}} \left[1-e^{-\frac{t-a}{b}}\right]^{c-1}}{b \left[1-pe^{-\frac{t-a}{b}}\right]^{c-1}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - \left[\frac{1-e^{-\frac{t-a}{b}}}{1-pe^{-\frac{t-a}{b}}} \right]^c$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{c(1-p)e^{-\frac{t-a}{b}} \left[1-e^{-\frac{t-a}{b}}\right]^{c-1}}{b \left[1-pe^{-\frac{t-a}{b}}\right]^{c-1} - \left[1-pe^{-\frac{t-a}{b}}\right]^c \left[1-e^{-\frac{t-a}{b}}\right]^c}$	IDHR for $p \in \left(\frac{c-1}{c+1}, 1\right)$ and $c > 1$ IHR for $p \in \left(0, \frac{c-1}{c+1}\right)$ and $c > 1$ DHR for otherwise.
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty t S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T>t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DIMRL or multiply bended when IDHR or DHR DMRL when IHR
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}$, $b > 0$, $c > 0$, $p \in (0, 1)$	

Table 30: Generalized Exponential 분포함수에 기반한 척도 함수

16.1.16 Gompertz Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\alpha e^{\beta t} e^{\frac{\alpha}{\beta}} [1 - e^{\beta t}]$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{\frac{\alpha}{\beta}} [1 - e^{\beta t}]$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\alpha e^{\beta t}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DMRL
k 차 적률	$E(T^k)$	작성중	
100 p 백분위수	t_p	작성중	
변수		$t \geq 0$	
파라메터		$\alpha \geq 0, \beta > 0$	

Table 31: Gompertz 분포함수에 기반한 척도 함수

16.1.17 Gompertz-Makeham Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\zeta \left[\frac{\alpha}{\beta} (1 - e^{\beta t}) - \zeta t \right] + \alpha e^{\beta t} e^{\frac{\alpha}{\beta}(1-e^{\beta t})-\zeta t}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{\frac{\alpha}{\beta}(1-e^{\beta t})-\zeta t}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\zeta + \alpha e^{\beta t}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq 0$	
파라메터		$\alpha > 0, \beta > 0, \zeta > 0$	

Table 32: Gompertz-Makeham 분포함수에 기반한 척도 함수

16.1.18 Hjorth Distribution(호스 분포)

Table 33: 호스 분포함수에 기반한 척도 함수

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{(1+\beta t)\delta t^2 + \theta}{(1+\beta t)^{\frac{\theta}{\beta}+1}} e^{-\frac{1}{2}\delta t^2}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\frac{e^{-\frac{1}{2}\delta t^2}}{(1+\beta t)^{\frac{\theta}{\beta}}}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\delta t + \frac{\theta}{1+\beta t}$	
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq 0$	
파라미터		$\delta \geq 0, \beta \geq 0, \theta \geq 0$	

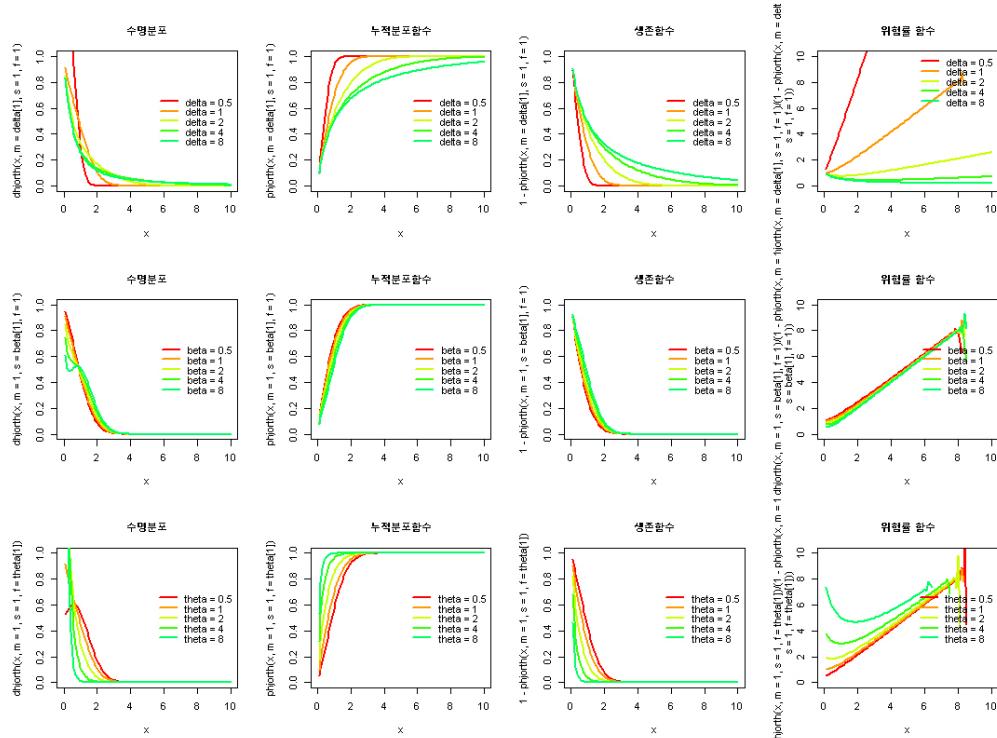


Figure 24: 호스분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 9. 호스분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```
1 ##### 호스 분포
2 ### f(y) = (1+s y)^(-f/s) exp(-(y/m)^2/2) (y/m^2+f/(1+s y))
3 ### y = x or t, s = beta, f = theta, m = delta
4 require(rmutil)
5
6 par(mfrow = c(3, 4))
7
8 ### parameter: delta
9 delta = c(0.5, 1, 2, 4, 8) # delta
10
11 ### Input Variable
12 x <- seq(0.1, 10, length.out = 101)
13
14 color = rainbow(10)
15
16 ### Life Distribution
17 plot(x, dhjorth(x, m=delta[1], s=1, f=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Life Distribution")
18 for (i in 2:5) { lines(x, dhjorth(x, m=delta[i], s=1, f=1), col=color[i], lwd=2); }
19 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('delta = 0.5', 'delta = 1', 'delta = 2', 'delta = 4', 'delta = 8'))
20
21 ### Cumulative Distribution
22 plot(x, phjorth(x, m=delta[1], s=1, f=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative Distribution")
23 for (i in 2:5) { lines(x, phjorth(x, m=delta[i], s=1, f=1), col=color[i], lwd=2); }
24 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('delta = 0.5', 'delta = 1', 'delta = 2', 'delta = 4', 'delta = 8'))
25
26 ### Survival Function
27 plot(x, 1-phjorth(x, m=delta[1], s=1, f=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
28 for (i in 2:5) { lines(x, 1-phjorth(x, m=delta[i], s=1, f=1), col=color[i], lwd=2); }
29 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('delta = 0.5', 'delta = 1', 'delta = 2', 'delta = 4', 'delta = 8'))
30
31 ### Hazard Function
32 plot(x, dhjorth(x, m=delta[1], s=1, f=1)/(1-phjorth(x, m=delta[1], s=1, f=1)), xlim=c(0, 10), ylim=c(0, 10), col=color[1], lwd=2,
      type = 'l', main="Hazard Function")
33 for (i in 2:5) { lines(x, dhjorth(x, m=delta[i], s=1, f=1)/(1-phjorth(x, m=delta[i], s=1, f=1)), col=color[i], lwd=2); }
34 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('delta = 0.5', 'delta = 1', 'delta = 2', 'delta = 4', 'delta = 8'))
35
36
37
38 ### parameter: beta
39 beta = c(0.5, 1, 2, 4, 8) # beta
40
41 ### Input Variable
42 x <- seq(0.1, 10, length.out = 101)
43
44 color = rainbow(10)
45
46 ### Life Distribution
47 plot(x, dhjorth(x, m=1, s=beta[1], f=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Life Distribution")
48 for (i in 2:5) { lines(x, dhjorth(x, m=1, s=beta[i], f=1), col=color[i], lwd=2); }
49 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta = 0.5', 'beta = 1', 'beta = 2', 'beta = 4', 'beta = 8'))
50
51 ### Cumulative Distribution
52 plot(x, phjorth(x, m=1, s=beta[1], f=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative Distribution")
53 for (i in 2:5) { lines(x, phjorth(x, m=1, s=beta[i], f=1), col=color[i], lwd=2); }
54 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta = 0.5', 'beta = 1', 'beta = 2', 'beta = 4', 'beta = 8'))
55
56 ### Survival Function
57 plot(x, 1-phjorth(x, m=1, s=beta[1], f=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
58 for (i in 2:5) { lines(x, 1-phjorth(x, m=1, s=beta[i], f=1), col=color[i], lwd=2); }
59 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta = 0.5', 'beta = 1', 'beta = 2', 'beta = 4', 'beta = 8'))
60
61 ### Hazard Function
62 plot(x, dhjorth(x, m=1, s=beta[1], f=1)/(1-phjorth(x, m=1, s=beta[1], f=1)), xlim=c(0, 10), ylim=c(0, 10), col=color[1], lwd=2,
      type = 'l', main="Hazard Function")
63 for (i in 2:5) { lines(x, dhjorth(x, m=1, s=beta[i], f=1)/(1-phjorth(x, m=1, s=beta[i], f=1)), col=color[i], lwd=2); }
64 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta = 0.5', 'beta = 1', 'beta = 2', 'beta = 4', 'beta = 8'))
65
66
67
68 ### parameter: theta
69 theta = c(0.5, 1, 2, 4, 8) #theta
70
71 ### Input Variable
72 x <- seq(0.1, 10, length.out = 101)
73
74 color = rainbow(10)
75
76 ### Life Distribution
77 plot(x, dhjorth(x, m=1, s=1, f=theta[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Life Distribution")
78 for (i in 2:5) { lines(x, dhjorth(x, m=1, s=1, f=theta[i]), col=color[i], lwd=2); }
79 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
```

```

81  ## Cumulative Distribution
82  plot(x, phjorth(x, m=1, s=1, f=theta[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative
83  Distribution")
84  for (i in 2:5) { lines(x, phjorth(x, m=1, s=1, f=theta[i]), col=color[i], lwd=2); }
85  legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
86  ## Survival Function
87  plot(x, 1-phjorth(x, m=1, s=1, f=theta[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function"
88  )
89  for (i in 2:5) { lines(x, 1-phjorth(x, m=1, s=1, f=theta[i]), col=color[i], lwd=2); }
90  legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
91  ## Hazard Function
92  plot(x, dhjorth(x, m=1, s=1, f=theta[1])/(1-phjorth(x, m=1, s=1, f=theta[1])), xlim=c(0, 10), ylim=c(0, 10), col=color[1], lwd=2,
93  type = 'l', main="Hazard Function")
94  for (i in 2:5) { lines(x, dhjorth(x, m=1, s=1, f=theta[i])/(1-phjorth(x, m=1, s=1, f=theta[i])), col=color[i], lwd=2); }
94  legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))

```

호스 분포는 3개의 모수를 가진 분포로, 모수들의 값에 따라 위험률 함수(고장률 함수)가 증가, 감소, 육조 모양을 가지는 매우 유동적인 수명 분포이다.

- $\theta = 0$: 레일리 분포
- $\theta = \beta = 0$: 지수 분포
- $\delta = 0$: 위험률(고장률)이 감소하는 분포(DFR)
- $\delta \geq \theta\beta$: 위험률(고장률)이 증가하는 분포(IFR)
- $0 < \delta < \theta\beta$: 위험률 함수(고장률 함수)가 육조 모양인 분포

16.1.19 Hyperbolic Secant Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{b\pi} \operatorname{sech}\left(\frac{t-a}{b}\right)$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - \frac{2}{\pi} \arctan\left(e^{\frac{t-a}{b}}\right)$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{\operatorname{sech}\left(\frac{t-a}{b}\right)}{b \left[\pi - 2 \arctan\left(e^{\frac{t-a}{b}}\right) \right]}$	IHR with $\lim_{t \rightarrow \infty} h(t) = \frac{1}{b}$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty tf(t)dt$ $= \int_0^\infty S(t)dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u)du}{S(t)}$	No closed form	DMRL with $\lim_{t \rightarrow \infty} m(t) = b$
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \in \mathbf{R}$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 34: Hyperbolic Secant 분포함수에 기반한 척도 함수

16.1.20 Laplace Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{2b} e^{-\frac{ t-a }{b}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - \frac{1}{2} \left[1 + \text{sign}(t-a) \left(1 - e^{-\frac{ t-a }{b}} \right) \right]$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{e^{-\frac{ t-a }{b}}}{b \left[2 - \left(1 + \text{sign}(t-a) \left(1 - e^{-\frac{ t-a }{b}} \right) \right) \right]}$	
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \in \mathbf{R}$	
파라미터		$a \in \mathbf{R}, b > 0$	

Table 35: Laplace 분포함수에 기반한 척도 함수

16.1.20.1 Log-Laplace Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\begin{cases} \frac{1}{b} \frac{cd}{c+d} \left(\frac{t}{b}\right)^{c-1} & \text{for } 0 \leq t \leq b \\ \frac{1}{b} \frac{cd}{c+d} \left(\frac{b}{t}\right)^{d+1} & \text{for } t \geq b \end{cases}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	$\begin{cases} 1 - \frac{d}{c+d} \left(\frac{t}{b}\right)^c & \text{for } 0 \leq t \leq b \\ \frac{c}{c+d} \left(\frac{b}{t}\right)^d & \text{for } t \geq b \end{cases}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\begin{cases} \frac{cd}{b} \left(\frac{t}{b}\right)^{c-1} [c+d-d\left(\frac{t}{b}\right)^c]^{-1} & \text{for } 0 \leq t \leq b \\ \frac{cd}{b} \left(\frac{b}{t}\right)^{d+1} \left[c\left(\frac{b}{t}\right)^d\right]^{-1} & \text{for } t \geq b \end{cases}$	
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T>t] = \frac{\int_t^\infty S(u) du}{S(t)}$	Complicated closed form	
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq 0$	
파라메터		$b > 0, c > 0, d > 0$	

Table 36: Log-Laplace 분포함수에 기반한 척도 함수

16.1.21 Linear Hazard Rate Distribution(선형 증가 분포)

Table 37: 선형증가 분포함수에 기반한 척도 함수

척도 함수	기호	내용
수명분포함수	$f(t)$	$(\alpha t + \beta)e^{-(\frac{1}{2}\alpha t^2 + \beta t)}$
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	$1 - e^{-(\frac{1}{2}\alpha t^2 + \beta t)}$
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{-\int_0^t (\alpha u + \beta) du} = e^{-(\frac{1}{2}\alpha t^2 + \beta t)}$
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\alpha t + \beta$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form
k 차 적률	$E(T^k)$	작성중
100p 백분위수	t_p	작성중
변수		$t \geq 0$
파라미터		$\alpha \geq 0, \beta > 0$

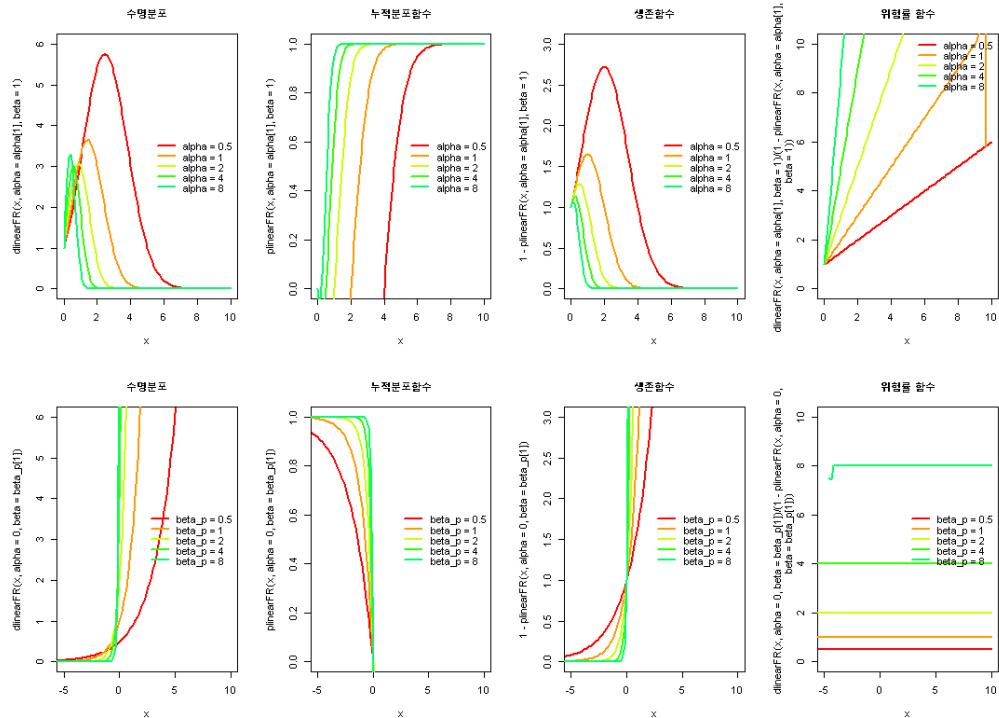


Figure 25: 선형증가분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 10. 선형증가분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```
1 ##### 선형 증가 분포
2 dlinearFR = function (x, alpha = 0, beta = 1, log = FALSE)
3 {
4   fx <- (alpha * x + beta) * exp(-(1/2) * alpha * x^2 + beta * x)
5   if (log)
6     return(log(fx))
7   else return(fx)
8 }
9
10 plinearFR = function (q, alpha = 0, beta = 1, lower.tail = TRUE, log.p = FALSE)
11 {
12   Fx <- 1 - exp(-(1/2) * alpha * x^2 + beta * x)
13   if (!lower.tail)
14     Fx <- 1 - Fx
15   if (log.p)
16     Fx <- log(Fx)
17   return(Fx)
18 }
19
20 # 검증 필요
21 # qlinearFR = function (p, scale = 1, location = 0, lower.tail = TRUE, log.p = FALSE)
22 # {
23   # if (log.p)
24   #   p <- exp(p)
25   # if (!lower.tail)
26   #   p <- 1 - p
27   # xF <- location - scale * log(-log(p))
28   # return(xF)
29 # }
30
31 # rlinearFR = function (n, alpha = 0, beta = 1)
32 # {
33 #   qlinearFR(runif(n), scale, location)
34 # }
35
36
37 par(mfrow = c(2, 4))
38
39 ### parameter: alpha
40 alpha = c(0.5, 1, 2, 4, 8) # shape
41
42 ### Input Variable
43 x <- seq(0, 10, length.out = 101)
44
45 color = rainbow(10)
46
47 ### Life Distribution
48 plot(x, dlinearFR(x, alpha=alpha[1], beta=1), xlim=c(0, 10), ylim=c(0, 6), col=color[1], lwd=2, type = 'l', main="Life
  Distribution")
49 for (i in 2:5) { lines(x, dlinearFR(x, alpha=alpha[i], beta=1), col=color[i], lwd=2); }
50 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('alpha = 0.5', 'alpha = 1', 'alpha = 2', 'alpha = 4', 'alpha = 8'))
51
52 ### Cumulative Distribution
53 plot(x, plinearFR(x, alpha=alpha[1], beta=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative
  Distribution")
54 for (i in 2:5) { lines(x, plinearFR(x, alpha=alpha[i], beta=1), col=color[i], lwd=2); }
55 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('alpha = 0.5', 'alpha = 1', 'alpha = 2', 'alpha = 4', 'alpha = 8'))
56
57 ### Survival Function
58 plot(x, 1-plinearFR(x, alpha=alpha[1], beta=1), xlim=c(0, 10), ylim=c(0, 3), col=color[1], lwd=2, type = 'l', main="Survival
  Function")
59 for (i in 2:5) { lines(x, 1-plinearFR(x, alpha=alpha[i], beta=1), col=color[i], lwd=2); }
60 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('alpha = 0.5', 'alpha = 1', 'alpha = 2', 'alpha = 4', 'alpha = 8'))
61
62 ### Hazard Function
63 plot(x, dlinearFR(x, alpha=alpha[1], beta=1)/(1-plinearFR(x, alpha=alpha[1], beta=1)), xlim=c(0, 10), ylim=c(0, 10), col=color[1],
  lwd=2, type = 'l', main="Hazard Function")
64 for (i in 2:5) { lines(x, dlinearFR(x, alpha=alpha[i], beta=1)/(1-plinearFR(x, alpha=alpha[i], beta=1)), col=color[i], lwd=2); }
65 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('alpha = 0.5', 'alpha = 1', 'alpha = 2', 'alpha = 4', 'alpha = 8'))
  )
66
67
68
69 ### parameter: beta_p
70 beta_p = c(0.5, 1, 2, 4, 8) #beta
71
72 ### Input Variable
73 x <- seq(-10, 10, length.out = 101)
74
75 color = rainbow(10)
76
77 ### Life Distribution
78 plot(x, dlinearFR(x, alpha=0, beta=beta_p[1]), xlim=c(-5, 10), ylim=c(0, 6), col=color[1], lwd=2, type = 'l', main="Life
  Distribution")
79 for (i in 2:5) { lines(x, dlinearFR(x, alpha=0, beta=beta_p[i]), col=color[i], lwd=2); }
80 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8')
```

```

81      ')))
82
83  #### Cumulative Distribution
84  plot(x, plinearFR(x, alpha=0, beta=beta_p[1]), xlim=c(-5, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative
85  Distribution")
86  for (i in 2:5) { lines(x, plinearFR(x, alpha=0, beta=beta_p[i]), col=color[i], lwd=2); }
87  legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8
88  '))
89
90  #### Survival Function
91  plot(x, 1-plinearFR(x, alpha=0, beta=beta_p[1]), xlim=c(-5, 10), ylim=c(0, 3), col=color[1], lwd=2, type = 'l', main="Survival
92  Function")
93  for (i in 2:5) { lines(x, 1-plinearFR(x, alpha=0, beta=beta_p[i]), col=color[i], lwd=2); }
94  legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8
95  '))
96
97  #### Hazard Function
98  plot(x, dlinearFR(x, alpha=0, beta=beta_p[1])/(1-plinearFR(x, alpha=0, beta=beta_p[1])), xlim=c(-5, 10), ylim=c(0, 10), col=color
99  [1], lwd=2, type = 'l', main="Hazard Function")
100 for (i in 2:5) { lines(x, dlinearFR(x, alpha=0, beta=beta_p[i])/(1-plinearFR(x, alpha=0, beta=beta_p[i])), col=color[i], lwd=2); }
101 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('beta_p = 0.5', 'beta_p = 1', 'beta_p = 2', 'beta_p = 4', 'beta_p = 8
102 '))

```

수명을 대표하는 연속확률변수 T 의 위험률 함수(고장률 함수)가 선형증가함수라고 가정한 경우이다.

만일 $\alpha = 0$ 이면, 이 분포는 지수분포가 된다.

만일 $\beta = 0$ 인 경우,

$$\begin{aligned}
 h(t) &= \alpha t \\
 S(t) &= e^{-\frac{1}{2}\alpha t^2} \\
 f(t) &= \alpha t e^{-\frac{1}{2}\alpha t^2} \\
 E(t) &= \frac{\sqrt{\pi}}{\sqrt{2\alpha}} \\
 Var(t) &= \frac{2}{\alpha} \left(1 - \frac{\pi}{4}\right)
 \end{aligned}$$

16.1.21.1 Generalized Linear Hazard Rate Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$c(\alpha + \beta t) \left[1 - e^{-(\alpha t + \frac{\beta}{2} t^2)}\right]^{c-1} e^{-(\alpha t + \frac{\beta}{2} t^2)}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - \left[1 - e^{-(\alpha t + \frac{\beta}{2} t^2)}\right]^c$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{c(\alpha + \beta t) e^{-(\alpha t + \frac{\beta}{2} t^2)}}{\left[1 - e^{-(\alpha t + \frac{\beta}{2} t^2)}\right] \left[\left(1 - e^{-(\alpha t + \frac{\beta}{2} t^2)}\right)^{-c} - 1\right]}$	
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq 0$	
파라메터		$\alpha \geq 0, \beta \geq 0$, but not $\alpha = \beta = 0, c > 0$	

Table 38: Generalized linear hazard rate 분포함수에 기반한 척도 함수

16.1.22 Logistic Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{b} \frac{e^{\frac{t-a}{b}}}{\left[1+e^{\frac{t-a}{b}}\right]^2}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\left[1 + e^{\frac{t-a}{b}}\right]^{-1}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{1}{b} \left[1 + e^{-\frac{t-a}{b}}\right]^{-1}$	IHR with $\lim_{t \rightarrow \infty} h(t) = \frac{1}{b}$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DMRL with $\lim_{t \rightarrow \infty} m(t) = b$
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \in \mathbf{R}$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 39: Logistic 분포함수에 기반한 척도 함수

16.1.22.1 Log-logistic Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{c}{b} \left(\frac{t-a}{b}\right)^{c-1} \left[1 + \left(\frac{t-a}{b}\right)^c\right]^{-2}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$[1 + \left(\frac{t-a}{b}\right)^c]^{-1}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{c}{b} \left(\frac{t-a}{b}\right)^{c-1} \left[1 + \left(\frac{t-a}{b}\right)^c\right]^{-1}$	
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0, c > 0$	

Table 40: Log-logistic 분포함수에 기반한 척도 함수

16.1.22.2 Half-logistic Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{2e^{\frac{t-a}{b}}}{b \left[1 + e^{\frac{t-a}{b}} \right]^2}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$2 \left[1 + e^{\frac{t-a}{b}} \right]^{-1}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\left[b \left(1 + e^{\frac{t-a}{b}} \right) \right]^{-1}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DMRL
k 차 적률	$E(T^k)$	작성중	
100 p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 41: Half-logistic 분포함수에 기반한 척도 함수

16.1.22.3 Generalized Logistic Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{B(c,d)} \frac{e^{-\frac{dt}{b}}}{b} \left[1+e^{-\frac{t}{b}}\right]^{c+d}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	$1 - I_{\left[1+e^{-\frac{t}{b}}\right]^{-1}}(c, d)$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	No closed form	IHR with $\lim_{t \rightarrow \infty} h(t) = \frac{d}{b}$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t] = \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DMRL with $\lim_{t \rightarrow \infty} m(t) = \frac{d}{b}$
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \in \mathbf{R}$	
파라메터		$b > 0, c > 0, d > 0$	

Table 42: Generalized logistic 분포함수에 기반한 척도 함수

16.1.23 Lomax Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{c}{b} \left(1 + \frac{t-a}{b}\right)^{-(c+1)}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\left(1 + \frac{t-a}{b}\right)^{-c}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{c}{t-a-b}$	DHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	$\begin{cases} \text{Does not exist for } 0 < c \leq 1 \\ \frac{t-a-b}{c-1} \text{ for } c > 1 \end{cases}$	IMRL for $c > 1$
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0, c > 0$	

Table 43: Lomax 분포함수에 기반한 척도 함수

16.1.23.1 Generalized Lomax Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{b} [1 + c \frac{t-a}{b}]^{-(1+\frac{1}{c})}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$(1 + c \frac{t-a}{b})^{-\frac{1}{c}}$ for $c \neq 0$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{1}{b} (1 + c \frac{t-a}{b})^{-1}$	IHR for $c < 0$ DHR for $c > 0$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	$\begin{cases} \frac{\int_t^\infty S(u) du}{S(t)} \text{ or } \frac{\int_t^\infty S(u) du}{S(t)}^{\frac{a-b}{c}} \\ \text{존재하지 않음 for } c \geq 1 \end{cases}$	DMRL for $c < 0$ IMRL or IDMRL for $0 < c < 1$
k 차 적률	$E(T^k)$	작성중	
$100p$ 백분위수	t_p	작성중	
변수		$\begin{cases} a \leq t \leq a - \frac{b}{c} & \text{for } c < 0 \\ t \geq a & \text{for } c > 0 \end{cases}$	
파라메터		$a \in \mathbf{R}, b > 0, c \in \mathbf{R}$	

Table 44: Generalized Lomax 분포함수에 기반한 척도 함수

이 분포는 generalized Pareto distribution of the second kind로도 알려져 있다.

또한 $\lim_{c \rightarrow 0} f(t) = \left(\frac{1}{b} e^{-\frac{t-a}{b}}\right)^{-\frac{1}{c}}$ 가 되며, 이는 exponential distribution이다.

16.1.24 Makeham Distribution(메이크햄 분포)

Table 45: 메이크햄 분포함수에 기반한 척도 함수

척도 함수	기호	내용
수명분포함수	$f(t)$	$[1 + \theta (1 - e^{-t})] e^{-[t+\theta(t+e^{-t}-1)]}$
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{-[t+\theta(t+e^{-t}-1)]}$
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$1 + \theta (1 - e^{-t})$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	작성중
k 차 적률	$E(T^k)$	작성중
100p 백분위수	t_p	작성중
변수		$t \geq 0$
파라미터		$\theta \geq 0$

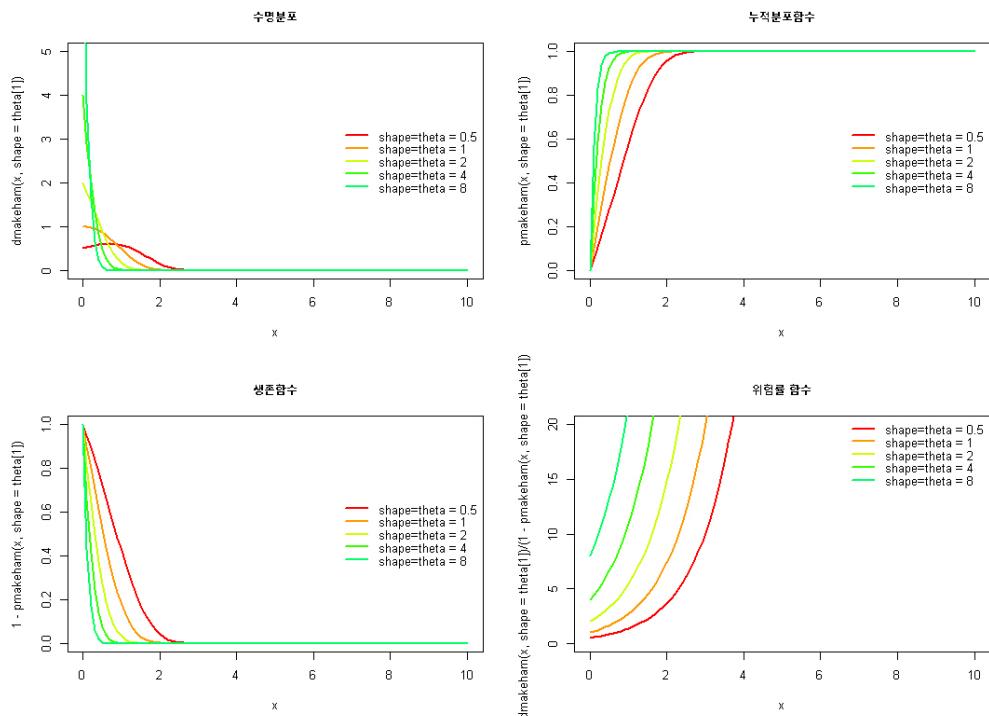


Figure 26: 메이크햄분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 11. 메이크햄분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```
1 ##### 메이크햄 분포
2 library(VGAM)
3
4 ### parameter
5 theta = c(0.5, 1, 2, 4, 8) # theta
6
7 ### Input Variable
8 x <- seq(0, 10, length.out = 101)
9
10
11 color = rainbow(10)
12 par(mfrow = c(2, 2))
13
14 ### Life Distribution
15 plot(x, dmakeham(x, shape=theta[1]), xlim=c(0, 10), ylim=c(0, 5), col=color[1], lwd=2, type = 'l', main="Life Distribution")
16 for (i in 2:5) { lines(x, dmakeham(x, shape=theta[i]), col=color[i], lwd=2); }
17 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('shape=theta = 0.5', 'shape=theta = 1', 'shape=theta = 2', 'shape=
    theta = 4', 'shape=theta = 8'))
18
19 ### Cumulative Distribution
20 plot(x, pmakeham(x, shape=theta[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative Distribution")
21 for (i in 2:5) { lines(x, pmakeham(x, shape=theta[i]), col=color[i], lwd=2); }
22 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('shape=theta = 0.5', 'shape=theta = 1', 'shape=theta = 2',
    'shape=theta = 4', 'shape=theta = 8'))
23
24 ### Survival Function
25 plot(x, 1-pmakeham(x, shape=theta[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
26 for (i in 2:5) { lines(x, 1-pmakeham(x, shape=theta[i]), col=color[i], lwd=2); }
27 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('shape=theta = 0.5', 'shape=theta = 1', 'shape=theta = 2',
    'shape=theta = 4', 'shape=theta = 8'))
28
29 ### Hazard Function
30 plot(x, dmakeham(x, shape=theta[1])/(1-pmakeham(x, shape=theta[1])), xlim=c(0, 10), ylim=c(0, 20), col=color[1], lwd=2, type = 'l',
    main="Hazard Function")
31 for (i in 2:5) { lines(x, dmakeham(x, shape=theta[i])/(1-pmakeham(x, shape=theta[i])), col=color[i], lwd=2); }
32 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('shape=theta = 0.5', 'shape=theta = 1', 'shape=theta = 2',
    'shape=theta = 4', 'shape=theta = 8'))
```

메이크햄 분포는 비모수적 검정 방법의 검정력 비교에 자주 이용된다.

16.1.25 Maxwell-Boltzmann Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{b} \sqrt{\frac{2}{\pi}} \left(\frac{t-a}{b}\right)^2 e^{-\frac{1}{2}\left(\frac{t-a}{b}\right)^2}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - F_{\chi_3^2}\left[\left(\frac{t-a}{b}\right)^2\right]$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	No closed form	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 46: Maxwell-Boltzmann 분포함수에 기반한 척도 함수

여기서, $F_{\chi_3^2}[\cdot]$ 은 자유도가 3인 χ^2 분포의 CDF

16.1.26 Muth Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{b} \left[e^{c(\frac{t-a}{b})} - c \right] e^{-\frac{1}{c}e^{c(\frac{t-a}{b})} + c(\frac{t-a}{b}) + \frac{1}{c}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{-\frac{1}{c}e^{c(\frac{t-a}{b})} + c(\frac{t-a}{b}) + \frac{1}{c}}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{1}{b} \left[e^{c(\frac{t-a}{b})} - c \right]$	IHR with $h(a) = \frac{1-c}{b} \forall c$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0, 0 < c \leq 1$	

Table 47: Much 분포함수에 기반한 척도 함수

16.1.27 Normal(Gaussian) Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{\sqrt{2\pi}b} e^{-\frac{(t-a)^2}{2b^2}} = \frac{1}{b}\phi\left(\frac{t-a}{b}\right)$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	$1 - \Phi\left(\frac{t-a}{b}\right) = \Phi\left(\frac{a-t}{b}\right)$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{1}{b}\phi\left(\frac{t-a}{b}\right) [\Phi\left(\frac{a-t}{b}\right)]^{-1}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t] = \frac{\int_t^\infty S(u)du}{S(t)}$	$a + b^2 h(t) - t$	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \in \mathbf{R}$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 48: Normal 분포함수에 기반한 척도 함수

16.1.27.1 Log-normal Distribution(대수 정규 분포, 로그 정규 분포)

Table 49: 로그정규분포함수에 기반한 척도 함수

척도 함수	기호	내용
수명분포함수	$f(t)$	$\psi\left(\frac{\ln t - \mu}{\sigma}\right) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln t - \mu}{\sigma}\right)^2}$
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	$\Phi\left(\frac{\ln t - \mu}{\sigma}\right)$
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	$1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{\psi\left(\frac{\ln t - \mu}{\sigma}\right)}{t\sigma[1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)]}$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중
평균수명	$E[T] = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt$	$e^{\left(\frac{\mu + \sigma^2}{2}\right)}$
평균잔여수명함수	$m(t) = E[T - t T > t] = \frac{\int_t^\infty S(u)du}{S(t)}$	작성중
k 차 적률	$E(T^k)$	작성중
100p 백분위수	t_p	작성중
변수		$t \geq 0$
파라미터	μ (위치 모수; location parameter), σ (척도 모수; scale parameter)	

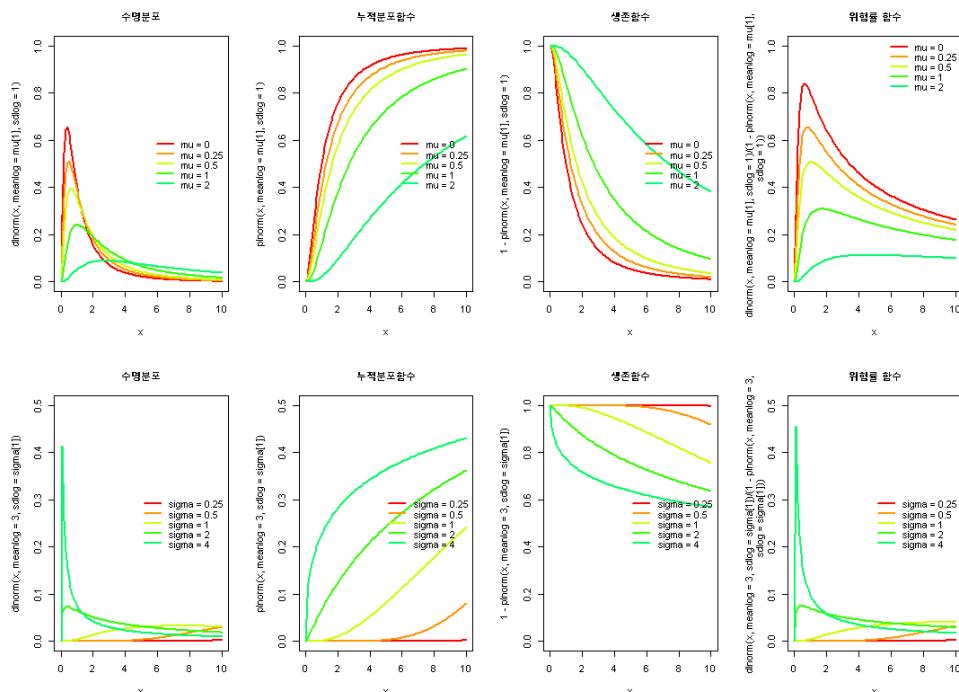


Figure 27: 로그정규분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 12. 로그정규분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```
1 ##### 로그 정규 분포
2 par(mfrow = c(2, 4))
3
4 ### parameter: mu
5 mu = c(0, 0.25, 0.5, 1, 2) # meanlog
6
7 ### Input Variable
8 x <- seq(0, 10, length.out = 101)
9
10 color = rainbow(10)
11
12 ### Life Distribution
13 plot(x, dlnorm(x, meanlog=mu[1], sdlog=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Life Distribution")
14 for (i in 2:5) { lines(x, dlnorm(x, meanlog=mu[i], sdlog=1), col=color[i], lwd=2); }
15 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0', 'mu = 0.25', 'mu = 0.5', 'mu = 1', 'mu = 2'))
16
17 ### Cumulative Distribution
18 plot(x, plnorm(x, meanlog=mu[1], sdlog=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative Distribution")
19 for (i in 2:5) { lines(x, plnorm(x, meanlog=mu[i], sdlog=1), col=color[i], lwd=2); }
20 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0', 'mu = 0.25', 'mu = 0.5', 'mu = 1', 'mu = 2'))
21
22 ### Survival Function
23 plot(x, 1-plnorm(x, meanlog=mu[1], sdlog=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
24 for (i in 2:5) { lines(x, 1-plnorm(x, meanlog=mu[i], sdlog=1), col=color[i], lwd=2); }
25 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0', 'mu = 0.25', 'mu = 0.5', 'mu = 1', 'mu = 2'))
26
27 ### Hazard Function
28 plot(x, dlnorm(x, meanlog=mu[1], sdlog=1)/(1-plnorm(x, meanlog=mu[1], sdlog=1)), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2,
      type = 'l', main="Hazard Function")
29 for (i in 2:5) { lines(x, dlnorm(x, meanlog=mu[i], sdlog=1)/(1-plnorm(x, meanlog=mu[i], sdlog=1)), col=color[i], lwd=2); }
30 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('mu = 0', 'mu = 0.25', 'mu = 0.5', 'mu = 1', 'mu = 2'))
31
32
33
34 ### parameter: sigma
35 sigma = c(0.25, 0.5, 1, 2, 4) #sdlog
36
37 ### Input Variable
38 x <- seq(0, 10, length.out = 101)
39
40 color = rainbow(10)
41
42 ### Life Distribution
43 plot(x, dlnorm(x, meanlog=3, sdlog=sigma[1]), xlim=c(0, 10), ylim=c(0, 0.5), col=color[1], lwd=2, type = 'l', main="Life Distribution")
44 for (i in 2:5) { lines(x, dlnorm(x, meanlog=3, sdlog=sigma[i]), col=color[i], lwd=2); }
45 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('sigma = 0.25', 'sigma = 0.5', 'sigma = 1', 'sigma = 2', 'sigma = 4'))
46
47 ### Cumulative Distribution
48 plot(x, plnorm(x, meanlog=3, sdlog=sigma[1]), xlim=c(0, 10), ylim=c(0, 0.5), col=color[1], lwd=2, type = 'l', main="Cumulative Distribution")
49 for (i in 2:5) { lines(x, plnorm(x, meanlog=3, sdlog=sigma[i]), col=color[i], lwd=2); }
50 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('sigma = 0.25', 'sigma = 0.5', 'sigma = 1', 'sigma = 2', 'sigma = 4'))
51
52 ### Survival Function
53 plot(x, 1-plnorm(x, meanlog=3, sdlog=sigma[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
54 for (i in 2:5) { lines(x, 1-plnorm(x, meanlog=3, sdlog=sigma[i]), col=color[i], lwd=2); }
55 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('sigma = 0.25', 'sigma = 0.5', 'sigma = 1', 'sigma = 2', 'sigma = 4'))
56
57 ### Hazard Function
58 plot(x, dlnorm(x, meanlog=3, sdlog=sigma[1])/(1-plnorm(x, meanlog=3, sdlog=sigma[1])), xlim=c(0, 10), ylim=c(0, 0.5), col=color[1], lwd=2, type = 'l', main="Hazard Function")
59 for (i in 2:5) { lines(x, dlnorm(x, meanlog=3, sdlog=sigma[i])/(1-plnorm(x, meanlog=3, sdlog=sigma[i])), col=color[i], lwd=2); }
60 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('sigma = 0.25', 'sigma = 0.5', 'sigma = 1', 'sigma = 2', 'sigma = 4'))
```

로그정규분포는 고장(재발)시간을 모형화하는데 자주 사용하는 분포이다.

로그정규분포는 정규분포와 밀접한 관련이 있는데, 만일 $T \sim N()$ 을 따른다면, $\ln T \sim N()$ 이 된다.

로그정규분포는 척도 모수 σ 가 커지면, 확률밀도함수의 그래프가 왼쪽으로 치우치는 경향이 있고, 척도 모수 σ 가 작아지면, 확률밀도함수의 그래프가 종 모양의 그래프가 되어간다.

이는, 척도 모수 σ 의 값이 크면, ”초기에” 고장(재발) 발생률이 높다는 것을 의미한다.

로그정규분포의 응용 분야는 다음과 같다.

- 동일하고 독립적인 분포를 하는 양적인 확률변수들의 곱으로 표현되는 확률변수를 나타낸다.
- 여러 개의 요인이 합의 형태가 아닌 곱이나 비례의 형태로 열화과정이 진행되는 경우를 모형화하는 데 적합하다.
- 금속의 피로(fatigue), 반도체나 다이오드와 같은 부속이나 전기절연체의 수명자료 분석에 광범위하게 사용된다.
- 모집단에서 작은 부분을 차지하는 불량품에 의해 고장률(위험률)이 감소하는 부품의 고장 시간을 모형화하는 데 적합하다.

16.1.27.2 Log-normal Distribution with Lower Threshold

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$e^{-\frac{[\ln(t-a)-\alpha]^2}{2\beta^2}} [\beta(t-a)\sqrt{2\pi}]^{-1}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\Phi\left[-\frac{\ln(t-a)-\alpha}{\beta}\right]$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\phi\left[-\frac{\ln(t-a)-\alpha}{\beta}\right] \left[\beta(t-a)\Phi\left(-\frac{\ln(t-a)-\alpha}{\beta}\right)\right]^{-1}$	IDHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty tf(t)dt$ $= \int_0^\infty S(t)dt$	작성중	
평균잔여수명함수	$m(t) = \frac{E[T-t T>t]}{\int_t^\infty S(u)du}$ $= \frac{E(T^k)}{S(t)}$	No closed form	DIMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, \alpha \in \mathbf{R}, \beta > 0$	

Table 50: Log-normal with lower threshold 분포함수에 기반한 척도 함수

16.1.27.3 Log-normal Distribution with Upper Threshold

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$e^{-\frac{[\ln(a-t)-\alpha]^2}{2\beta^2}} [\beta(a-t)\sqrt{2\pi}]^{-1}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\Phi\left[-\frac{\ln(a-t)-\alpha}{\beta}\right]$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\phi\left[-\frac{\ln(a-t)-\alpha}{\beta}\right] \left[\beta(a-t)\Phi\left(-\frac{\ln(a-t)-\alpha}{\beta}\right)\right]^{-1}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty tf(t)dt$ $= \int_0^\infty S(t)dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u)du}{S(t)}$	No closed form	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t < a$	
파라미터		$a \in \mathbf{R}, \alpha \in \mathbf{R}, \beta > 0$	

Table 51: Log-normal with upper threshold 분포함수에 기반한 척도 함수

16.1.27.4 Inverse Normal(Gaussian) Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\sqrt{\frac{b}{2\pi t^3}} e^{-\frac{b(t-a)^2}{2a^2 t}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\Phi\left[\sqrt{\frac{b}{t}}\left(1 - \frac{t}{a}\right)\right] - e^{\frac{2b}{a}} \Phi\left[-\sqrt{\frac{b}{t}}\left(\frac{t}{a} + 1\right)\right]$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	No closed form	IDHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DIMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t > 0$	
파라메터		$a > 0, b > 0$	

Table 52: Inverse-normal 분포함수에 기반한 척도 함수

이 분포는 Wald distribution이라고도 알려져 있다.

16.1.27.5 Half-normal Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{b} \sqrt{\frac{2}{\pi}} e^{-\frac{(t-a)^2}{2b^2}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$2\Phi\left(\frac{a-t}{b}\right)$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{1}{2\sqrt{2\pi}} \frac{e^{-\frac{(t-a)^2}{2b^2}}}{\Phi\left(\frac{a-t}{b}\right)}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 53: Half-normal 분포함수에 기반한 척도 함수

16.1.27.6 Trimmed(Truncated) Normal Distribution(절사 정규 분포)

Table 54: 절사정규분포 함수에 기반한 척도 함수

척도 함수	기호	내용
수명분포함수	$f(t)$	$\begin{cases} 0 & t < t_0 \\ \frac{1-\Phi(\frac{t-\mu}{\sigma})}{\sqrt{2\pi}\sigma[1-\Phi(\frac{t_0-\mu}{\sigma})]} & t \geq 0 \end{cases}$
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중
생존함수 (신뢰도함수)	$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \leq t) \\ &= 1 - F(t) \\ &= e^{-H(t)} \end{aligned}$	$\frac{1-\Phi(\frac{t-\mu}{\sigma})}{1-\Phi(\frac{t_0-\mu}{\sigma})} \quad t \geq 0$
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})}}{\sqrt{2\pi}\sigma[1-\Phi(\frac{t-\mu}{\sigma})]} \quad t \geq 0$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중
평균수명	$\begin{aligned} E[T] &= \int_{0}^{\infty} tf(t)dt \\ &= \int_{0}^{\infty} S(t)dt \end{aligned}$	작성중
평균잔여수명함수	$\begin{aligned} m(t) &= E[T - t T > t] \\ &= \frac{\int_t^{\infty} S(u)du}{S(t)} \end{aligned}$	작성중
k 차 적률	$E(T^k)$	작성중
100p 백분위수	t_p	작성중
변수		$-\infty \leq t \leq \infty$
파라메터		$\mu, \sigma > 0$

Code 13. 절사정규분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```

1 ##### 절사 정규 분포 (작성중)
2 require(truncnorm)
3 dtruncnorm(x, a=-Inf, b=Inf, mean = 0, sd = 1)
4 ptruncnorm(q, a=-Inf, b=Inf, mean = 0, sd = 1)
5 qtruncnorm(p, a=-Inf, b=Inf, mean = 0, sd = 1)
6 rtruncnorm(n, a=-Inf, b=Inf, mean = 0, sd = 1)
7 etruncnorm(a=-Inf, b=Inf, mean=0, sd=1)
8 vtruncnorm(a=-Inf, b=Inf, mean=0, sd=1)

```

16.1.28 Parabolic U-shaped Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{3}{2b} \left(\frac{t-a}{b}\right)^2$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	$\frac{1}{2} \left[1 - \left(\frac{t-a}{b} \right)^3 \right]$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{3(a-t)^2}{b^3 + (a-t)^3}$	DIHR with minimum at $t^* = a$ and $h(t^*) = 0$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t] = \frac{\int_t^\infty S(u) du}{S(t)}$	$\frac{\frac{3}{4}b^4 - b^3 t + \frac{1}{4}t^4}{b^3 - t^3}$	IDMRL with maximum at $t^* \approx a - 0.596072b$
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$a - b \leq t \leq a + b$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 55: Parabolic U-shaped 분포함수에 기반한 척도 함수

16.1.28.1 Parabolic Inverted U-shaped Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{3}{4b} \left[1 - \left(\frac{t-a}{b} \right)^2 \right]$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\frac{1}{2} - \frac{1}{4} \left[3 \left(\frac{t-a}{b} \right) - \left(\frac{t-a}{b} \right)^3 \right]$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	Complicated form	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T>t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	$b \left[0.75 - 2 \left(\frac{t-a}{b} \right)^2 - 0.25 \left(\frac{t-a}{b} \right)^4 \right] \left[\left(2 + \frac{t-a}{b} \right) \left(\frac{t-a}{b} - 1 \right) \right]^{-1}$	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$a - b \leq t \leq a + b$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 56: Parabolic Inverted U-shaped 분포함수에 기반한 척도 함수

16.1.29 Pareto Distribution of the first kind

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{c}{b} \left(\frac{t-a}{b}\right)^{-(c+1)}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\left(\frac{t-a}{b}\right)^{-c}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{c}{t-a}$	DHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	$\left\{ \begin{array}{l} \text{존재하지 않음 for } c \leq 1 \\ \frac{b}{c-1} \left(\frac{t-a}{b}\right) \end{array} \right\}$	IMRL for $c \geq 1$
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a + b$	
파라메터		$a \in \mathbf{R}, b > 0, c > 0$	

Table 57: Pareto 분포함수에 기반한 척도 함수

16.1.29.1 Generalized Pareto Distribution of the first kind

이 분포는 Generalized Lomax 분포로도 알려져 있다. Chapter 16.1.23.1를 참고하도록 하자.

16.1.30 Power Function Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{c}{b} \left(\frac{t-a}{b}\right)^{c-1}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - \left(\frac{t-a}{b}\right)^c$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{c}{a-t} \left[1 + \left[\left(\frac{t-a}{b}\right)^c - 1 \right]^{-1} \right]$	DIHR for $0 < c < 1$ IHR for $c \geq 1$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	$\frac{1}{c+1} \left[\frac{c(t-a-b)}{\left(\frac{t-a}{b}\right)^c - 1} - (t-a) \right]$	IDMRL for $0 < c \lesssim 0.815$ DMRL for $c \gtrsim 0.815$
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$a \leq t \leq a+b$	
파라메터		$a \in \mathbf{R}, b > 0, c > 0$	

Table 58: Power Function 분포함수에 기반한 척도 함수

16.1.31 Rayleigh Distribution(레이일리 분포)

Table 59: 레일리 분포함수에 기반한 척도 함수

척도 함수	기호	내용
수명분포함수	$f(t)$	$kte^{-\frac{kt^2}{2}}$
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	$1 - e^{-\frac{kt^2}{2}}$
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - F(t) = e^{-H(t)}$	$e^{-\frac{kt^2}{2}}$
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	kt
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중
평균수명	$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$	$\sqrt{\frac{\pi}{2k}}$
평균잔여수명함수	$m(t) = E[T - t T > t] = \frac{\int_t^\infty S(u) du}{S(t)}$	작성중
k 차 적률	$E(T^k)$	작성중
100p 백분위수	t_p	작성중
변수	$t \geq 0$	
파라메터	$k > 0$	

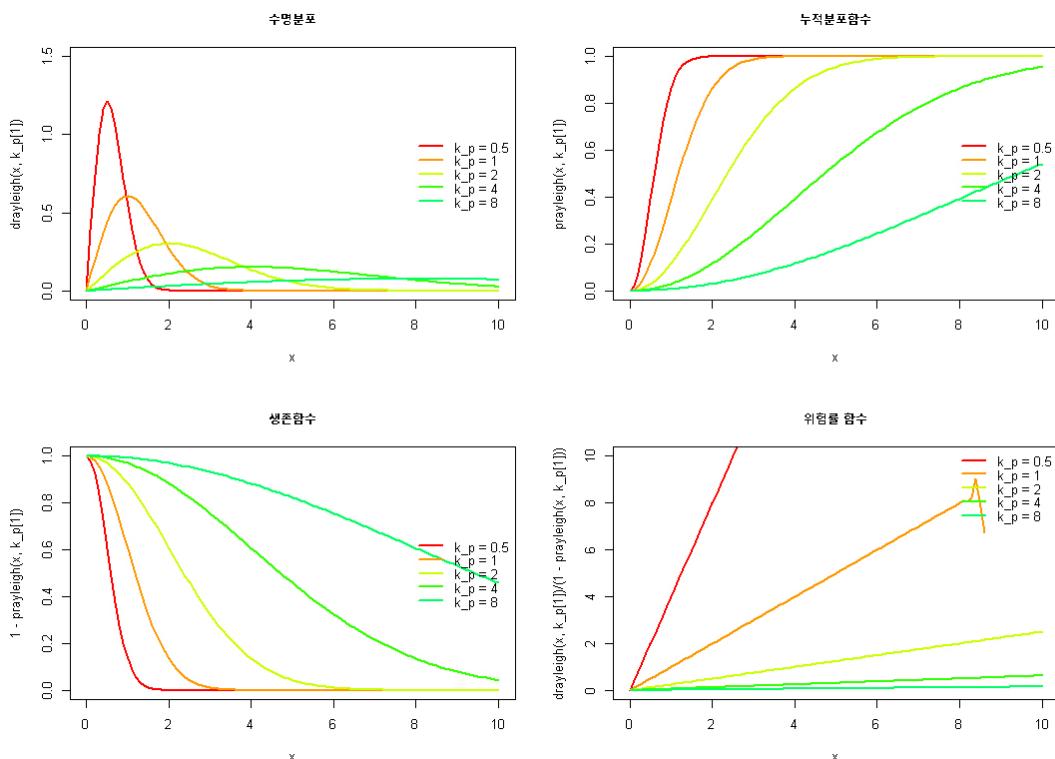


Figure 28: 레일리분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 14. 레일리분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```
1 ##### 레일리 분포
2 library(VGAM)
3
4 ### parameter
5 k_p = c(0.5, 1, 2, 4, 8) # k_p
6
7 ### Input Variable
8 x <- seq(0, 10, length.out = 101)
9
10
11 color = rainbow(10)
12 par(mfrow = c(2, 2))
13
14 ### Life Distribution
15 plot(x, drayleigh(x, k_p[1]), xlim=c(0, 10), ylim=c(0, 1.5), col=color[1], lwd=2, type = 'l', main="Life Distribution")
16 for (i in 2:5) { lines(x, drayleigh(x, k_p[i]), col=color[i], lwd=2); }
17 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('k_p = 0.5', 'k_p = 1', 'k_p = 2', 'k_p = 4', 'k_p = 8'))
18
19 ### Cumulative Distribution
20 plot(x, prayleigh(x, k_p[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative Distribution")
21 for (i in 2:5) { lines(x, prayleigh(x, k_p[i]), col=color[i], lwd=2); }
22 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('k_p = 0.5', 'k_p = 1', 'k_p = 2', 'k_p = 4', 'k_p = 8'))
23
24 ### Survival Function
25 plot(x, 1-prayleigh(x, k_p[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival Function")
26 for (i in 2:5) { lines(x, 1-prayleigh(x, k_p[i]), col=color[i], lwd=2); }
27 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('k_p = 0.5', 'k_p = 1', 'k_p = 2', 'k_p = 4', 'k_p = 8'))
28
29 ### Hazard Function
30 plot(x, drayleigh(x, k_p[1])/(1-prayleigh(x, k_p[1])), xlim=c(0, 10), ylim=c(0, 10), col=color[1], lwd=2, type = 'l', main="Hazard Function")
31 for (i in 2:5) { lines(x, drayleigh(x, k_p[i])/(1-prayleigh(x, k_p[i])), col=color[i], lwd=2); }
32 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('k_p = 0.5', 'k_p = 1', 'k_p = 2', 'k_p = 4', 'k_p = 8'))
```

레일리 분포는 와이블 분포의 특수한 경우로, 와이블 분포의 형상 모수(shape parameter)가 2일 때 레일리 분포가 된다. 레일리분포는 마모 특성을 모형화하는 데 유용한 분포이다.

이 분포의 특징은, 위험률 함수(고장률 함수)가 원점을 지나며, 시간이 지남에 따라 선형적으로 증가한다는 것이다. 따라서, 레일리 분포의 위험률 함수(고장률 함수)는 $r(t) = kt^{\alpha}$ 이다.

16.1.31.1 Inverse Rayleigh Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{2b}{(t-a)^3} e^{-\frac{b}{(t-a)^3}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	$1 - e^{-\frac{b}{(t-a)^3}}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$2b \left[(t-a)^3 \left(e^{\frac{b}{(t-a)^2}} - 1 \right) \right]^{-1}$	IDHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T>t] = \frac{\int_t^\infty S(u) du}{S(t)}$	$(t-a) \left[e^{-\frac{b}{(t-a)^2}} - 1 \right] + \sqrt{b\pi} \cdot \text{erf} \left(\frac{\sqrt{b}}{t-a} \right)$	DIMRL or DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 60: Inverse Rayleigh 분포함수에 기반한 척도 함수

16.1.31.2 Generalized Rayleigh Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$c \frac{t-a}{b^2} e^{-\frac{1}{2} \left(\frac{t-a}{b} \right)^2} \left[1 - e^{-\frac{1}{2} \left(\frac{t-a}{b} \right)^2} \right]^{c-1}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	$1 - \left[e^{-\frac{1}{2} \left(\frac{t-a}{b} \right)^2} \right]^c$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	작성중	DIHR for $0 < c < 0.5$ IHR for $c \geq 0.5$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T>t] = \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	IDMRL for $0 < c < 0.5$ DMRL for $c \geq 0.5$
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0, c > 0$	

Table 61: Generalized Rayleigh 분포함수에 기반한 척도 함수

$c = 1$ 이면 이 분포는 Rayleigh 분포가 된다.

16.1.32 Semi-elliptical Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{2}{b\pi} \sqrt{1 - \left(\frac{t-a}{b}\right)^2}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$\frac{1}{2} - \frac{1}{\pi} \left[\frac{t-a}{b} \sqrt{1 - \left(\frac{t-a}{b}\right)^2} + \arcsin \left(\frac{t-a}{b} \right) \right]$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{4\sqrt{1 - \left(\frac{t-a}{b}\right)^2}}{b \left[\pi - 2 \left(\left(\frac{t-a}{b} \right) \sqrt{1 - \left(\frac{t-a}{b}\right)^2} + \arcsin \left(\frac{t-a}{b} \right) \right) \right]}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$a - b \leq t \leq a + b$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 62: Semi-elliptical 분포함수에 기반한 척도 함수

16.1.33 t Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{\Gamma(\frac{\nu+1}{\nu})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	No closed form	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	No closed form	
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T > t] = \frac{\int_t^\infty S(u)du}{S(t)}$	No closed form	
k 차 적률	$E(T^k)$	작성중	
100 p 백분위수	t_p	작성중	
변수		$t \in \mathbf{R}$	
파라메터		$\nu > 0$	

Table 63: t 분포함수에 기반한 척도 함수

16.1.34 Teisser Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{b} \left[e^{\frac{t-a}{b}} - 1 \right] e^{\left[1 + \frac{t-a}{b} - e^{\frac{t-a}{b}} \right]}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{1 + \frac{t-a}{b} - e^{\frac{t-a}{b}}}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{1}{b} \left[e^{\frac{t-a}{b}} - 1 \right]$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	$\frac{1}{b} e^{-\frac{t-a}{b}}$	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 64: Teisser 분포함수에 기반한 척도 함수

16.1.35 Triangular Distribution(Continuous)

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\begin{cases} \frac{2(t-a)}{cb^2} & \text{for } a \leq t \leq a+cb \\ \frac{2(a+b-x)}{(1-c)b^2} & \text{for } a+cb \leq t \leq a+b \end{cases}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	$\begin{cases} 1 - \frac{(t-a)^2}{cb^2} & \text{for } a \leq t \leq a+cb \\ \frac{(a+b-x)^2}{(1-c)b^2} & \text{for } a+cb \leq t \leq a+b \end{cases}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\begin{cases} \frac{2(t-a)}{cb^2 - (t-a)^2} & \text{for } a \leq t \leq a+cb \\ \frac{2}{a+b-x} & \text{for } a+cb \leq t \leq a+b \end{cases}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T > t] = \frac{\int_t^\infty S(u)du}{S(t)}$	$\begin{cases} b \frac{c+c^2-3c(\frac{t-a}{b})+(\frac{t-a}{b})^3}{3[c-(\frac{t-a}{b})^2]} & \text{for } 0 \leq \frac{t-a}{b} \leq c \\ b \frac{1-\frac{t-a}{b}}{3} & \text{for } c \leq \frac{t-a}{b} \leq 1 \end{cases}$	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수 파라메터		$t \geq a$ $a \in \mathbf{R}, b > 0, 0 < c < 1$	

Table 65: Triangular 분포함수에 기반한 척도 함수

Triangular distribution은 c 파라메터에 따라 다음과 같은 분포가 된다.

- $c = 0.5$: Symmetric Triangular Distribution
- $c = 0$: Right-angled and Positively Skewed Triangular Distribution
- $c = 1$: Right-angled and Negatively Skewed Triangular Distribution

16.1.36 Uniform Distribution(Continuous)

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{b}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - \frac{t-a}{b}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{1}{a+b-x}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^{\infty} tf(t)dt$ $= \int_0^{\infty} S(t)dt$	작성중	
평균잔여수명함수	$m(t) = E[T-t T > t]$ $= \frac{\int_t^{\infty} S(u)du}{S(t)}$	$\frac{b-t-a}{2}$	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수	$a \leq t \leq a + b$		
파라메터	$a \in \mathbf{R}, b > 0$		

Table 66: Uniform 분포함수에 기반한 척도 함수

16.1.37 V-shaped Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\begin{cases} \frac{2(2a+b-2t)}{b^2} & \text{for } a \leq t \leq a + \frac{b}{2} \\ \frac{2(2t-2a-b)}{b^2} & \text{for } a + \frac{b}{2} \leq t \leq a + b \end{cases}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \leq t) \\ &= 1 - F(t) \\ &= e^{-H(t)} \end{aligned}$	$\begin{cases} 1 - \frac{2(t-a)(a+b-t)}{b^2} & \text{for } a \leq t \leq a + \frac{b}{2} \\ 0.5 - \frac{(2t-2a-b)^2}{b^2} & \text{for } a + \frac{b}{2} \leq t \leq a + b \end{cases}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\begin{cases} \frac{2(2a+b-2t)}{b^2+2(a-t)(a+b-t)} & \text{for } a \leq t \leq a + \frac{b}{2} \\ \frac{2a+b-2t}{(a-t)(a+b-t)} & \text{for } a + \frac{b}{2} \leq t \leq a + b \end{cases}$	DIHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$\begin{aligned} m(t) &= E[T - t T > t] \\ &= \frac{\int_t^\infty S(u) du}{S(t)} \end{aligned}$	$\begin{cases} b \frac{3-2(\frac{t-a}{b})[3+(\frac{t-a}{b})\{2(\frac{t-a}{b})-3\}]}{6[1+2(\frac{t-a}{b})\{(\frac{t-a}{b})-1\}]} & \text{for } 0 \leq (\frac{t-a}{b}) \leq 0.5 \\ b \frac{[(\frac{t-a}{b})-1]^2[(\frac{t-a}{b})+0.5]}{3(\frac{t-a}{b})[1-(\frac{t-a}{b})]} & \text{for } 0.5 \leq (\frac{t-a}{b}) \leq 1 \end{cases}$	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수 파라미터		$\begin{aligned} t &\geq a \\ a &\in \mathbf{R}, b > 0 \end{aligned}$	

Table 67: V-shaped 분포함수에 기반한 척도 함수

16.1.38 Wald Distribution

이 분포는 inverse normal(Gaussian) distribution이라고도 알려져 있다. 자세한 것은 Chapter 16.1.27.4을 참고하도록 하자.

16.1.39 Weibull Distribution(와이블 분포)

지수분포는 무기역성 때문에 응용 분야가 제한될 뿐만 아니라, 위험률(고장률)이 나이에 관계없이 일정하다는 성질이 비현실적이며, 수명자료의 분석에 부적합한 경우도 종종 있다.

와이블 분포는 위험률이 상수인 경우, 증가하는 경우, 감소하는 경우 등을 모형화하는 수명분포로, 다양한 형태의 수명자료를 분석하는 데 널리 활용되고 있다.

16.1.39.1 Weibull Distribution with 2 Parameters(2-모수 와이블 분포)

Table 68: 2모수 와이블 분포함수에 기반한 척도 함수

척도 함수	기호	내용
수명분포함수	$f(t)$	$\frac{\alpha}{\theta^\alpha} t^{\alpha-1} e^{-(\frac{t}{\theta})^\alpha}$
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	$1 - e^{-(\frac{t}{\theta})^\alpha}$
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{-(\frac{t}{\theta})^\alpha}$
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{\alpha}{\theta^\alpha} t^{\alpha-1}$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	$(\frac{t}{\theta})^\alpha$
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(u) du$	$\theta \Gamma(1 + \frac{1}{\alpha})$
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	$\frac{\int_t^\infty e^{-(\frac{u}{\theta})^\alpha} du}{e^{-(\frac{t}{\theta})^\alpha}}$
k 차 적률	$E(T^k)$	$\theta^k \Gamma(1 + \frac{k}{\alpha}) \quad k \geq 1$
100p 백분위수	t_p	$\theta [-\ln(1-p)]^{\frac{1}{\alpha}}$
변수		$t \geq 0$
파라메터		$\alpha > 0$ (형상 모수; shape parameter), $\theta > 0$ (척도 모수; scale parameter)

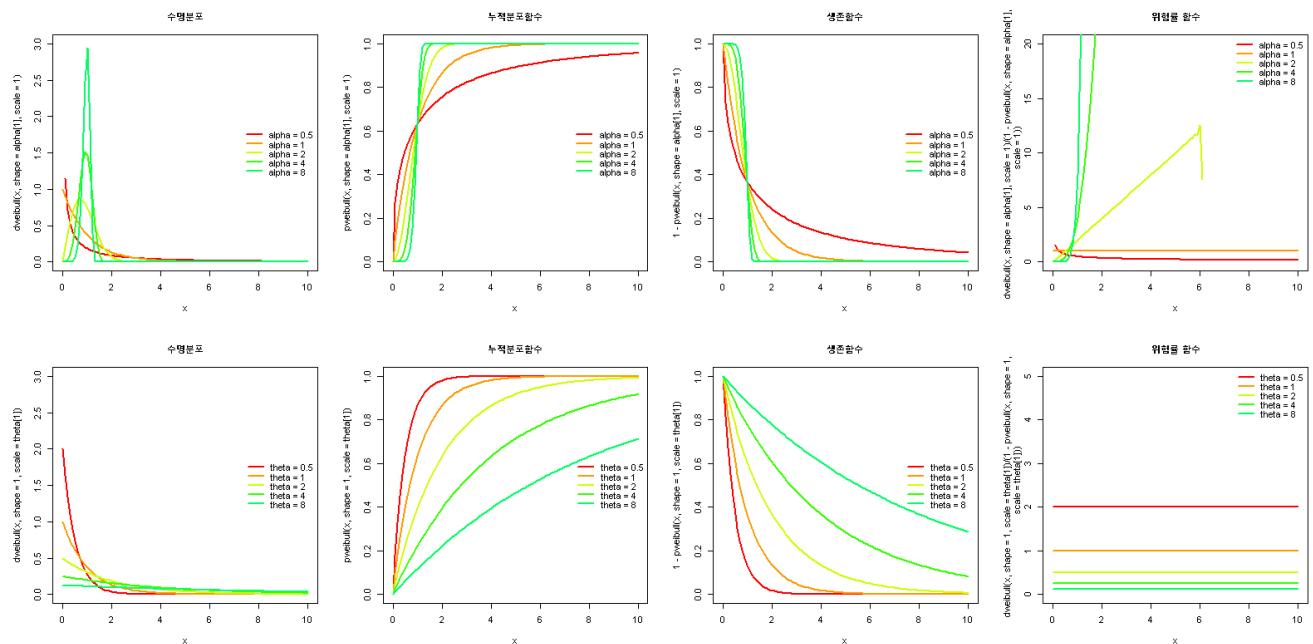


Figure 29: 2모수 와이블분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 15. 2모수 와이블분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```

1 ##### Weibull Distribution with 2 parameters
2 par(mfrow = c(2, 4))
3
4 ### parameter: alpha
5 alpha = c(0.5, 1, 2, 4, 8) # shape
6
7 ### Input Variable
8 x <- seq(0, 10, length.out = 101)
9
10 color = rainbow(10)
11
12 ### Life Distribution
13 plot(x, dweibull(x, shape=alpha[1], scale=1), xlim=c(0, 10), ylim=c(0, 3), col=color[1], lwd=2, type = 'l', main="Life
   Distribution")
14 for (i in 2:5) { lines(x, dweibull(x, shape=alpha[i], scale=1), col=color[i], lwd=2); }
15 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('alpha = 0.5', 'alpha = 1', 'alpha = 2', 'alpha = 4', 'alpha = 8'))
16
17 ### Cumulative Distribution
18 plot(x, pweibull(x, shape=alpha[1], scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative
   Distribution")
19 for (i in 2:5) { lines(x, pweibull(x, shape=alpha[i], scale=1), col=color[i], lwd=2); }
20 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('alpha = 0.5', 'alpha = 1', 'alpha = 2', 'alpha = 4', 'alpha = 8'))
21
22 ### Survival Function
23 plot(x, 1-pweibull(x, shape=alpha[1], scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival
   Function")
24 for (i in 2:5) { lines(x, 1-pweibull(x, shape=alpha[i], scale=1), col=color[i], lwd=2); }
25 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('alpha = 0.5', 'alpha = 1', 'alpha = 2', 'alpha = 4', 'alpha = 8'))
26
27 ### Hazard Function
28 plot(x, dweibull(x, shape=alpha[1], scale=1)/(1-pweibull(x, shape=alpha[1], scale=1)), xlim=c(0, 10), ylim=c(0, 20), col=color[1],
   lwd=2, type = 'l', main="Hazard Function")
29 for (i in 2:5) { lines(x, dweibull(x, shape=alpha[i], scale=1)/(1-pweibull(x, shape=alpha[i], scale=1)), col=color[i], lwd=2); }
30 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('alpha = 0.5', 'alpha = 1', 'alpha = 2', 'alpha = 4', 'alpha = 8'))
31
32
33
34 ### parameter: theta
35 theta = c(0.5, 1, 2, 4, 8) #scale
36
37 ### Input Variable
38 x <- seq(0, 10, length.out = 101)
39
40 color = rainbow(10)
41
42 ### Life Distribution
43 plot(x, dweibull(x, shape=1, scale=theta[1]), xlim=c(0, 10), ylim=c(0, 3), col=color[1], lwd=2, type = 'l', main="Life
   Distribution")
44 for (i in 2:5) { lines(x, dweibull(x, shape=1, scale=theta[i]), col=color[i], lwd=2); }
45 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
46
47 ### Cumulative Distribution
48 plot(x, pweibull(x, shape=1, scale=theta[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative
   Distribution")
49 for (i in 2:5) { lines(x, pweibull(x, shape=1, scale=theta[i]), col=color[i], lwd=2); }
50 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
51
52 ### Survival Function
53 plot(x, 1-pweibull(x, shape=1, scale=theta[1]), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival
   Function")
54 for (i in 2:5) { lines(x, 1-pweibull(x, shape=1, scale=theta[i]), col=color[i], lwd=2); }
55 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
56
57 ### Hazard Function
58 plot(x, dweibull(x, shape=1, scale=theta[1])/(1-pweibull(x, shape=1, scale=theta[1])), xlim=c(0, 10), ylim=c(0, 5), col=color[1],
   lwd=2, type = 'l', main="Hazard Function")
59 for (i in 2:5) { lines(x, dweibull(x, shape=1, scale=theta[i])/(1-pweibull(x, shape=1, scale=theta[i])), col=color[i], lwd=2); }
60 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))

```

척도 모수(scale parameter)는 종종 $\lambda = \frac{1}{\theta}$ 로 표시되는 경우도 있다.

Figure 29를 통해, 형상 모수 α 에 따라 $h(t)$ 의 형태가 바뀌는 모습을 확인할 수 있다.

- $\alpha = 1$: 위험률(고장률) $h(t)$ 가 상수
- $\alpha > 1$: 위험률(고장률) $h(t)$ 가 증가 함수
- $\alpha < 1$: 위험률(고장률) $h(t)$ 가 감소 함수

Note 25. 2모수 와이블분포에 기반한 평균잔여수명함수

$$m(t) = \frac{\int_t^\infty S(u)du}{S(t)} = \frac{\int_t^\infty e^{-(\frac{u}{\theta})^\alpha} du}{e^{-(\frac{t}{\theta})^\alpha}}$$

이 때,

- $\alpha = 1$ 인 경우 $m(t) = 0$
- $\alpha = 2$ 인 경우

$$\begin{aligned} \int_t^\infty e^{-(\frac{u}{\theta})^2} du &= \sqrt{\pi}\theta \int_t^\infty \frac{1}{\sqrt{\pi}\theta} e^{-\frac{u^2}{\theta^2}} du \\ &= \sqrt{\pi}\theta [1 - \Phi(t)] \end{aligned}$$

이므로,

$$m(t) = e^{-(\frac{t}{\theta})^2} \sqrt{\pi}\theta [1 - \Phi(t)]$$

여기서, $\Phi(t) = P(Z \leq t)$: 표준정규분포의 누적분포함수

그러나 일반적인 α 의 값에 대한 $m(t)$ 의 공식은 계산할 수 없으며, 수치적인 방법을 이용하여야 한다.

Note 26. 2모수 와이블분포에 기반한 k 차 적률

$$\begin{aligned} E(T^k) &= \int_0^\infty t^k f(t) dt \\ &= \int_0^\infty t^k \frac{\alpha}{\theta^\alpha} t^{\alpha-1} e^{-(\frac{t}{\theta})^\alpha} dt \\ &= \int_0^\infty t^k \frac{\alpha}{\theta} \left(\frac{t}{\theta}\right)^{\alpha-1} e^{-(\frac{t}{\theta})^\alpha} dt \\ &= \int_0^\infty t^{k+\alpha-1} \frac{1}{\theta^\alpha} \alpha e^{-(\frac{t}{\theta})^\alpha} dt \\ &= \int_0^\infty \left(y^{\frac{1}{\alpha}}\right)^{k+\alpha-1} \frac{1}{\theta^\alpha} \alpha e^{-\frac{y}{\theta^\alpha}} \frac{1}{\alpha \left(y^{\frac{1}{\alpha}}\right)^{\alpha-1}} dy \quad (\text{substituting } t^\alpha = y) \\ &= \frac{1}{\theta^\alpha} \Gamma\left(1 + \frac{k}{\alpha}\right) \theta^{\alpha(1+\frac{k}{\alpha})} \\ &= \theta^k \Gamma\left(1 + \frac{k}{\alpha}\right) \end{aligned}$$

Note 27. 100p 백분위수

$$\begin{aligned} S(t_p) &= e^{-\left(\frac{t_p}{\theta}\right)^\alpha} \\ &= 1 - p \\ \Rightarrow \left(\frac{t_p}{\theta}\right)^\alpha &= -\ln(1 - p) \\ \Rightarrow t_p &= \theta [-\ln(1 - p)]^{\frac{1}{\alpha}} \end{aligned}$$

와이블 분포의 기본적인 성질은 다음과 같다.

1. 만일 T_1, T_2, \dots, T_n 이 서로 독립이고, 각각 형상 모수 α 와 척도 모수 θ 가 동일한 와이블 분포를 가진다고 하자.

이 경우, $T = \min(T_1, T_2, \dots, T_n) \sim Weibull(\alpha, \theta n^{-\frac{1}{\alpha}})$ 가 된다.

Proof: T 의 생존함수 $S(t)$ 가 다음과 같이 계산된다.

$$\begin{aligned} S(t) &= P[\min(T_1, T_2, \dots, T_n) > t] \\ &= P[T_1 > t, T_2 > t, \dots, T_n > t] \\ &= P[T_1 > t]P[T_2 > t] \cdots P[T_n > t] \\ &= e^{-(\frac{t}{\theta})^\alpha} e^{-(\frac{t}{\theta})^\alpha} \cdots e^{-(\frac{t}{\theta})^\alpha} \\ &= e^{-n(\frac{t}{\theta})^\alpha} \\ &= e^{-\left(\frac{n^{\frac{1}{\alpha}} t}{\theta}\right)^\alpha} \sim Weibull(\alpha, \theta n^{-\frac{1}{\alpha}}) \end{aligned}$$

Note 28. 2모수 와이블분포를 따르는 완전자료에 대한 모수 추정

- 점추정

시험 중인 n 개의 부품(환자)에 대한 고장(재발)시간 t_1, t_2, \dots, t_n 이 관측되었다고 가정하고, 이 완전자료를 이용하면, 2모수 와이블분포의 형상모수 α 와 척도모수 θ 의 최대가능도추정량(MLE)은 다음과 같이 계산된다.

$$\begin{aligned} L(t; \alpha, \theta) &= \prod_{i=1}^n \frac{\alpha}{\theta^\alpha} t_i^{\alpha-1} e^{-\sum_{i=1}^n \left(\frac{t_i}{\theta}\right)^\alpha} \\ &= \left(\frac{\alpha}{\theta^\alpha}\right)^n \prod_{i=1}^n t_i^{\alpha-1} e^{-\sum_{i=1}^n \left(\frac{t_i}{\theta}\right)^\alpha} \end{aligned}$$

양변에 자연로그를 취하면

$$\ln L(t; \alpha, \theta) = n(\ln \alpha - \alpha \ln \theta) + (\alpha - 1) \sum_{i=1}^n t_i - \sum_{i=1}^n \left(\frac{t_i}{\theta}\right)^\alpha \quad (8)$$

식 (8)에 대해 α 와 θ 에 대한 1차도함수를 구하고 0으로 놓으면, 다음과 같은 두 개의 방정식을 얻는다.

$$\frac{n}{\alpha} - nl\theta + \sum_{i=1}^n lt_i - \sum_{i=1}^n \left(\frac{t_i}{\theta}\right)^\alpha l\left(\frac{t_i}{\theta}\right) = 0 \quad (9)$$

$$\frac{-n\alpha}{\theta} + \frac{\alpha}{\theta^{\alpha+1}} \sum_{i=1}^n t_i^\alpha = 0 \quad (10)$$

식 (9), (10)을 동시에 만족시키는 α 와 θ 의 값이 MLE가 된다.

우선, 식 (10)를 θ 에 대하여 풀면

$$\hat{\theta} = \left[\frac{\sum_{i=1}^n t_i^\alpha}{n} \right]^{\frac{1}{\alpha}}$$

이 $\hat{\theta}$ 를 식 (9)에 대입하면, 다음과 같은 α 에 대한 방정식을 얻는다.

$$\frac{1}{\alpha} + \frac{1}{n} \sum_{i=1}^n \ln t_i - \frac{\sum_{i=1}^n t_i^\alpha \ln t_i}{\sum_{i=1}^n t_i^\alpha} = 0 \quad (11)$$

식 (11)에 대한 해 $\hat{\alpha}$ 가 MLE이며, Newton-Raphson 방법을 이용하여 수치적으로 계산된다.
일단 $\hat{\alpha}$ 가 결정되면, θ 의 MLE $\hat{\theta}$ 는 다음과 같다.

$$\hat{\theta} = \left[\frac{\sum_{i=1}^n t_i^{\hat{\alpha}}}{n} \right]^{\frac{1}{\hat{\alpha}}}$$

- 구간 추정

$\hat{\alpha}$ 와 $\hat{\theta}$ 의 정확한 분포를 모르는 경우, α 와 θ 에 대한 소표본 신뢰구간을 구할 수 없다.

그러나 대표본인 경우에는 α 와 θ 에 대한 $100(1-\nu)\%$ 신뢰구간이 근사적으로 계산될 수 있다.
우선, 다음과 같이 세팅을 하자.

$$S_0 = \sum_{i=1}^n t_i^{\hat{\alpha}} \quad S_1 = \sum_{i=1}^n t_i^{\hat{\alpha}} (\ln t_i) \quad S_2 = \sum_{i=1}^n t_i^{\hat{\alpha}} (\ln t_i)^2$$

그러면, $\hat{\alpha}$, $\hat{\theta}^{\hat{\alpha}}$ 에 대한 분산은 각각 다음과 같이 근사적으로 계산될 수 있다.

$$\begin{aligned} Var(\hat{\alpha}) &= \frac{\hat{\alpha}^2 S_0^2}{n \left(S_0^2 + \hat{\alpha}^2 S_0 S_2 - \hat{\alpha}^2 S_1^2 \right)} \\ Var(\hat{\theta}^{\hat{\alpha}}) &= \frac{S_0}{n^2} \left(\frac{S_0}{\hat{\alpha}^2} + S_2 \right) Var(\hat{\alpha}) \end{aligned}$$

따라서, α 에 대한 신뢰구간은 다음과 같다.

$$\left[\hat{\alpha} - z_{\frac{\nu}{2}} \sqrt{Var(\hat{\alpha})} < \alpha < \hat{\alpha} + z_{\frac{\nu}{2}} \sqrt{Var(\hat{\alpha})} \right]$$

또한, $\theta^{\hat{\alpha}}$ 에 대한 신뢰구간은 다음과 같다.

$$\left[\hat{\theta}^{\hat{\alpha}} - z_{\frac{\nu}{2}} \sqrt{Var(\hat{\theta}^{\hat{\alpha}})} < \theta^{\hat{\alpha}} < \hat{\theta}^{\hat{\alpha}} + z_{\frac{\nu}{2}} \sqrt{Var(\hat{\theta}^{\hat{\alpha}})} \right]$$

θ 에 대한 신뢰구간은 $\theta = (\theta^{\hat{\alpha}})^{\frac{1}{\hat{\alpha}}}$ 의 관계를 이용하여 구한다.

Note 29. 2모수 와이블분포를 따르는 Type II 우중도절단자료(정수중단자료)에 대한 모수 추정

2모수 와이블 수명분포를 따르는 n 개의 부품(환자)을 가지고 수명시험을 수행하여 $r(\leq n)$ 개의 고장(재발)이 발생한 시점에서 시험을 중단하였을 때, 얻어진 고장(재발)자료를 크기 순서대로 나열하여 $t_{(1)} \leq t_{(2)} \leq \dots \leq t_{(r)}$ 으로 나타낸다. 나머지 $(n - r)$ 개의 부품(환자)에 대한 고장(재발) 자료는 시점 $t_{(r)}$ 에서 결단된다.

Type II 우중도절단자료(정수중단자료)가 주어진 경우의 α 와 θ 에 대한 가능도 함수는 다음과 같다.

$$\begin{aligned} L(t; \theta) &= \frac{n!}{(n-r)!} \left[\prod_{i=1}^r \frac{\alpha}{\theta} \left(\frac{t_{(i)}}{\theta} \right)^{\alpha-1} e^{-\left(\frac{t_{(i)}}{\theta} \right)^\alpha} \right] e^{-\left(\frac{t_{(r)}}{\theta} \right)^\alpha} \\ &= \frac{n!}{(n-r)!} \left(\frac{\alpha}{\theta} \right)^r \prod_{i=1}^r \left(\frac{t_{(i)}}{\theta} \right)^{\alpha-1} e^{-\sum_{i=1}^r \left(\frac{t_{(i)}}{\theta} \right)^\alpha - (n-r) \left(\frac{t_{(r)}}{\theta} \right)^\alpha} \end{aligned}$$

양변에 자연로그를 취하면

$$\ln L(t; \alpha, \theta) = n(\ln \alpha - \alpha \ln \theta) + (\alpha - 1) \sum_{i=1}^r t_i - \sum_{i=1}^r \left(\frac{t_i}{\theta} \right)^\alpha \quad (12)$$

식 (12)에 대해 α 와 θ 에 대한 1차도함수를 구하고 0으로 놓으면, 다음과 같은 두 개의 방정식을 얻는다.

$$-r + \frac{1}{\theta^\alpha} \sum_{i=1}^r t_{(i)}^\alpha + (n-r)t_{(r)}^\alpha = 0 \quad (13)$$

$$\frac{r}{\alpha} + \sum_{i=1}^r t_{(i)}^\alpha - \frac{1}{\theta^\alpha} \sum_{i=1}^r t_{(i)}^\alpha \ln t_{(i)} = (n-r)t_{(i)}^\alpha \ln t_{(r)} = 0 \quad (14)$$

식 (13)로부터 얻어진 θ^α 의 값을 식 (14)에 대입하면

$$\frac{\sum_{i=1}^r t_{(i)}^\alpha \ln t_{(i)} + (n-r)t_{(r)}^\alpha \ln t_{(r)}}{\sum_{i=1}^r t_{(i)}^\alpha + (n-r)t_{(r)}^\alpha} - \frac{1}{r} \sum_{i=1}^r t_{(i)} - \frac{1}{\alpha} = 0 \quad (15)$$

식 (15)을 만족하는 α 의 값이 MLE $\hat{\alpha}$ 가 된다. 이 $\hat{\alpha}$ 의 값은 Newton-Raphson 방법에 의해 수치적으로 찾아질 수 있으며, 일단 $\hat{\alpha}$ 의 값이 결정되면 θ 의 MLE $\hat{\theta}$ 는 다음과 같다.

$$\hat{\theta} = \left[\frac{\sum_{i=1}^r t_{(i)}^{\hat{\alpha}} + (n-r)t_{(r)}^{\hat{\alpha}}}{r} \right]^{\frac{1}{\hat{\alpha}}}$$

참고로 $\hat{\alpha}$, $\hat{\theta}$ 의 분산은 시뮬레이션으로 구할 수 있다.[12]

16.1.39.2 Weibull Distribution with 3 Parameters(3-모수 와이블 분포)

Table 69: 3모수 와이블 분포함수에 기반한 척도 함수

척도 함수	기호	내용
수명분포함수	$f(t)$	$\frac{\alpha}{\theta} \left(\frac{t-\gamma}{\theta}\right)^{\alpha-1} e^{-\left(\frac{t-\gamma}{\theta}\right)^\alpha}$
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	$1 - e^{-\left(\frac{t-\gamma}{\theta}\right)^\alpha}$
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{-\left(\frac{t-\gamma}{\theta}\right)^\alpha}$
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{\alpha}{\theta^\alpha} (t - \gamma)^{\alpha-1}$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty t S(t) dt$	작성중
평균잔여수명함수	$m(t) = \frac{E[T-t T>t]}{\int_t^\infty S(u) du}$	작성중
k 차 적률	$E(T^k)$	작성중
100p 백분위수	t_p	작성중
변수		$t \geq 0$
파라메터	$\alpha > 0$ (형상 모수; shape parameter), $\theta > 0$ (척도 모수; scale parameter) $0 < \gamma < \infty$ (위치 모수; location parameter), $t \geq \gamma$	

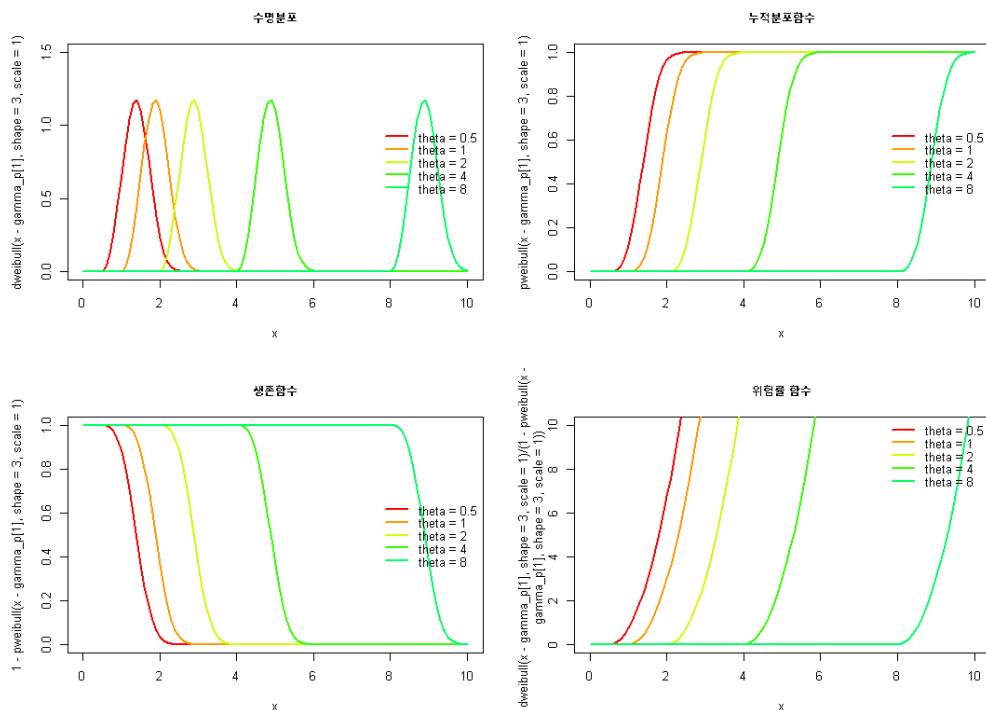


Figure 30: 3모수 와이블분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수

Code 16. 3모수 와이블분포에 기반한 수명분포함수, 누적분포함수, 생존함수, 위험률 함수(R 코드)

```
1 ##### Weibull Distribution with 3 parameters
2 ### parameter: alpha
3 gamma_p = c(0.5, 1, 2, 4, 8) # location
4
5 ### Input Variable
6 x <- seq(0, 10, length.out = 101)
7
8 color = rainbow(10)
9 par(mfrow = c(2, 2))
10
11
12 ### Life Distribution
13 plot(x, dweibull(x-gamma_p[1], shape=3, scale=1), xlim=c(0, 10), ylim=c(0, 1.5), col=color[1], lwd=2, type = 'l', main="Life
   Distribution")
14 for (i in 2:5) { lines(x, dweibull(x-gamma_p[i], shape=3, scale=1), col=color[i], lwd=2); }
15 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
16
17 ### Cumulative Distribution
18 plot(x, pweibull(x-gamma_p[1], shape=3, scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Cumulative
   Distribution")
19 for (i in 2:5) { lines(x, pweibull(x-gamma_p[i], shape=3, scale=1), col=color[i], lwd=2); }
20 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
21
22 ### Survival Function
23 plot(x, 1-pweibull(x-gamma_p[1], shape=3, scale=1), xlim=c(0, 10), ylim=c(0, 1), col=color[1], lwd=2, type = 'l', main="Survival
   Function")
24 for (i in 2:5) { lines(x, 1-pweibull(x-gamma_p[i], shape=3, scale=1), col=color[i], lwd=2); }
25 legend('right', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
26
27 ### Hazard Function
28 plot(x, dweibull(x-gamma_p[1], shape=3, scale=1)/(1-pweibull(x-gamma_p[1], shape=3, scale=1)), xlim=c(0, 10), ylim=c(0, 10), col=
   color[1], lwd=2, type = 'l', main="Hazard Function")
29 for (i in 2:5) { lines(x, dweibull(x-gamma_p[i], shape=3, scale=1)/(1-pweibull(x-gamma_p[i], shape=3, scale=1)), col=color[i], lwd
   =2); }
30 legend('topright', bty = 'n', lwd=2, col=color[1:5], legend = c('theta = 0.5', 'theta = 1', 'theta = 2', 'theta = 4', 'theta = 8'))
```

시스템의 수명이 일정기간 동안 보장되는 경우에는, 3모수 와이블분포를 적용하는 것이 보다 적절하다.

척도 모수(scale parameter)는 종종 $\lambda = \frac{1}{\theta}$ 로 표시되는 경우도 있다.
위치모수 $\gamma = 0$ 이면 2모수 와이블 분포가 된다.

16.1.39.3 Log-Weibull Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{1}{b} e^{\frac{t-a}{b}} - e^{\frac{t-a}{b}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$e^{-e^{\frac{t-a}{b}}}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{1}{b} e^{\frac{t-a}{b}}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수	$t \geq 0$		
파라메터	$a \in \mathbf{R}, b > 0$		

Table 70: Log-Weibull 분포함수에 기반한 척도 함수

16.1.39.4 Double Weibull Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{c}{2b} \left \frac{t-a}{b} \right ^{c-1} e^{-\left \frac{t-a}{b} \right ^c}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - \frac{1}{2} e^{-\left(\frac{a-t}{b} \right)^c}$ for $x \leq a$ $\frac{1}{2} e^{-\left(\frac{t-a}{b} \right)^c}$ for $x \geq a$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{c \left(\frac{a-t}{b} \right)^{c-1} e^{-\left(\frac{a-t}{b} \right)^c}}{b \left[2 - e^{-\left(\frac{a-t}{b} \right)^c} \right]}$ for $x \leq a$ $\frac{c \left(\frac{t-a}{b} \right)^{c-1} e^{-\left(\frac{t-a}{b} \right)^c}}{b e^{-\left(\frac{t-a}{b} \right)^c}}$ for $x \geq a$	
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \in \mathbf{R}$	
파라메터		$a \in \mathbf{R}, b > 0, c > 0$	

Table 71: Double Weibull 분포함수에 기반한 척도 함수

만일 $c = 2$ 가 되면, 이 분포는 Laplace 분포가 된다.

16.1.39.5 Inverse Weibull Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{c}{b} \left(\frac{t-a}{b}\right)^{-c-1} e^{-\left(\frac{t-a}{b}\right)^{-c}}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - e^{-\left(\frac{t-a}{b}\right)^{-c}}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{c \left(\frac{t-a}{b}\right)^{-c-1} e^{-\left(\frac{t-a}{b}\right)^{-c}}}{b \left[1 - e^{-\left(\frac{t-a}{b}\right)^{-c}}\right]}$	IDHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	작성중	
k 차 적률	$E(T^k)$	작성중	
100 p 백분위수	t_p	작성중	
변수		$t \geq a$	
파라메터		$a \in \mathbf{R}, b > 0, c > 0$	

Table 72: Inverse Weibull 분포함수에 기반한 척도 함수

16.1.39.6 Reflected Weibull Distribution

척도 함수	기호	내용	구분
수명분포함수	$f(t)$	$\frac{c}{b} \left(\frac{a-t}{b}\right)^{c-1} e^{-\left(\frac{a-t}{b}\right)^c}$	
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	작성중	
생존함수 (신뢰도함수)	$S(t) = P(T > t)$ $= 1 - P(T \leq t)$ $= 1 - F(t)$ $= e^{-H(t)}$	$1 - e^{-\left(\frac{a-t}{b}\right)^c}$	
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{c\left(\frac{a-t}{b}\right)^{c-1}}{b\left[e^{\left(\frac{a-t}{b}\right)^c} - 1\right]}$	IHR
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중	
평균수명	$E[T] = \int_0^\infty t f(t) dt$ $= \int_0^\infty S(t) dt$	작성중	
평균잔여수명함수	$m(t) = E[T - t T > t]$ $= \frac{\int_t^\infty S(u) du}{S(t)}$	No closed form	DMRL
k 차 적률	$E(T^k)$	작성중	
100p 백분위수	t_p	작성중	
변수		$t \leq a$	
파라메터		$a \in \mathbf{R}, b > 0$	

Table 73: Reflected Weibull 분포함수에 기반한 척도 함수

16.1.40 Wigner's Semi-circle Distribution

이 분포는 Semi-elliptical distribution으로도 알려져 있다. Chapter 16.1.32를 참고하도록 하자.

16.2 이산화률분포

- The Basic Representative of a Discrete Distribution is its PMF

$$P_i = P(X = i) \quad i = 0, 1, 2, \dots \text{ or } 1, 2, 3, \dots$$

- The Survival Function

$$S_i = P(X \geq i) = \sum_{j \geq i} P_j$$

- The Hazard Rate

$$h_i = \frac{P_i}{S_i}$$

- The Mean Residual Life Function

$$L_i = E[X - i | X \geq i]$$

16.2.1 Binomial Distribution

16.2.1.1 Positive Binomial Distribution

16.2.2 Extreme Value Distribution(극치 분포)

16.2.2.1 Weibull Distribution of Type I

16.2.2.2 Weibull Distribution of Type II

16.2.2.3 Weibull Distribution of Type III

Table 74: Type III 극치분포함수에 기반한 척도 함수

척도 함수	기호	내용
수명분포함수	$f(t)$	$\frac{m}{\beta} \left(\frac{t-\mu}{\beta}\right)^{m-1} e^{-\left(\frac{t-\mu}{\beta}\right)^m}$
누적분포함수 (불신뢰도함수)	$F(t) = P(T \leq t)$	$1 - e^{-\left(\frac{t-\mu}{\beta}\right)^m}$
생존함수 (신뢰도함수)	$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-H(t)}$	$e^{-\left(\frac{t-\mu}{\beta}\right)^m}$
위험률 함수 (고장률 함수)	$h(t) = \frac{f(t)}{S(t)}$	$\frac{m}{\beta} \left(\frac{t-\mu}{\beta}\right)^{m-1}$
누적 위험률 함수 (누적 고장률 함수)	$H(t) = -\log S(t)$	작성중
평균수명	$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$	작성중
평균잔여수명함수	$m(t) = E[T-t T>t] = \frac{\int_t^\infty S(u) du}{S(t)}$	작성중
k 차 적률	$E(T^k)$	작성중
100 p 백분위수	t_p	작성중
변수		$t \geq 0$
파라메터		$t \geq \mu, -\infty < \mu < \infty$ (위치 모수; location parameter) $m > 0$ (형상 모수; shape parameter), $\beta > 0$ (척도 모수; scale parameter)

Type III 극치분포는 부하분담(load-sharing) 시스템의 고장 강도를 모형화하는 등의 용도로 사용된다.

Notation을 다르게 사용하기는 했지만, 3모수 Weibull 분포이다. 따라서 Chapter 16.1.39.2을 참고하도록 하자.

16.2.3 Geometric Distribution

16.2.3.1 Positive Geometric Distribution

16.2.3.2 Zero-inflated Geometric Distribution

16.2.4 Hypergeometric Distribution

16.2.4.1 Positive Hypergeometric Distribution

16.2.5 Logarithmic Distribution

16.2.5.1 Right-truncated Logarithmic Distribution

16.2.6 Matching Distribution

16.2.7 Negative Binomial Distribution

16.2.8 Negative Hypergeometric Distribution

16.2.9 Occupancy Distribution

16.2.10 Poisson Distribution

16.2.10.1 Positive Poisson Distribution

16.2.11 Polya Distribution

16.2.12 Runs Distribution

16.2.13 Salvia-Bollinger's Distribution

16.2.13.1 Salvia-Bollinger's DHR Distribution

16.2.13.2 Salvia-Bollinger's Generalized DHR Distribution

16.2.13.3 Salvia-Bollinger's Generalized IHR Distribution

16.2.14 Triangular Distribution

16.2.14.1 Right-angled and Negatively Skew Triangular Distribution

16.2.14.2 Right-angled and Positively Skew Triangular Distribution

16.2.14.3 Right-angled and Symmetric Triangular Distribution

16.2.15 Uniform Distribution(Discrete)

16.2.16 Yule Distribution

16.2.17 Zeta Distribution

16.2.17.1 Zeta Distribution of Zipf

16.2.17.2 Zeta Distribution of Haight

Part VII

Cox 비례위험모형

만일 생존시간 데이터의 분포에 대한 정보가 있다면, 모수회귀모형을 이용하면 된다.

그러나 이러한 분포에 대한 가정이 적절하지 못할 경우에, 모수회귀모형을 사용하면, 추정된 계수에 대한 정확성에 대해 신뢰성을 잃게 된다.

모수회귀모형과 비례위험모형의 차이점을 간략하게 보면 다음과 같다.

모수회귀모형:

$$\log T_i = Z_i \beta + \epsilon_i$$

Covariate Z_i 가 반응변수 $\log T_i$ 와 선형관계로 표현된다.

비례위험모형:

$$h(t|Z_i) = h_0(t) \exp(Z_i \beta) \quad (16)$$

* 회귀계수 β 가 covariate 변화에 따른 log(생존시간)의 관계를 나타낸다는 점에 있어서는 동일하다.

비례위험모형(proportional hazard model, 이하 PH model)은, 위험함수에 covariate에 대한 회귀식을 포함하는 모형으로, 다음과 같은 특징이 있어 널리 사용되고 있다.

- 기저 분포에 대한 가정이 필요하지 않다.
- time-varying covariate을 회귀모형에 포함시킬 수 있다.
- 추정된 회귀계수는 covariate과 위험함수의 관계를 나타낸다.

Cox 모형의 기본 가정은 다음과 같다.

- Relative hazard는 시간에 관계 없이, 시간과는 독립적으로 일정하다는 비례 위험의 가정

식 (16)으로부터 $\frac{h_x(t)}{h_0(t)} = e^{Z_i \beta}$ 가 되는데, $e^{Z_i \beta}$ 가 상수값처럼 취급된다고 했기 때문이다.

이를 확인하기 위해, Kaplan-Meier 생존곡선을 그려서, 곡선이 교차하지 않고 평행하게 가는지를 확인하여야 한다. 여기에서 relative hazard가 상수라고 해도, 반드시 기차길처럼 평행할 필요는 없다. 생존곡선에서 두 개의 곡선의 차이가 점차적으로 증가하면 비례가정을 만족하는 것이고, 두 개의 곡선이 서로 교차하거나 만나면 비례가정에 위배되는 것이다. 이에 대한 판단은 생존 곡선을 보고 주관적으로 할 수밖에 없다.

이외에도 log minus survival plot을 그려보고, 두 곡선 사이에 일정하게 수직으로 차이가 있으면, 비례 가정을 만족한다고 판단할 수도 있다.

이 비례가정이 만족되지 않는 경우, Cox 모형을 이용할 수 없으며, 비례가정을 만족하지 않을 경우 time dependent covariate approach²⁷를 사용하여야 한다.

- Hazard function과 covariate 사이의 log-linear 관계

hazard function과 covariate 사이에는 기본적으로 log-linear 관계가 있어야 한다. 즉, hazard function에 자연로그를 취한 값은, 각 covariate의 회귀계수와는 linear relation(선형 관계)가 있어야 한다.

이를 확인하기 위해서는 cumulative hazard function으로 Martingale residuals를 구하여 확인할 수 있다. Martingale residual과 covariate가 웬만큼 직선으로 그려지만, 가정을 만족한다고 판단한다.

²⁷가령 Extended Cox 모형 등을 이용할 수 있다.

17 Cox 비례위험모형의 원리

이 Chapter는 김재균과 서대철(2009)[1]의 문서를 옮겨놓았다.

생존분석(survival analysis)은 사건-시간 분석(time to event analysis)이라고도 불리는데, 이는 사건이 일어날 때까지의 시간을 대상으로 분석하는 통계방법이라는 의미이다. 임상적으로 가장 의미있는 사건은 환자의 사망이지만, 사망 이외의 사건도 대상이 된다. 즉, 동맥류 색전술 후의 재개통이나, 뇌동맥 삽관술 후의 뇌경색 또는 스텐드 재협착도 의미 있는 사건이 될 수 있으며, 따라서 신경중재치료의학 분야에서 유용하게 사용할 수 있고, 또 사용하고 있는 통계 기법이다.

환자마다 연구에 참여하는 시점이 다르므로, 생존분석에서 실제로 다루는 것은 연구 시작 시점에서 사건이 일어날 때까지의 기간이다. 연구 종료시점 전에 사망하는 환자들의 생존기간 자료를 완결(complete) 자료라고 하며, 연구종료 후에도 살아있는 환자의 생존기간 자료를 결단(censored) 자료라고 한다. **연구 도중에 추적이 안되거나 탈락한 환자의 자료는 마지막으로 확인한 시점까지는 살아 있었으므로, 결단자료로 취급한다.** 특히, **전자의 경우 우절단(right censored or suspended) 자료라고 하는데, 시간 축의 가장 우측에서 결단되기 때문에 그렇게 불린다.**

환자들의 다양한 생존기간(완결되거나 결단된)자료를 모아놓았을 때, 이들은 정규분포를 보이지 않으므로, 선형회귀분석이나 분산분석을 이용해서 통계분석을 할 수 없다. 대신에 **hazard $h(t)$** 를 이용하여 이를 비교 분석한다. Hazard는 t 시점까지 생존한 사람이 t 시점 바로 직후에 순간적으로 사망할 확률로 정의된다.

Hazard는 사건이 일어날 때 순간적으로 증가했다가, 사건이 없을 때는 0이 되기 때문에, 자료로부터 직접 측정하기는 어렵다. 대신에 **cumulative hazard function $H(t)$** 를 이용하는데, 이는 t 시점까지의 전체 hazard와 같으며, 이 함수의 기울기가 바로 hazard가 된다.

연구기간 동안에 hazard 또는 cumulative hazard function이 일정하지 않고 시간에 따라 변할 수 있기 때문에, 이를 직접 다루기보다는 이들의 비(ratio)를 다루는 것이 편리하다. 특정 위험인자에 노출된 군과 그렇지 않은 군의 hazard ratio를 사건이 일어나는 매 시점마다 2×2 table을 만들어서 구하고, 전 기간에 걸쳐서 평균을 내면 weighted average hazard ratio를 구할 수 있는데, 이를 **Mantel-Cox estimate of the hazard ratio(HR_{MC})**라고 한다. 최종적으로 HR_{MC} = 1을 널가설로 하는 χ^2 test를 통해서, 특정 위험인자에 노출된 군과 그렇지 않은 군의 차이를 통계적으로 검정할 수 있게 된다. 이러한 검정 방법을 **Mantel-Cox χ^2 test**라고 하며, **log-rank test**라고도 알려져 있다.

사건 발생과 관계되는 인자가 하나일 때는 log-rank test로도 충분하지만, 실제로는 둘 이상인 경우가 많다. 예를 들어, 치료 방법에 따른 생존의 차이를 보고자 할 때, 치료방법 뿐만 아니라 나이, 성별, 환자가 가지고 있는 질환, 다른 위험인자 등이 직간접적으로 생존에 영향을 미치므로, 이러한 변수(potential confounders)들을 보정해야 한다.

시간보다는, 사건 발생 여부에 초점을 두는 통계기법인 로지스틱 회귀분석(logistic regression)은 odds ratio를 종속변수로 하면서, 여러 개의 독립변수들을 동시에 보정하는 통계기법이다.

생존분석은 odds ratio와 유사한 hazard ratio를 다루므로, 로지스틱 회귀분석의 알고리즘을 차용할 수 있으며, 이러한 분석 방법을 **Cox regression**이라고 한다. Cox regression은 시간에 관계없이 hazard ratio가 일정하다는 가정(proportional hazard assumption) 하에서 이루어진다.(Proportional hazard assumption).

Proportional hazard assumption 하에서는

$$\frac{H_1(t)}{H_0(t)} = \text{상수}$$

이므로, 양변에 자연로그를 취하면

$$\ln[H_1(t)] - \ln[H_0(t)] = \ln(\text{상수})$$

가 되고, 노출된 군의 위험함수와 그렇지 않은 군의 위험함수가 서로 평행하게 된다.

이 때 $H(t) = -\ln[S(t)]$ 이므로, $\ln[-\ln S(t)]$ 그래프가 proportional hazard assumption을 검정하는 데 쓰이게 된다.

만약 hazard ratio가 시간에 따라 변할 때에는 위의 방법을 사용할 수 없게 되므로, 변형된 방법을 사용하면 된다. 변수와 시간 사이의 교호작용 항(interaction term)을 모델에 추가하는 방법이 있으며, 이는 **time dependent Cox regression**라고 불린다.

Cox model은 Kaplan-Meier method에서 쓰이는 log-rank test와는 달리, 생존율의 차이를 hazard ratio라는 수치로 계량화해줄 수 있으며, 이는 odds ratio와 유사하므로 임상적으로 해석하기가 용이하다.

18 Covariate가 1개인 비례위험모형

18.1 데이터 구조 - covariate가 1개인 경우

Covariate 1개를 포함하므로, 데이터는 $(T_i, \delta_i, Z_i(t))$ ($i = 1, \dots, n$)으로 표기한다.
여기서

- T_i 는 i 번째 object(개인)가 연구에 참여해있는 시간
- δ_i 는 i 번째 object에 대한 사건발생/중도절단 지시변수
$$\delta_i = \begin{cases} 1 & \text{관측된 경우} \\ 0 & \text{중도절단된 경우} \end{cases}$$
- $Z_i(t)$ 는 t 시점에서 i 번째 object에 대한 위험인자 또는 covariate이다.
- $t_1 < \dots < t_n$ 은 순서대로 정렬한 순서통계량

18.2 준모수적 Cox 비례위험모형

Cox 비례위험모형은 covariate에서는 모수적 형태를 가정하고, 기저함수에는 모수적 가정을 가정하지 않으므로 준모수적(semi-parametric) 방법으로 불린다.

t 시점에서 위험인자 Z 를 가진 i 번째 object에 대한 위험함수는 다음과 같다.

$$\begin{aligned} h(t|Z_i) &= h_0(t)\psi(Z; \beta) \\ &= h_0(t)\exp(Z; \beta) \end{aligned}$$

- $h(t|Z)$ 는 t 시점에서 위험인자 또는 covariate Z 를 가진 object에 대한 위험률이다.
- $h_0(t)$ 는 분포 가정이 주어져 있지 않은 기저위험함수(baseline hazard function)이다.

일반적으로 $\psi(Z; \beta) = \exp(Z; \beta)$ 인 지수함수를 고려한다.

Covariate Z_i 가 1단위 증가할 때마다 $\exp(\beta)$ 만큼 위험률이 증가한다.

위험률은 음수가 될 수 없으므로, 지수함수(음수를 갖지 않는)가 음이 아닌 함수값을 보장해주는 중요한 역할을 하며 널리 이용된다.

- 그 외에는

선형함수 $\psi(Z) = 1 + Z\beta$,

로지스틱 함수 $\psi(Z) = \log(1 + e^{Z\beta})$

등이 있다. 자세한 것은, Chapter 16를 참고하도록 하자.

1개의 covariate를 고려한 경우, 두 object 간 사건 발생 위험비를 비교해 보자.
object i 와 object j 의 사건발생 위험비는

$$\begin{aligned} \frac{h(t|Z_i)}{h(t|Z_j)} &= \frac{h_0(t)\exp(Z_i\beta)}{h_0(t)\exp(Z_j\beta)} \\ &= \exp[(Z_i - Z_j)\beta] \end{aligned}$$

즉, 이러한 위험비는 시간에 의존하지 않고, 회귀계수 β 와 covariate 값의 차이 $(Z_i - Z_j)$ 에 의존한다는 것을 알 수 있다.

Covariate이 $Z = 1$ 또는 $Z = 0$ 인 binary 데이터고 하자.

두 object i 와 j 에 대해 covariate $Z_i = 1, Z_j = 0$ 인 경우, 사건 발생 위험비는

$$\begin{aligned} \frac{h(t|Z_i)}{h(t|Z_j)} &= \frac{h_0(t)\exp(Z_i\beta)}{h_0(t)\exp(Z_j\beta)} \\ &= \exp[(Z_i - Z_j)\beta] \\ &= \exp(\beta) \end{aligned}$$

이러한 위험비는 $\exp(\beta)$ 는 시간이 변하더라도 불변(invariant)으로, 시점에 의존하지 않는다.
그래서 Cox가 ”비례위험(proportional hazard)”이라고 불렀다.

18.3 회귀계수에 대한 추론

회귀계수 β 를 추정하기 위해서는 부분가능도함수(partial likelihood function)를 이용한다.

가능도함수는 관측 데이터셋의 확률에 비례하지만, 반면 부분가능도함수는 그렇지 않다. 그럼에도 불구하고, 부분가능도함수는 근사적인 추론에서 가능도함수와 마찬가지로 다를 수 있어서 유용하다.



동점이 없는 생존시간 데이터에 대한 부분가능도함수는

$$\begin{aligned} PL(\beta) &= \prod_{i=1}^n \left[\frac{Y_i \exp(Z_i \beta)}{\sum_{\ell \in R_i} Y_\ell \exp(Z_\ell \beta)} \right]^{\delta_i} \\ &= \prod_{i=1}^n \left[\frac{Y_i r_i(\beta, t)}{\sum_{\ell \in R_i} Y_\ell r_\ell(\beta, t)} \right]^{\delta_i} \end{aligned}$$

여기서, $r_i(\beta, t) = \exp[Z_i \beta] = r_i(t)$ 은 i 번째 object에 대한 위험점수(risk score)이다.



여기서 로그를 취하면

$$\begin{aligned} p\ell(\beta) &= \log PL(\beta) \\ &= \sum_{i=1}^n \delta_i \left[\log \{Y_i \exp(Z_i \beta)\} - \log \left\{ \sum_{\ell \in R_i} Y_\ell \exp(Z_\ell \beta) \right\} \right] \end{aligned}$$



로그 부분가능도함수 $p\ell(\beta)$ 를 β 에 대해 미분하면, score function $U(\beta)$ 를 얻을 수 있다.

$$\begin{aligned} U(\beta) &= \frac{\partial [p\ell(\beta)]}{\partial \beta} \\ &= \sum_{i=1}^n \delta_i \left[\frac{Z_i Y_i \exp(Z_i \beta)}{Y_i \exp(Z_i \beta)} - \frac{\sum_{\ell \in R_i} Z_\ell Y_\ell \exp(Z_\ell \beta)}{\sum_{\ell \in R_i} Y_\ell \exp(Z_\ell \beta)} \right] \\ &= \sum_{i=1}^n \delta_i \left[Z_i - \frac{\sum_{\ell \in R_i} Z_\ell Y_\ell \exp(Z_\ell \beta)}{\sum_{\ell \in R_i} Y_\ell \exp(Z_\ell \beta)} \right] \end{aligned}$$

최대부분가능도추정량 $\hat{\beta}$ 은 다음의 부분가능도방정식을 풀어 구한다.

$$U(\hat{\beta}) = 0$$



이렇게 구한 해 $\hat{\beta}$ 은 β 에 대한 일치추정량이 되며, 근사적으로 정규분포 $N(\beta, E[I(\beta)]^{-1})$ 를 따른다. 여기서 information matrix $I(\beta)$ 는

$$\begin{aligned} I(\beta) &= -\frac{\partial^2 [p\ell(\beta)]}{\partial \beta^2} \\ &= \sum_{i=1}^n \delta_i \left[\frac{Z_i Y_i \exp(Z_i \beta)}{Y_i \exp(Z_i \beta)} - \frac{\sum_{\ell \in R_i} Z_\ell Y_\ell \exp(Z_\ell \beta)}{\sum_{\ell \in R_i} Y_\ell \exp(Z_\ell \beta)} \right] \end{aligned}$$



함수 $U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta}$ 에 대해 테일러 공식을 적용하면, 근사적으로

$$\begin{aligned} 0 &= U(\hat{\beta}) \\ &\approx U(\beta) + U'(\beta)(\hat{\beta} - \beta) \end{aligned}$$



$U'(\beta) = \frac{\partial^2 [p\ell(\beta)]}{\partial \beta^2}$ 이고, $-U'(\beta) = I(\beta)$ 으로,

$$\hat{\beta} = \beta[U'(\beta)]^{-1}$$

$$U(\beta) \approx \beta + I^{-1}(\beta) \cdot U(\beta)$$



적당한 초기값 $\hat{\beta}^{(0)}$ 를 이용하여, 다음과 같이 Newton-Raphson 알고리즘을 사용하여 반복적인 계산을 통해 $\hat{\beta}$ 를 구한다.

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + I^{-1}(\hat{\beta}^{(k)})U(\hat{\beta}^{(k)})$$

로그부분가능도함수가 수렴할 때까지, 즉 $\ell(\hat{\beta}^{(k+1)}) \approx \ell(\hat{\beta}^{(k)})$ 일 때까지 반복적인 계산을 하여 해를 구한다.

✓ 이렇게 구한 최대부분가능도추정량 $\hat{\beta}$ 을 기반으로 누적위험함수를 구하기 위해, 다음과 같은 Breslow 추정량을 이용한다.

$$\hat{H}(t|Z) = \sum_{t_i \leq t} \frac{d_i}{\sum_{\ell \in R_i} \exp(Z_\ell \hat{\beta})}$$

여기서, $d_i = \sum_{\ell=1}^n I(t_\ell = t_i)$

✓ 회귀계수 추정량 $\hat{\beta}$ 의 분포는, 마팅게일 중심극한정리를 이용한다.

$$\hat{\beta} - \beta \sim N(0, I^{-1}(\beta))$$

회귀계수추정량 $\hat{\beta}$ 은 근사적으로 정규분포를 따른다.

✓ 회귀계수 β 에 대한 근사적인 $(1 - \alpha) \times 100\%$ 신뢰구간은 다음과 같다.

$$\hat{\beta} \pm z_{\alpha/2} [I^{-1}(\beta)]^{\frac{1}{2}}$$

19 Covariate가 여러 개인 비례위험모형

19.1 데이터 구조 - covariate가 여러 개인 경우

Covariate vector를 포함하므로, 데이터는 $(T_i, \delta_i, Z_i(t))$ (단, $i = 1, \dots, n$)으로 표기한다.
여기서

- T_i 는 i 번째 object(개인)가 연구에 참여해있는 시간
- δ_i 는 i 번째 object에 대한 사건발생/중도절단 지시변수
$$\delta_i = \begin{cases} 1 & \text{관측된 경우} \\ 0 & \text{중도절단된 경우} \end{cases}$$
- $Z_i(t) = [Z_{i1}(t), \dots, Z_{ip}(t)]'$ 는 t 시점에서 i 번째 object에 대한 위험인자 또는 covariate
 $Z_{ij}(t)$ 는 t 시점에서 i 번째 object의 j 번째 위험인자 또는 covariate
Covariate가 시간 t 에 의존하지 않을 경우에는 $Z_{ij}(t) = Z_{ij}$ 로 표기하기도 한다.

19.2 준모수적 Cox 비례위험모형

Cox 비례위험모형은 covariate에서는 모수적 형태를 가정하고, 기저함수에는 모수적 가정을 가정하지 않으므로 준모수적(semi-parametric) 방법으로 불린다.

t 시점에서 위험인자벡터 $Z = [Z_{i1}(t), \dots, Z_{ip}(t)]'$ 를 가진 i 번째 object에 대한 위험함수는 다음과 같다.

$$\begin{aligned} h(t|Z_i) &= h_0(t) \exp(Z_i' \beta) \\ &= h_0(t) \exp(Z_{i1}\beta_1 + \dots + Z_{ip}\beta_p) \end{aligned}$$

- $h(t|Z)$ 는 t 시점에서 위험인자 또는 covariate Z 를 가진 object에 대한 위험률이다.
- $h_0(t)$ 는 분포 가정이 주어져 있지 않은 기저위험함수(baseline hazard function)이다.
- $\beta = (\beta_1, \dots, \beta_p)'$ 는 $p \times 1$ 회귀계수벡터로, covariate의 효과를 추정하는 회귀계수로 구성된다.
- Covariate Z_i 가 1단위 증가할 때마다 $\exp(\beta)$ 만큼 위험률이 증가한다.

p개의 covariate를 고려한 경우, 두 object 간 사건 발생 위험비를 비교해 보자.
object i 와 object j 의 사건발생 위험비는

$$\begin{aligned} \frac{h(t|Z_i)}{h(t|Z_j)} &= \frac{h_0(t) \exp(Z_{i1}\beta_1 + \dots + Z_{ip}\beta_p)}{h_0(t) \exp(Z_{j1}\beta_1 + \dots + Z_{jp}\beta_p)} \\ &= \exp \left[\sum_{k=1}^p (Z_{ik} - Z_{jk})\beta_k \right] \end{aligned}$$

즉, 이러한 위험비는 시간에 의존하지 않고, covariate 값에 비례한다.

또한, 이 식을, 위험인자 $Z_i = (Z_{i1}, \dots, Z_{ip})'$ 와 $Z_j = (Z_{j1}, \dots, Z_{jp})'$ 를 가진 object 간의 상대 위험도(relative risk) 또는 위험비(hazard ratio)라고 부른다.

19.3 회귀계수와 누적위험함수의 추정

Covariate가 여러 개인 경우, covariate vector를 이용해 데이터 구조와 모형을 $(T_i, \delta_i, Z_i(t))$, $Z_i = (Z_{i1}, \dots, Z_{ip})'$ (단, $i = 1, \dots, n$)으로 표기한다.

- $\delta_i = \begin{cases} 1 & \text{관측된 경우} \\ 0 & \text{중도절단된 경우} \end{cases}$
- $Z_i = (Z_{i1}, \dots, Z_{ip})'$ 는 $p \times 1$ covariate vector
- $t_1 < \dots < t_n$ 은 순서대로 정렬한 순서통계량
- t_j 시점에 사건을 가진 object의 covariate는 Z_j 로 표기한다.

위험그룹 $R_j = R(t_j)$ 는 t_j 바로 전 시점까지 사건을 경험하지 않고 살아있는 object들의 그룹이다.

19.3.1 비례위험모형

비례위험모형에서 Cox의 비례위험함수는 다음과 같다.

$$\begin{aligned} h(t|Z_i) &= h_0(t) \exp(Z\beta) \\ &= h_0(t) \exp(Z_1\beta_1 + \cdots + Z_p\beta_p) \end{aligned} \quad (17)$$

- $h(t|Z)$ 는 t 시점에서 위험인자 또는 covariate Z 를 가진 object에 대한 위험률이다.
- $h_0(t)$ 는 분포 가정이 주어져 있지 않은 기저위험함수(baseline hazard function)이다.
- $\beta = (\beta_1, \dots, \beta_p)'$ 는 $p \times 1$ 회귀계수벡터로, covariate의 효과를 추정하는 회귀계수로 구성된다.
- $Z = (Z_i, \dots, Z_p)'$ 는 $p \times 1$ covariate vector

i 번째 object에 대한 위험함수는 다음과 같다.

$$\begin{aligned} h(t|Z_i) &= h_0(t) \exp(Z'_i\beta) \\ &= h_0(t) \exp(Z_{i1}\beta_1 + \cdots + Z_{ip}\beta_p) \end{aligned}$$

19.3.2 부분가능도함수

한 object에 대한 부분가능도함수(partial likelihood)는 다음과 같이 구한다.

$$\begin{aligned} L_j^*(\beta) &= P(A|B) \\ &= \frac{h_0(t_j) \exp(Z_j\beta)}{\sum_{\ell \in R(t_j)} h_0(t_\ell) \exp(Z_\ell\beta)} \\ &= \frac{\exp(Z_j\beta)}{\sum_{\ell \in R(t_j)} \exp(Z_\ell\beta)} \end{aligned}$$

✓ 생존시간 데이터에 대한 부분가능도함수는 다음과 같다.

$$\begin{aligned} PL(\beta) &= \prod_{i=1}^n \left[\frac{Y_i \exp(Z'_i\beta)}{\sum_{\ell \in R_i} Y_\ell \exp(Z'_\ell\beta)} \right]^{\delta_i} \\ &= \prod_{i=1}^n \left[\frac{Y_i r_i(\beta, t)}{\sum_{\ell \in R_i} Y_\ell r_\ell(\beta, t)} \right]^{\delta_i} \\ &= \prod_{i=1}^n \left[\frac{Y_i (Z_{i1}\beta_1 + \cdots + Z_{ip}\beta_p)}{\sum_{\ell \in R_i} Y_\ell (Z_{\ell 1}\beta_1 + \cdots + Z_{\ell p}\beta_p)} \right]^{\delta_i} \\ &= \prod_{i=1}^n [L_i^*]^{\delta_i} \end{aligned} \quad (18)$$

여기서 $r_i(\beta, t) = \exp[Z'_i\beta] \equiv r_i(t)$ (i 번째 object에 대한 위험접수)

✓ $p\ell(\beta) = \log PL(\beta)$ 을 이용하여 score function $U(\beta_k)$ 를 다음과 같이 구한다.

$$\begin{aligned} U(\beta_k) &= \frac{\partial[p\ell(\beta)]}{\partial \beta_k} \\ &= \sum_{i=1}^n \delta_i \left[Y_i Z_{ik} - \frac{\sum_{\ell \in R_i} Y_\ell \exp(Z'_\ell\beta) Z_{\ell k}}{\sum_{\ell \in R_i} Y_\ell \exp(Z'_\ell\beta)} \right] \end{aligned}$$

19.3.3 정보행렬(information matrix)

모수에 대한 -2차 미분을 구하여 다음과 같은 정보행렬 $I(\beta)$ 을 얻는다.

$$I(\beta) = [I_{gh}(\beta)]_{p \times p}$$

$$= - \begin{bmatrix} \frac{\partial^2 \ell(\beta)}{\partial \beta_1^2} & \frac{\partial^2 \ell(\beta)}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 \ell(\beta)}{\partial \beta_1 \partial \beta_p} \\ \cdots & \frac{\partial^2 \ell(\beta)}{\partial \beta_2^2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \frac{\partial^2 \ell(\beta)}{\partial \beta_p^2} \end{bmatrix}$$

여기서 $g, h = 1, \dots, p$ 에 대하여,

$$\begin{aligned} I_{gh}(\beta) &= \frac{\partial^2 \ell(\beta)}{\partial \beta_g \partial \beta_h} \\ &= \sum_{i=1}^n \frac{\sum_{j \in R(t_i)} Z_{jg} z_{jh} \exp[\sum_{k=1}^p \beta_k Z_{jk}]}{\sum_{j \in R(t_i)} \exp[\sum_{k=1}^p \beta_k Z_{jk}]} \\ &\quad - \sum_{i=1}^D \left[\frac{\sum_{j \in R(t_i)} z_{jh} \exp(\sum_{k=1}^p \beta_k z_{jk})}{\sum_{j \in R(t_i)} \exp(\sum_{k=1}^p \beta_k z_{jk})} \right] \left[\frac{\sum_{j \in R(t_i)} z_{jh} \exp(\sum_{k=1}^p \beta_k z_{jk})}{\sum_{j \in R(t_i)} \exp(\sum_{k=1}^p \beta_k z_{jk})} \right] \end{aligned}$$

✓ 적당한 초기값 $\hat{\beta}^{(0)}$ 를 이용하여, 다음과 같이 Newton-Raphson 알고리즘을 사용하여, 반복적인 계산을 통해 $\hat{\beta}$ 를 구한다.

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + I^{-1}(\hat{\beta}^{(k)})U(\hat{\beta}^{(k)})$$

로그부분가능도함수가 수렴할 때까지, 즉 $\ell(\hat{\beta}^{(k+1)}) \approx \ell(\hat{\beta}^{(k)})$ 일 때까지 반복적인 계산을 하여 해를 구한다.

✓ 이렇게 구한 최대부분가능도추정량 $\hat{\beta}$ 을 기반으로 누적위험함수를 구하기 위해, 다음과 같은 Breslow 추정량을 이용한다.

$$\hat{H}(t|Z) = \sum_{t_i \leq t} \frac{d_i}{\sum_{\ell \in R_i} \exp(Z_\ell \hat{\beta})}$$

여기서, $d_i = \sum_{\ell=1}^n \delta_i I(t_\ell = t_i)$

Note 30. Cox 비례위험모형 추정량의 통계적 성질

이러한 회귀계수추정량 $\hat{\beta}$ 는 다음의 통계적 성질을 만족한다.

[일치성(consistency)] 표본의 크기가 클 수록 $\hat{\beta}$ 은 참값 β 로 수렴한다.

[근사적 정규성(asymptotic normality)] $\hat{\beta}$ 는 근사적으로 정규분포를 따른다.

[효율성(efficiency)] β 에 대한 추정량들 중 MPLE(maximum partial likelihood estimator)가 최소분산을 갖는다.

20 동점 처리(handling ties)

비례위험모형에서 유도된 부분가능도함수 식은, 모든 시점들이 서로 동일하지 않다는 가정 하에서 유도된다.

그러나 실제 데이터에서 생존시간이 동일한 경우는 종종 발생할 것이다.

이러한 동점 생존시간이 존재할 경우, 부분가능도함수를 생성하는 여러가지 동점 처리 방법들이 제안되었다.

설명의 편의를 위해, 예를 하나 만들어놓자. 다음의 데이터셋에서 부분가능도함수를 계산해보자. 환자 1과 환자 2의 시간 데이터 값이 동점인 경우이다.

($t_1 = t_2 < t_3 < t_4 < t_5$ 라고 하자.)

ID	t	δ	Z
1	t_1	1	z_1
2	t_2	1	z_2
3	t_3	0	z_3
4	t_4	1	z_4
5	t_5	1	z_5

환자 1과 환자 2 사이의 순서를 알 수 없으므로, 가능한 순서는 $2!$ 이다.

- A_1 은 환자 1이 환자 2보다 먼저 사건이 발생하는 경우
- A_2 은 환자 2이 환자 1보다 먼저 사건이 발생하는 경우

그러면, $L_1(\beta) = P(A_1 \cup A_2) = P(A_1) + P(A_2)$ 가 된다.

20.1 Exact Method

가능한 위험집합에서 부분가능도함수를 계산하여 동점에 대해 처리하는 방법을 말한다.

이제 앞서 만든 예를 바탕으로 exact method를 생각해보자. 동점인 두 관측값 $t_1 = t_2$ 의 부분가능도함수 기여값은 다음과 같이 계산할 수 있다.

$$P(A_1) = \left(\frac{e^{z_1\beta}}{e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \right) \left(\frac{e^{z_2\beta}}{e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \right)$$

$$P(A_2) = \left(\frac{e^{z_2\beta}}{e^{z_2\beta} + e^{z_1\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \right) \left(\frac{e^{z_1\beta}}{e^{z_1\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \right)$$

$L_j(\beta)$ 를 j 번째 서로 다른 값 $\{t_1, t_4, t_5\}$ 에서의 부분가능도함수라고 할 때 (t_3 는 중도절단이므로, 부분가능도계산에서 제외됨)

$$L(\beta) = L_1(\beta)L_2(\beta)L_3(\beta)$$

여기서 $L_1(\beta) = P(A_1) + P(A_2)$ 로 계산한다.

일단 부분가능도함수가 계산된 후에는 모두 추정 방법은 동일하다.

$$L_2(\beta) = \frac{e^{z_1\beta}}{e^{z_1\beta} + e^{z_5\beta}}$$

$$L_3(\beta) = \frac{e^{z_5\beta}}{e^{z_5\beta}} = 1$$

20.2 Breslow's Approximation

Breslow가 제안한 통계량으로, 근사 정도가 좋은 경우에 유용하며, 매우 간단한 방법이다.

Exact method에서 사용한 식(1)을 식을 다음과 같은 근사를 이용하여 $P(A_1)$, $P(A_2)$ 를 계산한다.

$$\left(\frac{e^{z_2\beta}}{e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \right) \approx \left(\frac{e^{z_2\beta}}{e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \right)$$

$$\left(\frac{e^{z_1\beta}}{e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \right) \approx \left(\frac{e^{z_1\beta}}{e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \right)$$

\implies

$$P(A_1) = \left(\frac{e^{z_1\beta}}{e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \right) \left(\frac{e^{z_2\beta}}{e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \right)$$

$$= \frac{e^{(z_1+z_2)\beta}}{[e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}]^2}$$

$$P(A_2) = \left(\frac{e^{z_2\beta}}{e^{z_2\beta} + e^{z_1\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \right) \left(\frac{e^{z_1\beta}}{e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \right)$$

$$= \frac{e^{(z_1+z_2)\beta}}{[e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}]^2}$$

따라서,

$$L_1(\beta) \approx \frac{\exp \left[\beta \sum_{\ell \in R_i} z_\ell \right]}{\left[\sum_{\ell \in R_i} \exp(z_\ell \beta) \right]^{d_j}}$$

즉, 동점값이 모두 분모에 들어가게 되어

$$\left(\frac{r_1}{r_1 + r_2 + r_3 + r_4 + r_5} \right) \left(\frac{r_2}{r_1 + r_2 + r_3 + r_4 + r_5} \right)$$

으로 나타낸다.

20.3 Efron's Approximation

부분가능도에 가까운 근사법이다.

Exact method에 의하면 다음과 같이 쓸 수 있다.

$$\begin{aligned} L_1(\beta) &= \frac{bc}{a(a-b)} + \frac{bc}{a(a-c)} \\ &\approx \frac{2bc}{a} \frac{1}{\left(a - \frac{b+c}{2}\right)} \end{aligned}$$

여기서

$$\begin{aligned} a &= e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta} \\ b &= e^{z_1\beta} \\ c &= e^{z_2\beta} \end{aligned}$$

이러한 근사식에 착안하여 1번쨰 동점 데이터에 대한 부분가능도함수의 일반적인 근사식은 다음과 같다.

$$L_1(\beta) \approx \frac{\exp \left[\sum_{\ell \in R_i} e^{z_\ell \beta} \right]}{\prod_{j=1}^{d_1} \left[\sum_{\ell \in R_1} e^{z_\ell \beta} - \frac{j-1}{d_1} \sum_{\ell \in R_i} e^{z_\ell \beta} \right]}$$

만약 2개의 동점이 있다면, 분모에 두 관측값의 평균을 사용한다.

$$\left(\frac{r_1}{r_1 + r_2 + r_3 + r_4 + r_5} \right) \left(\frac{r_2}{0.5r_1 + 0.5r_2 + r_3 + r_4 + r_5} \right)$$

앞선 예의 경우,

- 2개의 동점이 있다면 가중치는 $(1, \frac{1}{2})$
- 3개의 동점이 있다면 가중치는 $(1, \frac{1}{2}, \frac{1}{3})$
- 4개의 동점이 있다면 가중치는 $(1, \frac{3}{4}, \frac{2}{4}, \frac{1}{4})$
- r 개의 동점이 있다면, 가중치는 $(1, \frac{r-1}{r}, \frac{r-2}{r}, \dots, \frac{1}{r})$

Part VIII

Performance Evaluation Metrics

본 Chapter의 지표들에 대한 카테고리는 추순규(2015)[6], 석진미(2017)[4], Ping Wang et al.(2017)[51], 임성빈(????)[5]을 참고하여 정리하였다.

임상 분야에서 질병의 유무를 예측함에 있어,
반응 변수(response variable)가 단순히 ”질병의 유/무” 구분을 의미하는 binary data이면

- Receive operating characteristic curve(**ROC Curve**) & Area under the curve(**AUC**)
- Net reclassification improvement(**NRI**)
- Mean Square Error(**MSE**) & Mean Absolute Error(**MAE**)
- Cox and Snell의 결정계수(R^2) & Nagelkerke의 결정계수(R^2)
- Brier Score
- Hosmer-Lemeshow Test

등을 사용할 수 있다.

하지만, 반응 변수(response variable)가 생존 시간일 경우에, 위 지표는 생존 자료가 가질 수 있는 정보를 고려하지 못하는 제약이 있으므로 그에 맞는 다음과 같은 평가 지표를 사용해야 한다.

- Ignoring time to event **C-index**)
- Chambliss and Diao의 **C statistic**)
- Gonen and Heller의 **K**)
- Harrell의 Concordance Index(**C-index**)
- Heagerty의 intergrated AUC
- Uno의 **C-index**
- Pencina and Uno의 **NRI**
- **Ctd Index**

다만, 반응 변수(response variable)가 binary data인 경우와, 생존 시간인 경우 각각 사용해야하는 평가 지표가 반드시 위와 같이 염밀하게 구분하여 사용되는 것은 아니고, 위 구분도 단지 survey 과정에서 편의를 위하여 임의로 구분한 것임을 염두에 두자. 어차피 앞으로 연구 목적에 따라 사용할 지표가 정해지게 될 것이며, 위 분류는 단지 참고용으로만 알고 있도록 하자.

21 반응 변수가 binary data인 경우

21.1 Receive operating characteristic curve(ROC Curve) & Area under the curve(AUC)

ROC curve는 민감도(sensitivity)와 특이도(specificity)를 이용해 구할 수 있다.

[민감도(sensitivity, true positive rate)] 실제 질병이 있는 대상을 검사 결과, 질병이 있다고 판단할 확률

[특이도(specificity, false negative rate)] 실제 질병이 없는 대상을 검사 결과, 질병이 없다고 판단할 확률

예측 모형을 통해 얻어진 질병이 있을 확률을 p 라고 할 때, p 값이 "어느 정도"²⁸ 이상이면, 질병이 있다고 판단할 수 있다. 특히, 특이도보다 민감도가 더 중요한 상황에서는 p 가 낮아도 질병이 있다고 판단하게 된다.

ROC curve는 질병이 있다고 판단할 확률 p 를 이동하면서, 민감도와 특이도를 이용하여 구할 수 있다.

예를 들어, $p = 0.1$ 을 기준으로 질병의 유무를 판단한다고 하면, 그 때의 민감도와 특이도를 구하고 y 축을 민감도, x 축을 (1-특이도)를 이용하여 구할 수 있다.

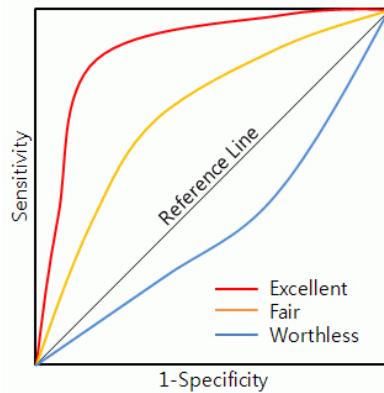


Figure 31: ROC curve

그림 31에서 Excellent, Fair, Worthless는 단지 참고용이며, 어느 질병이냐에 따라 기준이 달라질 수 있음을 염두에 두자.

AUC는 ROC curve의 아래 면적을 말하며, 0에서 1 사이의 값을 갖는다. 1에 가까울수록 모형의 예측력이 뛰어나다고 볼 수 있고, 1에 가까워진다는 것을 그림 31에 빗대어 표현하면, ROC curve가 reference line에서 빨간색 선 방향에 가까워진다는 것을 의미한다.

Note 31. AUC를 이용하여 두 모형을 비교할 경우, 통계적 유의성 검정

모형이 2개가 있고, 이 두 모형을 비교하는 경우, AUC가 1에 더 가까운 모형이 예측력이 뛰어나다고 평가한다.

또한, Henley and McNeil(1983)[31]에 따르면, 이 두 모형의 예측력 차이가 통계적으로 유의한지를 평가할 수 있는데, 첫 번째 모형을 Model₁, 두 번째 모형을 Model₂, 그리고 각 모형에서 얻어진 AUC를 AUC₁, AUC₂라고 하였을 때, 다음과 같은 관계가 있다.

$$\frac{AUC_1 - AUC_2}{SE[AUC_1 - AUC_2]} \sim Z(0, 1)$$

표준오차에 대한 식은 다음과 같다.

$$\begin{cases} \text{두 모형이 독립일 경우} & SE[AUC_1 - AUC_2] = \sqrt{SE^2(AUC_1) + SE^2(AUC_2)} \\ \text{두 모형이 독립이 아닐 경우} & SE[AUC_1 - AUC_2] = \sqrt{SE^2(AUC_1) + SE^2(AUC_2) - 2rSE^2(AUC_1)SE(AUC_2)} \end{cases}$$

단, r 은 두 AUC의 상관성을 의미한다.

²⁸"어느 정도"의 기준은, 상황에 따라 달라지게 된다. 어느 질병은 0.1 이상이어도 질병이 있다고 판단하게 되지만, 어느 질병은 0.7 이상이어야 질병이 있다고 판단하게 되기도 한다.

21.2 Net reclassification improvement(NRI)

NRI는 새로운 factor가 기존 모형에 추가되었을 때, 예측력 향상을 평가하기 위한 목적으로 제안되었다.

따라서, 이 방법은, 하나의 모형이 다른 모형에 nested되어있는 경우, 즉, 두 모형이 독립이 아닌 경우에 대해서 예측력을 비교할 때 사용된다.

기존 예측모형과, 새로운 인자가 포함된 예측모형을 각각 Model₁, Model₂라고 하면, 각 예측모형의 로지스틱 회귀모형은 다음과 같다.

$$\text{Model}_1 = \log \left(\frac{p_1(Y=1)}{1 - P_1(Y=1)} \right) = \alpha_1 + \beta_{(1,1)}x_1 + \cdots + \beta_{(n,1)}x_n$$

$$\text{Model}_2 = \log \left(\frac{p_2(Y=1)}{1 - P_2(Y=1)} \right) = \alpha_2 + \beta_{(1,2)}x_1 + \cdots + \beta_{(n,2)}x_n + \beta_{(n+1,2)}x_{n+1}$$

$Y = 1$ 인 경우를 질병이 있는 집단(event)이라고 할 때, Model₁, Model₂를 통해, event 집단이 될 예측 확률 p_1 , p_2 를 갖는다.

$$p_1(Y=1) = \frac{1}{1 + e^{-(\alpha_1 + \beta_{(1,1)}x_1 + \cdots + \beta_{(n,1)}x_n)}}$$

$$p_2(Y=1) = \frac{1}{1 + e^{-(\alpha_2 + \beta_{(1,2)}x_1 + \cdots + \beta_{(n,2)}x_n + \beta_{(n+1,2)}x_{n+1})}}$$

각 i 번째 대상은 모형에서의 예측 확률을 통해 event 집단에 속할 확률을 가지며, 어떤 cut point를 기준으로 event 집단에 속할 확률을 ”낮음, 중간, 높음”과 같이 구분할 수 있다. Model_k ($k = 1, 2$)에서 구해진 예측확률 p_k 를 cut point를 기준으로 분류표를 작성하는 것은 Note 32를 참고하자.

Note 32. Model_k ($k = 1, 2$)에서 구해진 예측확률 p_k 를 cut point를 기준으로 작성한 분류표와 이에 따른 NRI 계산

Table 75: 재분류표

Event		Model ₂			Non-event		Model ₂		
		낮음	중간	높음			L	M	H
Model ₁	낮음	a	b	c	Model ₁	낮음	j	k	l
	중간	d	e	f		중간	m	n	o
	높음	g	h	i		높음	p	q	r

Table 75은 cut point를 기준으로 하여, 범주를 총 3개 지점으로 잡은 경우의 재분류표를 나타낸다. NRI는 일반적으로 cut point를 기준으로 범주를 2 3개로 나누어 이용한다. 재분류표를 이용하여 기존 예측모형에 비해 새로운 인자가 포함된 예측 모형이 event 집단과 non-event 집단을 더 잘 분류했는지를 판단한다.

분류표를 이용한 NRI의 정의는 다음과 같다.

$$\text{NRI} = [P(\text{up}|Y=1) - P(\text{down}|Y=1)] - [P(\text{up}|Y=0) - P(\text{down}|Y=0)]$$

$P(\text{up}|Y=1)$ 은 event 집단에서 Model₂의 예측 확률이 더 높게 분류된 확률을 의미한다.

반대로 $P(\text{down}|Y=1)$ 은 event 집단에서 Model₂의 예측 확률이 더 낮게 분류된 확률을 의미한다.

$$\hat{P}(\text{up}|Y=1) = \frac{b + c + f}{a + b + c + d + e + f + g + h + i} \quad \hat{P}(\text{down}|Y=1) = \frac{d + g + h}{a + b + c + d + e + f + g + h + i}$$

$$\hat{P}(\text{up}|Y=0) = \frac{d + g + h}{j + k + l + m + n + o + p + q + r} \quad \hat{P}(\text{down}|Y=0) = \frac{m + p + q}{j + k + l + m + n + o + p + q + r}$$

추정된 확률들을 이용하여 다음과 같이 NRI를 나타낼 수 있다.

$$\begin{aligned} \hat{\text{NRI}} &= [\hat{P}(\text{up}|Y=1) - \hat{P}(\text{up}|Y=0)] + [\hat{P}(\text{up}|Y=0) - \hat{P}(\text{up}|Y=0)] \\ &= \left[\left(\frac{b + c + f}{n_{\text{event}}} \right) - \left(\frac{d + g + h}{n_{\text{event}}} \right) \right] + \left[\left(\frac{m + p + q}{n_{\text{non-event}}} \right) - \left(\frac{k + l + o}{n_{\text{non-event}}} \right) \right] \end{aligned}$$

새로운 예측모형의 예측력이 더 뛰어나다면, NRI의 값은 0보다 커질 것이다. NRI 값이 0보다 큰 값을 갖는지 검정하기 위한 검정통계량은 다음과 같다.[48]

$$z = \frac{\hat{\text{NRI}}}{\hat{SE}(\text{NRI})}$$

$$\text{where } \hat{SE}(\text{NRI}) = \sqrt{\frac{\hat{P}(\text{up}|Y=0) - \hat{P}(\text{up}|Y=0)}{n_{\text{event}}} + \frac{\hat{P}(\text{up}|Y=1) - \hat{P}(\text{up}|Y=1)}{n_{\text{non-event}}}}$$

한편, continuous NRI(혹은 category-free NRI)도 있으며, 이는 반응변수에 절대적인 범주의 수가 존재하지 않는다고 볼 수 있을 때 사용한다.

사실 절대적인 범주의 수가 존재한다고는 할 수 없거나, 범주를 나눌 때 cut point가 논쟁이 될 수 있는데, 이 문제를 해결하기 위해 Pencina et al.(2011)[49]이 범주를 나누지 않는 방법을 제안하였다. 이 방법은, 기존 예측모형보다 새로운 예측모형이 예측 확률이 높으면 up, 낮으면 down으로 분류한다. 범주를 나누지 않기 때문에 continuous NRI 혹은 category-free NRI라고 이름이 붙었으며, cNRI라고 간략하게 표기한다.

$$cNRI = E[\text{sign}(p_2 - p_1 | Y = 1)] - E[-\text{sign}(p_2 - p_1 | Y = 0)]$$

p_1, p_2 는 각각 기존 예측모형과 새로운 예측모형으로부터 얻어진 예측확률을 의미한다. event 집단과 non-event 집단의 독립을 가정하면, cNRI의 검정통계량은 다음과 같이 나타낼 수 있다.

$$z_{cNRI} = \frac{c\hat{NRI}}{4 \left[\frac{\hat{P}(\text{up}|\text{event}) - \{1 - \hat{P}(\text{up}|\text{event})\}}{n_{\text{event}}} + \frac{\hat{P}(\text{up}|\text{non-event}) - \{1 - \hat{P}(\text{up}|\text{non-event})\}}{n_{\text{non-event}}} \right]}$$

검정통계량을 통해 유의한 결과를 얻게 되면, 새로운 예측모형과 기존 예측모형의 예측력에 유의한 차이가 있다고 할 수 있다.

21.3 Mean Square Error(MSE) & Mean Absolute Error(MAE)

작성중...

21.4 Cox and Snell의 결정계수(R^2) & Nagelkerke의 결정계수(R^2)

주어진 자료에서 구축된 특정 모형이 설명할 수 있는 종합적인 척도로, 모형이 설명하는 변동량(explained variation, R^2)을 사용할 수 있다. 즉, R^2 를 사용하여, 결과변수에 대한 독립변수의 영향력을 확인할 수 있으며, 해당 값이 클수록 모형에 사용된 독립변수들의 결과변수에 대한 설명력이 높다고 할 수 있다.

작성중...

21.5 Brier Score

Brier score는, 로지스틱 회귀모형에서 Nagelkerke의 결정계수와 함께 종합적 모형 수행력 척도로 많이 사용되는 지표이다. 각 관찰 값에 대해 실제 자료 y 와 예측 결과 \hat{p} 간의 차이를 식 19과 같이 계산한 뒤, 이들을 평균한 값을 의미한다.

$$y^*(1 - \hat{p})^2 + (1 - y)^*\hat{p}^2 \quad (19)$$

Brier score는 0에서 1 사이의 값을 가지며, y 와 \hat{p} 의 결과가 일치할 수록 0, 일치하지 않을수록 1에 가까워진다.

Brier score는 자료 y 의 발생률에 의존하며, 발생률이 작으면 brier score의 최대값도 작아진다는 단점이 있다.

이러한 단점을 보완하기 위해 scaled Brier score도 제안되었다[60]. Scaled Brier score는 식 19의 최대값을 사용하여, 다음 식 20과 같이 0부터 100 사이의 값을 가지도록 정의된다. 값이 클수록 모형의 적합도가 좋음을 의미한다.

$$1 - \frac{\text{Brier Score}}{\max[\text{Brier Score}]} \quad (20)$$

21.6 Hosmer-Lemeshow Test

Hosmer-Lemeshow 검정은 로지스틱 회귀모형으로부터 예측 확률을 계산한 후, 이를 10개의 quantiles로 나눈 후, 각 그룹 내에서 기대 빈도와 관찰 빈도 간의 교차표를 대상으로 χ^2 검정을 실시하는 것이다.

만일 검정 결과가 유의하지 않다면, 모형의 적합도가 높음을 의미한다.

Hosmer-Lemeshow 검정의 단점은, 표본 수가 커지는 경우, 관찰 값과 예측 값 사이에 유의한 차이가, 실제로는 없음에도 불구하고 마치 존재하는 것으로 평가되는 제1종 오류가 증가한다는 점이다.

22 반응 변수가 생존 시간인 경우

22.1 Ignoring time to event (logistic C)

이 방법은 censored 부분을 그냥 버리고, event가 발생한 survival time을 이용하여 ROC curve를 구하는 방법이다. 당연히 censored가 많을 때에는 적용하기 어려워진다.

22.2 Chambless and Diao의 C statistic

작성중... (참고문헌 : [21])

22.3 Gonen and Heller의 K

작성중... (참고문헌 : [28])

22.4 Harrell의 Concordance Index(C-index)

Harrell의 C-index는 binary 종속 변수에서 Hanley and McNeil(1982)[30]의 C-index를 구하는 비모수적인 방법을 생존 자료에 적용시킨 방법이다.

생존 시간을 \hat{T}_i , 중도절단 시간을 C_i 라고 정의하면, 관찰된 생존 시간은 $T_i = \min(\hat{T}_i, \delta_i)$ 로 정의할 수 있고, 중도절단 여부는 다음과 같다고 하자. (식 2의 리마인더이다.)

$$\begin{aligned} & \{(T_i, \delta_i), i = 1, \dots, n\} \\ & \text{단, } T_i = \min(\hat{T}_i, \delta_i) \text{ and } \delta_i = I(\hat{T}_i < C_i) \\ & \text{즉, } \delta_i = \begin{cases} 1 & \text{if } \hat{T}_i < C_i \text{ (즉, 관측된 경우)} \\ 0 & \text{if } \hat{T}_i \geq C_i \text{ (즉, right censored만 고려한다는 가정 하에 중도절단된 경우)} \end{cases} \end{aligned}$$

Harrell의 C-index를 정의하기 위해서는, 비교 가능한 쌍과 비교 불가능한 쌍을 정의해야 한다.

- i 번째 대상과 j 번째 대상이 모두 중도절단된 경우이거나, 모두 중도절단되지 않은 경우에 생존 시간의 비교가 가능하다.
- i 번째 대상과 j 번째 대상 둘 중 하나만 중도절단된 경우는 2가지 경우로 나뉘게 된다.
 - $T_i < T_j, \delta_i = 0$ 또는 $T_i > T_j, \delta_i = 0$ 인 경우는 생존 시간 비교가 불가능하다.
 - $T_i < T_j, \delta_i = 1$ 또는 $T_i > T_j, \delta_i = 1$ 인 경우는 생존 시간 비교가 가능하다.

비교 가능한 쌍에서

- $T_i < T_j, \delta_i = 1$ 일 경우 점수를 1
- $T_i > T_j, \delta_i = 1$ 일 경우 점수를 0

으로 두면, Harrell의 C-index는 다음과 같이 정의된다.

$$C_{\text{Harrell}} = \frac{\text{비교 가능한 쌍에서의 점수의 합}}{\text{비교 가능한 쌍의 수}}$$

조금 다르게 표현하면,

$$T_c = P(r(X_i) > r(X_j) \text{ and } T_i > T_j) + P(r(X_i) > r(X_j) \text{ and } T_i < T_j)$$

$$T_d = P(r(X_i) > r(X_j) \text{ and } T_i < T_j) + P(r(X_i) > r(X_j) \text{ and } T_i > T_j)$$

여기서, $r(X_k) = \text{risk score}$

라고 하면

$$C = \frac{T_c}{T_c + T_d}$$

C-index는 0.5에서 1 사이의 값을 가지며, 1에 가까워질수록 model discrimination이 잘 되는 것이라고 이야기한다.

만약, 다른고자 하는 데이터에 중도절단된 대상이 없다면, 모든 쌍들이 비교 가능한 쌍이 되며, C-index는 Mann-Whitney Statistics와 같아지게 된다.[43]

Note 33. C-index를 이용하여 두 모형을 비교할 경우, 통계적 유의성 검정

기존 모형의 C-index와, 새로운 factor가 추가된 모형의 C-index를 각각 구하여, 새로운 모형에서의 예측력이 더 향상되었는지 판단하기 위해서는, 두 C-index의 차이가 0인지의 여부를 검정하면 된다.

검정을 위해서는 두 모형에서 얻어진 C-index의 차이에 대한 분산을 구하여 t검정을 진행할 수 있다.

그러나 두 모형을 독립이라고 간주할 수 없기 때문에, C-index의 차이에 대한 분산을 구하기가 쉽지 않다.

따라서 Bootstrap 방법을 이용하여, 두 C-index의 차이에 대한 신뢰구간을 구하여, 95% 신뢰구간에 0이 포함되는지 여부로 예측력 향상을 판단할 수 있을 것이다.

Note 34. C-index의 95% 신뢰구간 구하기

1. Nam's Method 작성중... (참고문헌 : [54])

2. Modified Kendall's τ

3. Simplified Method

작성중...

22.5 Heagerty의 intergrated AUC

Heagerty and Zheng(2005)[32]는 x 축을 시점 t , y 축을 시점 t 에서의 ROC curve의 아래 면적인 AUC 값으로 갖는 그래프를 그려 그 아래 면적을 적분하여 iAUC(integrated AUC) 값을 구하였다.

ROC curve는 모든 가능한 지점에서 민감도(sensitivity)와 특이도(specificity)의 관계를 표현한 곡선이다. 어떤 시점 t 에서의 민감도와 특이도는 다음과 같이 정의된다.

$$\text{sensitivity}^I(c, t) = P(M_i > c | T_i = t) = P(M_i > c | dN_i^*(t) = 1)$$

$$\text{specificity}^D(c, t) = P(M_i \leq c | T_i > t) = P(M_i \leq c | dN_i^*(t) = 0)$$

위 식에서, M_i 는 공변량 벡터와 회귀계수 벡터의 곱인 $\beta'x$ 를 의미한다. 또한 $dN_i^*(t) = I(T_i \leq t) - I(T_i \leq t-)$ 를 나타내며, 시점 t 에서 사건이 발생한 경우 1, 아닌 경우 0을 의미한다.

시간 T 에 따른 ROC curve를 시간 종속 ROC curve라 하며, 작성중...

22.6 Uno의 C-index

작성중...

22.7 Pencina and Uno의 NRI

작성중...

22.8 Ctd Index

작성중... (참고문헌 : [7])

Part IX

Results of Survey: Read Paper and Run Software

[Neural Network with Life Data]

임성빈(????)[5]에 따르면, David Faraggi and Richard Simon(1995)[22]의 연구가 'Survival Data에 적용된 최초의 신경망 모형'이라고 소개되고 있다.

다만, Ping Wang et al.[51]에 따르면, David Faraggi and Richard Simon의 전후에 출판된 페이퍼들도 소개되고 있으며, 내가 직접 찾은 페이퍼도 있다. 이들도 함께 살펴보고자 한다.

[Neural Network with Image Data]

또한, 위 연구들은 임상 데이터나 유전자 데이터에 집중하고 있으나, 우리는 이미지 데이터도 고려하고 있다.

이미지 데이터를 이용한 Survival Prediction을 연구한 페이퍼 목록을 만들 수 있었다.

다만, 관련 연구가 많이 있을 것으로 기대되지만, 기대만큼 많은 시간을 투입하여 만든 목록은 아니므로, 조금 더 찾아볼 필요는 있다.

[Feature Extraction of Image Data]

한편, 이미지 데이터를 넣는다고 해도, 어떤 이미지를 넣어야하는지가 관건이다. 이에 임성빈 박사는 이미지 데이터의 feature extraction에 대한 고민을 하고 있었다.

나 또한 이에 공감하고 있으며, 현재는 이미지에 들어있는 주요 object에 대한 patch를 뽑아내는 방안을 생각해 보았다.

* 가급적이면 본 노트의 notation에 맞춰서 수정하고자 하였지만, 여전히 페이퍼의 notation을 따를 수도 있다. 이를 감안해서 보도록 하자.

23 Neural Network with Life Data

23.1 Peter M. Ravdin et al.(1992), A practical application of neural network analysis for predicting outcome of individual breast cancer patients

Peter M. Ravdin and Gary M. Clark(1992).

A practical application of neural network analysis for predicting outcome of individual breast cancer patients, *Breast Cancer Research and Treatment*, Vol. 22, No. 3, 285–293.

23.1.1 데이터셋

Dataset 1. Nichols Institute Oncology Research Network의 breast cancer 환자 데이터

여기에서 사용된 데이터셋은 Nichols Institute Oncology Research Network로부터 얻은 breast cancer 환자 데이터이다. 1373명의 환자 중 920명분을 training set, 453명의 데이터를 test set²⁹으로 삼았다. 이 데이터셋을 어디선가 다운로드할 수 있을 거라고 생각하지만, 아직 찾지는 못하였다.

23.1.2 요약

Peter M. Ravdin et al.[52]은, 분포 가정 없이 alive/death 상태 추정이 이루어진다는 점에서 non-parametric approach라고 할 수 있다. 논문에서도 Kaplan-Meier과 성능 비교를 진행한다.

Table 76: Peter M. Ravdin et al.에서 사용한 데이터셋의 변수

Variable	expl/resp	meaning	measure	Transform.
A	explanatory	Ploidy		
B		S-Phase	%	$\log(S\text{-Phase} + 1)$
C		ER	fmol/mg	$\log(ER + 3)$
D		PgR	fmol/mg	$\log(PgR + 5)$
E		Tumor Size	cm	$\log(Tumor\ Size + 0.5)$
F		Node positive	#	$\log(\# + 0.01)$
G		Age	years	
T		Time Interval		
S	response	Survival Status	0 or 1	

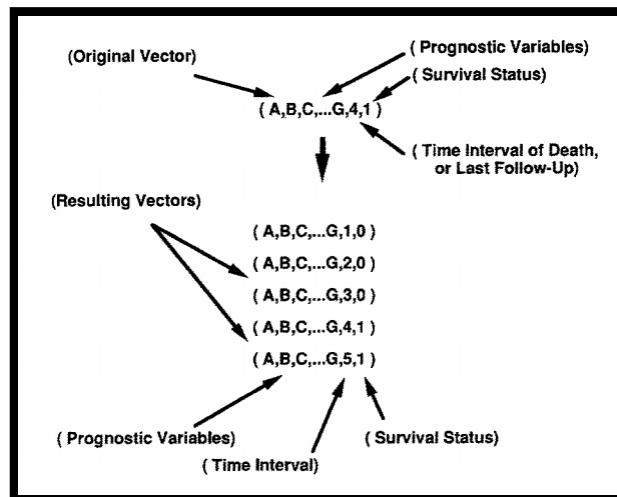


Figure 32: Peter M. Ravdin et al. 페이지에서의 데이터 전처리

Peter M. Ravdin et al.[52]은 data vector를 $(A, B, C, D, E, F, G, T, S)$ 로 정의하였으며, 각 element의 의미는 Table 76과 같고, 이들을 전처리하는 과정은 Figure 32와 같이 표현되어 있다.

²⁹페이지에서는 "validation set"이라고 표현하였음.

여기서,

$$T = \begin{cases} 1 & 90 \text{ percent at 12 months} \\ 2 & 80 \text{ percent at 18 months} \\ 3 & 70 \text{ percent at 27 months} \\ 4 & 60 \text{ percent at 40 months} \\ 5 & 50 \text{ percent at 60 months} \end{cases}$$

$$S = \begin{cases} 0 & \text{if surviving at maximum follow up} \\ 1 & \text{if dead with maximum follow up} \end{cases}$$

만일 $S = 0$ 이면, T 는 마지막으로 follow-up한 시간이 되며, $S = 1$ 이면 T 는 사망이 관측된 시간을 의미한다.

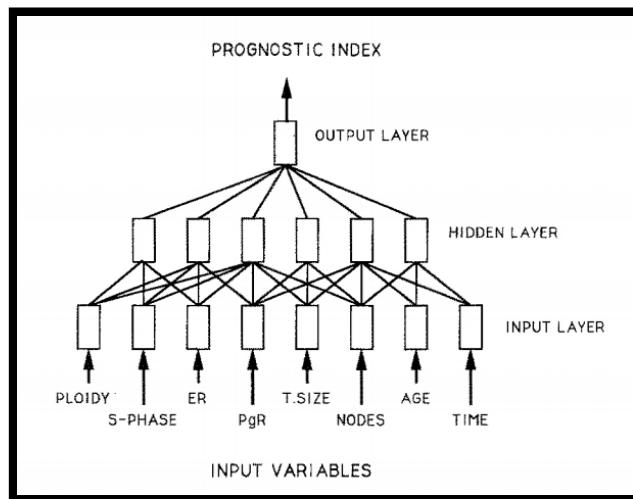


Figure 33: Peter M. Ravdin et al. 페이퍼에서의 신경망 모형

이렇게 전처리한 데이터를 Figure 33과 같이 적합하였다.

23.1.3 의견

알려진 생존분석 방법이 아닌, 순수한 neural network 방법에 life data를 적용한 방법이다. Neural network 자체만으로는 censored data에 대한 control이 된다는 보장이 없음에도 이를 검증하지 않고 적용했다는 비판을 할 수 있다.

23.2 Knut Liestol et al.(1994), Survival analysis and neural nets

Knut Liestbl, Per Kragh Andersen and Ulrich Andersen(1994), **Survival analysis and neural nets**, *Statistics in medicine*, Vol. 13, No. 12, 1189–1200.

23.2.1 데이터셋

Dataset 2. Drzewiecki and Andersen(1982)의 피부암 데이터

여기에서 사용된 데이터셋은 Drzewiecki and Andersen(1982)[23]의 연구를 통해 소개된 피부암 데이터이다. 205명의 환자 데이터이며, R의 'timereg' 패키지에 내장되어 있다.

이 데이터셋에 대한 자세한 요약은 Chatper 25.2.1를 참고하도록 하자.

23.2.2 요약

이 페이퍼에서는 본격적으로 Cox 비례위험모형을 소개하고 있다.

이 연구에서는 우선 feed forward neural network를 hidden layer가 없는 경우(one-layer network)와 hidden layer가 1개 있는 경우(two-layer) 두 경우를 소개하고 있다. 이는 이후에 나올 **regression models for survival data as neural net**과 관련이 깊으므로, 이 노트에도 실어놓았다.

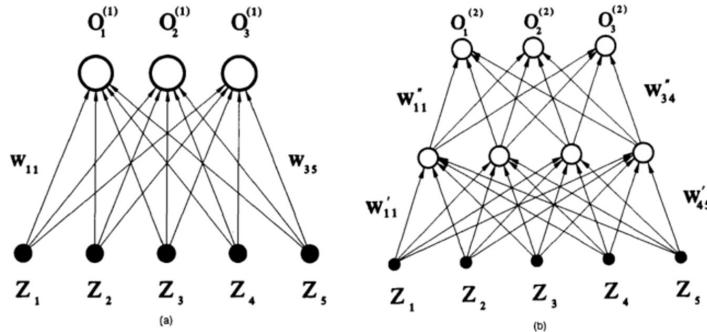


Figure 34: Liestbl et al.(1994)에서 소개하는 One-layer network와 Two-layer network

1. One-layer Network

$$O_k^{(1)}(z; w) = g \left(\sum_j w_{kj} z_j + w_{k0} \right) \quad (21)$$

여기서,

- $z = (z_1, z_2, \dots, z_p)$: Covariates
- $w_{k\ell}$: Weights for the connection from input node j to output node k

이에 따른 cost function은 $E^{(1)}$ 으로 정의된다.

$$E^{(1)} = \sum_{\ell} \sum_k [O_k(z_{\ell}; w) - Y_{k\ell}]^2$$

2. Two-layer Network

$$O_k^{(2)}(z; w) = g \left[\sum_{\ell} w''_{k\ell} h \left(\sum_j w'_{\ell j} z_j + w'_{\ell 0} \right) + w''_{k0} \right]$$

이에 따른 cost function은 $E^{(1)}$ 으로 정의된다.

$$E^{(2)} = - \sum_{\ell} \sum_k [Y_{k\ell} \log O_k(z_{\ell}; w) + (1 - Y_{k\ell}) \log(1 - O_k(z_{\ell}; w))]$$

이 페이퍼에서는 위의 cost function을 최소화하는 w 를 찾는 것이 목표라고 서술하고 있다.

3. 그리고 regression models for survival data as neural net을 다음과 같이 정의하고 있다.

우선, hazard function이 $\lambda(t)$ 이고, 반응 변수인 survival time이 T 일 때,

$$\lambda(t) = \frac{T < t + \delta t | T \geq t}{\delta t}$$

그러면 Cox 비례위험모형은 다음과 같이 서술할 수 있다.

$$\lambda(t) = \lambda_0(t) e^{\beta^T z}$$

여기서,

- $z = (z_1, \dots, z_p)^T$: Covariates
- $\lambda_0(t)$: 알려지지 않은 baseline hazard for individuals with $z = 0$
- $\beta = (\beta_1, \dots, \beta_p)^T$: regression coefficients

T 가 disjoint intervals I_k ($k = 1, \dots, m$)을 이용하여 그룹을 나누었다고 하자.

이 때 I_k 는 $t_{k-1} \leq t < t_k$ ($0 = t_0 < t_1 < \dots < t_m < \infty$)로 표시할 수 있으며, q_k 는 다음과 같은 조건부 생존률(conditional survival probabilities)로 나타낸다.

$$q_k = P(T \geq t_k | T \geq t_{k-1})$$

2개의 그룹으로 나누어진 경우, 다음과 같은 Cox 비례위험 모형을 고려할 수 있다.

$$\frac{p_k(z)}{q_k(z)} = \frac{p_k(0)}{q_k(0)} e^{\beta^T z} = e^{\beta^T z + \theta_k} \quad (22)$$

$$\text{여기서, } p_k = 1 - q_k, \quad \theta_k = \log \left(\frac{p_k(0)}{q_k(0)} \right)$$

Prentice and Gloeckler(1978)[52]와 Kalbfleisch and Prentice(1980)[41]의 방법에 따라, 식 (22)이 다음과 같다.

$$-\log(q_k(z)) = \Lambda_{0k} e^{\beta^T z} = e^{\beta^T z + \theta_k}$$

$$\text{여기서, } \Lambda_{0k} = e^{\theta_k} = \int_{t_{k-1}}^{t_k} \lambda_0(t) dt$$

따라서,

$$E = \sum_{i=1}^n \sum_{k=1}^{m_i} [D_{ki} \log p_k(z_i) + (1 - D_{ki} \log q_k(z_i))]$$

여기서,

- D_{ki} : the indicator for individual i dying in the interval I_k
- $m_i \leq m$: the number of intervals in which individual i is observed

이제 여기에 neural net을 붙이기 위해, 다음과 같은 가정을 하였다.

$$w_{1j} = w_{2j} = \dots = w_{mj} = \beta_j \quad (j = 1, \dots, p)$$

$\beta = (\beta_1, \dots, \beta_p)^T$, $\theta_k = w_{k0}$ 라고 하자. 그러면 식 (21)를 다음과 같이 출력할 수 있다.

$$O_k^{(1)}(z; \beta, \theta_k) = g(\beta^T z + \theta_k)$$

23.2.3 의견

이 페이퍼에서는 위험이 일정한 것으로 가정하고(즉, 수명 함수가 지수 분포를 따르는 것으로 가정) 출력 노드가 각 간격마다 설정된 시간 간격으로 그룹화하였다. Hidden layer가 없고 동일한 입력 노드에서 모든 출력 노드까지 동일한 가중치 조건에서 neural network는, 각 개인의 조건부 이벤트 확률을 예측하도록 학습되며, 그 결과는 grouped Cox 비례위험모형 버전과 동일하다.

동일한 데이터의 시간에 따른 ”일정하지 않은 위험“은, 각 출력 노드에 대한 가중치가 다를 때까지 모델링될 수 있다. Hidden layer를 추가하면, 비선형 공변량 효과가 있는 Cox neural network라는 하이브리드 형식으로 만들어진다. 비선형성 정도는 Hidden layer 수와 activation function의 선택에 달려 있다.

23.3 David Faraggi and Richard Simon(1995), A Neural Network Model for Survival Data

David Faraggi and Richard Simon(1995),
A Neural Network Model for Survival Data, Statistics in Medicine, Vol. 14, 73-82.

23.3.1 데이터셋

Dataset 3. Byar and Green(1980)의 전립선암 데이터

여기에서 사용된 데이터셋은 Byar and Green(1980)[20]의 연구를 통해 소개된 전립선암 데이터이다. 506명의 환자 데이터이며, R의 'clustMD' 패키지에 내장되어 있으나, clustMD 패키지에서는 475명의 환자 데이터를 제공하여 준다. 이 데이터셋에 대한 자세한 요약은 Chatper 25.2.2를 참고하도록 하자.

다만, David Faraggi and Richard Simon은 238명의 환자 데이터를 임의로 선택하여 training set으로, 나머지 237명의 데이터를 test set³⁰으로 삼은 상황이므로, 506명의 환자 데이터가 갖추어졌더라도 완벽한 재현은 불가능할 것으로 여겨진다.

23.3.2 요약

이 페이지에서 또한 Cox 비례위험모형에 neural network를 적용하고 있다.

1. Neural Network

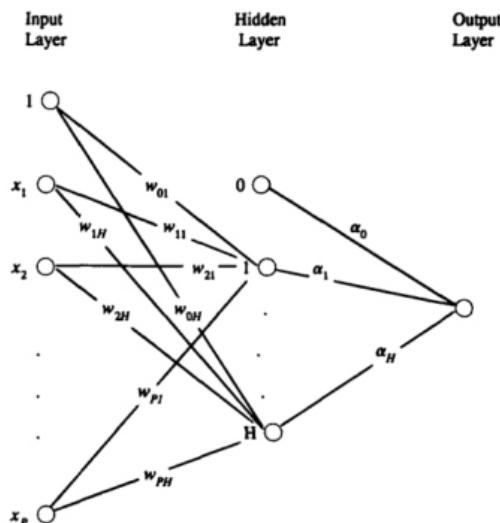


Figure 35: Single hidden layer neural network to single output

$$g(x_i, \theta) = \alpha_0 + \sum_{h=1}^H \alpha_h f(w'_h x_i) = \alpha_0 + \sum_{h=1}^H \frac{\alpha_h}{1 + e^{-w'_h x_i}}$$

여기서,

- Covariates: $x_i = (x_{i0}, x_{i1}, \dots, x_{iP})'$ ($i = 1, \dots, n$), ($p = 0, \dots, P$)
- Squashing function: $f()$ ³¹
- Hidden node: $h = 1, \dots, H$
- The vector of weights: $w_h = (w_{0h}, w_{1h}, \dots, w_{Ph})$
- The vector of unknown parameters:³²

$$\theta = (w_{01}, w_{11}, \dots, w_{P1}, w_{02}, \\ w_{12}, \dots, w_{P2}, \dots, w_{0H}, \\ w_{1H}, \dots, w_{PH}, \\ \alpha_0, \alpha_1, \dots, \alpha_H)'$$

³⁰페이지에서는 "for validating the models"이라고 표현하였음

³¹이 논문에서는 activation function을 squashing function이라고 표현하고 있다.

³²이 파라메터의 개수는 다음과 같이 계산할 수 있다. $m = (H + 1) + (P + 1)H$

2. Neural Network Survival Models

Hazard function이 다음과 같다.

$$h(t, x_i) = h_0(t)e^{\beta x_i}$$

파라메터의 벡터인 β 는 다음과 같은 부분 가능도 함수(maximizing the partial likelihood function)를 최대화하여 추정할 수 있다.

$$L_c(\beta) = \prod_{i \in w} \frac{e^{\beta x_i}}{\sum_{j \in R_i} e^{\beta x_j}}$$

이 논문에서는 이 partial likelihood를 network $g(x_i, \theta)$ 로 바꾸었다. 즉 Hazard function이 다시 다음과 같이 정의된다.

$$h(t, x_i) = h_0(t)e^{g(x_i, \theta)}$$

또한, 추정해야 할 부분 가능도 함수(partial likelihood function)도 다음과 같이 수정된다.

$$L_c(\theta) = \prod_{i \in w} \frac{\exp \left[\sum_{h=1}^H \frac{\alpha_h}{1+e^{-w_h' x_i}} \right]}{\sum_{j \in R_i} \exp \left[\sum_{h=1}^H \frac{\alpha_h}{1+e^{-w_h' x_j}} \right]}$$

이 논문에서는 부분 최대 가능도 추정량을 구하기 위해 Newton-Raphson Method를 사용하였다고 소개하고 있다.

또한, 이러한 방법은 가속수명모형(the accelerated failure time model), Buckley-James model에도 적용할 수 있다고 소개하고 있다.

23.3.3 의견

Cox 비례위험모형에 covariate의 일반적인 선형 조합 대신, 비선형 함수를 허용하도록 한 방법이다. 이 방법은 Cox 모델의 비례 위험 요소를 유지하지만, 단순한 Cox 모델에서 누락될 수 있는 입력 데이터의 복잡성과 interaction을 모델링하는 기능을 제공한다.

23.4 Stephen F. Brown et al.(1997), On the use of artificial neural networks for the analysis of survival data

Stephen F. Brown, Alan J. Branford and William Moran(1997),
On the use of artificial neural networks for the analysis of survival data,
Neural Networks, IEEE Transactions Vol. 8, Mo. 5, 1071–1077.

23.4.1 데이터셋

이 페이퍼에서는 Choong et al.³³과 deSilva et al.(1994)³⁴의 방법을 개선하면서, 이 두 페이퍼에서 사용한 피부암 및 유방암 데이터를 그대로 사용하고 있다. 다만 이 두 논문은 검색하기가 쉽지 않았으며, 데이터셋 또한 무엇인지 정확히 알 수 없었다.

23.4.2 요약

여기에서는 Choong et al.(1993)과 deSilva et al.(1994)의 방법을 follow up하고, 이들의 방법을 적용할 시 bias가 발생한다는 점을 개선하기 위한 새로운 접근을 시도하고 있다.

이 페이퍼에서는 Kaplan-Meier와 Cox 비례위험모형에 neural network를 붙이는 방식이다.

1. Some Results from Survival Statistics

- Survival Function:

$$S(t) = P(T \geq t) = e^{-\int_0^t h(t') dt'} \quad (23)$$

* $S(0) = 1, S(\infty) = 0$

- Hazard function:

$$h(t) = \lim_{\Delta t \rightarrow 0+} \frac{P(T \leq t + \Delta t | T \geq t)}{\Delta t}$$

- 생존 곡선은 Kaplan-Meier 방법에 기반한 최대 가능성 추정법(maximum likelihood approach)으로 추정될 수 있다.

$$\hat{S}(t) = \prod_{k|t_k < t} \left(\frac{n_k - d_k}{n_k} \right)$$

여기서,

– d_k : The number of subjects that fail at time t_k .

– n_k : The total number of subjects that fail or are censored at time t_k or later.

- x 를 입력 벡터(a vector of inputs), β 를 파라메터 벡터(a vector of parameters)라고 하고, Cox 비례위험모형에 기반한 생존 함수(survival function), 위험 함수(hazard function)를 다음과 같이 정의할 수 있다.

$$S(t; x) = S_0(t) e^{\beta^T x}$$

○] 때, $S_0(t) = e^{-\int_0^t h_0(t') dt'}$

$$h(t; x) = h_0(t) e^{\beta^T x}$$

○] 때, $h_0(t)$: arbitrary baseline hazard function

³³P. L. Choong, C. J. S. deSilva, J. Taran, P. Heenan and H. Dawkins(1993),
Survival analysis using artificial neural network, Proc. 1st Australia and New Zealand Conf. Intell. Inform. Syst., 283-287.

³⁴C. J. S. deSilva, P. L. Choong and Y. Attikiouzel(1994),
Artificial neural networks and breast cancer prognosis, Australian Comput. J., Vol. 26, 78-81.

2. Previous ANN Approaches

$$E = \frac{1}{2} \sum_{i=1}^m (t_i - t_{out})^2$$

$$\implies t_{out} = \sum_{i=1}^m \frac{t_i}{m}$$

여기서,

- t_i : the training set outputs
- m : the sample size

Choong et al.(1993)과 deSilva et al.(1994)는, artificial neural network(ANN)를 사용하여, 피부암과 유방암에 따른 사망률(skin and breast cancer mortality)을 연구했다. 그들은 failure time과 censoring time을 사용하여 네트워크를 훈련시켰다고 한다. 그러나 ANN을 사용하여 중도절단된 시간보다 긴 생존 시간을 예측할 때, 예측 오류가 0으로 간주되고, 네트워크 가중치가 업데이트 되지 않는 문제가 있었다고 한다. 이는 censored data에 의해 도입되는 bias를 다소 제거하는 효과를 주겠지만, 여전히 남아있는 bias의 크기를 평가할 방법을 주지 않는다.

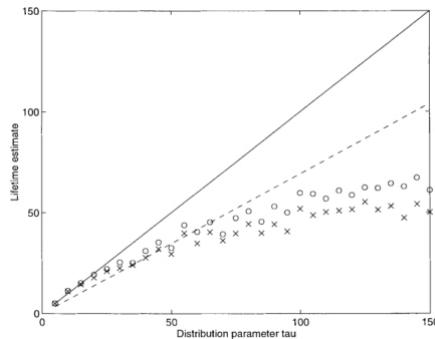


Figure 36: Choong et al.(1993)의 방법으로 추정된 observed failure의 평균(O 표시)와 censoring time의 평균(X 표시). 실선은 true mean이고 점선은 true median이다.

이 페이퍼에서는 위 두 페이퍼에서 논의된 bias의 효과를 보다 명확하게 보기 위해, 수명 분포가 지수분포를 갖는 집단으로부터 시뮬레이션을 수행했다.³⁵

$$S(t) = e^{-\frac{t}{\tau}} \quad (\text{여기서, } \tau > 0)$$

이 τ 값을 5부터 150까지 5씩 증가시키면서, 각각의 값마다 100번의 censoring time이 무작위로 배치된다. 또한 경과 시간이 100에 도달하기 전에 failed되지 않거나(not failed) censored된 객체는 censored로 간주되었다.

시뮬레이션의 결과는 Figure 36에서 확인할 수 있다.

- $\tau < 30$ 일 때 ANN는 비교적 평균을 잘 추정하는 모습을 볼 수 있다.
- $30 < \tau < 60$ 일 때, 추정 결과는 평균과는 거리가 멀어졌지만, 중앙값과는 가까운 모습을 볼 수 있다.
- $60 > \tau$ 일 때, 추정 결과가 평균과 중앙값 모두에 미치지 못하는 것을 볼 수 있다.

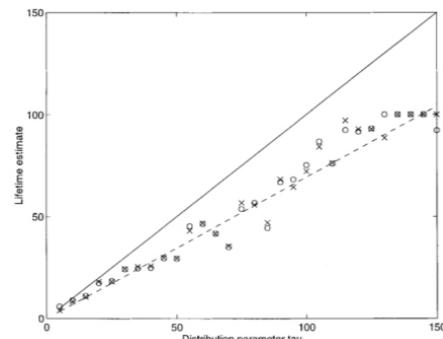


Figure 37: Kaplan-Meier 방법으로 추정된 observed failure의 평균(O 표시)와 censoring time의 평균(X 표시). 실선은 true mean이고 점선은 true median이다.

한편, 전통적인 non-parametric 방법인 Kaplan-Meier 방법으로 추정할 경우, Figure 37와 같이 추정량이 median을 비교적 정확하게 추정하는 것을 볼 수 있다. Choong et al.(1993)으로 추정한 Figure 36와 비교되는 부분이다.

³⁵Chapter 16.1.11에서는 지수분포에 대한 생존함수를 $S(t) = e^{-\lambda t}$ 로 서술하고 있고, 별도로 $\lambda = \frac{1}{\theta}$ 인 경우에 대해서도 서술하고 있다. 또한 MLE의 성질에 따라 $\hat{\lambda}_{MLE} = \frac{1}{\theta_{MLE}}$ 라고 할 수 있다. 이 페이퍼에서는 notation을 θ 대신 τ 로 기술한 것으로 보면 된다.

3. A New Approach

이 논문에서는 새로운 접근법으로, 객체의 수명을 추정하는 대신에, 객체의 $S(t)$, 즉 생존 함수를 추정하고자 시도하였다. $S(t)$ 는 ANN을 이용하여 hazard function을 계산한 뒤, 이를 $S(t)$ 와의 관계를 이용하여 구할 수 있다.³⁶

식 (23)의 discretized version은 다음 식 (24)과 같이 표현할 수 있다.

$$\begin{aligned}\tilde{S}(t_j) &= e^{-\sum_{k=1}^j \tilde{h}_k} \\ &= \prod_{k=1}^j (1 - h_k) \\ &= \tilde{S}(t_{j-1})(1 - h_j)\end{aligned}\quad (24)$$

이 때,

- $\tilde{h}_k \geq 0$
- $h_k = 1 - e^{-\tilde{h}_j}$
- $\tilde{S}(0) = 1$

이제 i 번째 객체 각각에 대해, 경험적 생존 함수(empirical survival function)³⁷ $S_i(t)$ 를 고려해보자.

t_f 가 어느 객체가 fail이 일어난 시간이라면,

$$S_i(t) = \begin{cases} 1 & t \leq t_f \\ 0 & t > t_f \end{cases} \quad (25)$$

식 (24)의 \tilde{S} 와 비교한다면, 식 (25)은 다음을 만족하는 경험적 위험률 요소(empirical hazard components) h_j^i 를 요구한다는 것이다.

- $j < j_{\text{crit}}$ must be zero
- One of $j = j_{\text{crit}}$ must be zero
where j_{crit} is the smallest value of j such that $t_f < t_j$

그리고 i 번째 객체 각각에 대해, t_c 가 중도절단시간이라고 할 때, 경험적 생존 함수는 다음과 같이 표시할 수 있다.

$$S_i(t) = \begin{cases} 1 & t \leq t_c \\ \text{unknown} & t > t_c \end{cases}$$

이 때 h_j^i 는 다음과 같이 요구된다.

- $j < j_{\text{crit}}$ must be zero
where j_{crit} is the smallest value of j such that $t_c < \frac{1}{2}(t_{j-1} + t_j)$

이제 추정된 위험률을 h_j 라고 할 때, 앞으로 최소화할 대상이 될 error term은 원래 다음과 같이 정의될 것이다.

$$E = \frac{1}{2} \sum_{i=1}^m (h_j^i - h_j)^2$$

여기서, m : sample size

이 논문에서는 이를 다음과 같이 확장하였다.

$$E = \left[\frac{1}{2} \sum_{k=1}^{n_f} (1 - h_j)^2 + \frac{1}{2} \sum_{k=1}^{n_c} (0 - h_j)^2 + \frac{1}{2} \sum_{k=1}^n (0 - h_j)^2 \right]$$

여기서,

- n_f is the number of subjects(객체) with failure times between $(j-1)\Delta t$ and $j\Delta t$
- n_c is the number of subjects(객체) that are censored between times $(j-\frac{1}{2})\Delta t$ and $j\Delta t$
- n is the number of subjects that have not failed or been censored by time $j\Delta t$

h_j 의 최소값은 $h_j = \frac{n_f}{n+n_f+n_c}$

23.4.3 의견

절단이 없는 부분에 대한 추정 결과와 중도절단된 부분에 대한 추정 결과를 분리해서 본 것이 적절하여 보였다.
데이터와 참고문헌을 구할 수 없어서, 논문 자체가 재현할 수가 없는 상황이다.

³⁶생존 함수와 위험률 함수와의 관계는 Chapter 10을 참고하자.

³⁷왜 ”경험적”이라는 말이 붙었는지는 아직 모르겠다.

23.5 Elia Biganzoli et al.(1998), Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach

Elia Biganzoli, Patrizia Boracchi, Luigi Mariani and Ettore Marubini(1998),
Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach, *Statistics in medicine*, Vol. 17, No. 10, 1169–1186.

23.5.1 데이터셋

- Head and neck cancer trial from Efron B.(1988)[24]

Efron B.(1988)[24]의 페이퍼³⁸에 데이터가 그대로 실려있다.

- Veteran's administration lung cancer trial from Kalbfleisch and Prentice (1980)[41] (veteran)

이 페이퍼에서는 R의 randomSurvivalForest 패키지의 veteran 함수로 제공이 된다고 설명하고 있다. 다만, randomSurvivalForest는 현재³⁹ 기준으로 randomForestSRC 패키지로 이름이 변경되었다.

23.5.2 코드

Venable and Ripley(2004)[65]⁴⁰이 제공한 R의 nnet 패키지를 활용하였다고 소개하고 있다. nnet의 raw code는 다음 링크를 통해 확인할 수 있다.

<https://github.com/cran/nnet/blob/master/R/nnet.R>

23.5.3 요약

이 논문의 첫 문장에서 "flexible modeling in survival analysis"라는 표현이 있는데, 이는 "time-varying model"을 고려한다는 의미가 된다. Giuliana Cortese et al.(2010)[29]을 참고하자.⁴¹

여기에서는 partial logistic regression에 neural network를 붙이는 접근을 하였다.

1. Artificial Neural Networks and Generalized Regression Models

우선 feed forward ANN에 대해서 이 챕터에서는 짚고 있다. Feed forward ANN은 non-linear multivariate regression method와 동치이다.

Node $h = 1, 2, \dots, H$, input node $j = 1, 2, \dots, J$, output node $k = 1, 2, \dots, K$ 라고 하고, observed input and observed responses를 y_{ik}^o 라고 표시하자. 모델에 의한 추정량(outputs)은 \hat{y}_{ik} 로 표기한다. 각 노드에서는 input x_{ij} 에 대한 weights w_{jh} 와 constant(bias) α_h 계산된다.

$$\hat{y}_k(x_i, w) = \phi_o \left(\alpha_k + \sum_{h=1}^H w_{hk} \phi_h \left(\alpha_h + \sum_{j=1}^J w_{jh} x_{ij} \right) \right)$$

Hidden node를 위한 activation function ϕ_h 는 다음과 같다.

$$\phi_h(u) = \frac{\exp(u)}{1 + \exp(u)}$$

GLM에서는 ϕ 를 link function이라고 부르며, 분포 가정에 따라 정의되는 형태가 다르다. 또한 이 link function은 위 activation function과 동치이다.

³⁸참고로 이 페이퍼는 다음 링크에서 재현이 되기도 하였다. 그냥 테이블만 보고는 데이터 재현이 곤란하다면, 이를 참고하도록 하자.

<http://blog.revolutionanalytics.com/2016/04/reading-efron-with-r-1.html>

³⁹2017/12/29

⁴⁰내가 참고한 것은 2004년 4th edition이며, 이 노트에도 이를 기재하였다.

다만, 저자들은 1994년에 발간된 초판을 참고한 것으로 보인다.

⁴¹Cox regression은 생존분석에서 Proportional Hazard 모형을 의미한다. 이는 소위 partial likelihood estimation을 통해 covariate의 계수를 추정하는데, partial Likelihood estimation은 쉽게 말해서, risk set을 고려하는 MLE의 일종이다. partial likelihood estimation을 하면 time dependent한 변수를 고려하기가 쉽다.

우선, time fixed는 fixed over time이라는 의미이다. 즉, 시간에 따라 fixed되어있다는 의미이다. 그리고, time dependent는 varying over time이라는 의미인데 time varying 변수라고 부르기도 한다.

예를 들어 어떤 sample이 운동을 하지 않다가 t 시간부터 운동을 하기 시작했다고 할 때, 이를 고려하는 변수는 time dependent 변수이고, 이를 고려하지 않으면 time fixed 변수이다. 즉, 그 sample에 대해 운동이라는 변수에 1이나 0으로 coding하면 time fixed 변수이다. t 를 고려해서 t 이전에는 0, t 이후에는 1로 coding하는 식이면 time dependent 변수이다.

Time dependent를 고려하게 되면 더욱 많은 정보(information)를 통해 분석하는 것이므로 더욱 적절한 분석이 된다. 즉, time dependent의 성격이 강하면 time dependent하도록 모형추정을 해야 한다.

더러는 time-dependent한 변수와 time-independent한 변수가 혼재되어 있는 상황이라면, time dependent 한 세팅에서 특정 변수는 fixed 되도록 설정하고 몇 가지 변수만 time varying 하게 한 뒤 분석을 수행하면 된다. 즉, time-dependent한 변수에 대해서만 time varying하도록 고려하면 됩니다.

생존분석은 $\log(time) = \dots$ 와 같은 모형인데 종속변수가 time이므로, 이와 같이 기본적으로 time varying covariate를 생각할 필요가 있다. Time varying covariate를 고려하기 위해서는 소위 counting process방식으로 dataset을 변경한 다음 분석하거나 프로그래밍할 때, 즉 partial Likelihood를 구성할 때 time varying하도록 고려하면 된다.

모델의 파라메터인 weights w 를 추정하기 위해 error function을 최소화하는 과정을 거친다. 가장 많이 사용되는 quadratic error는 다음과 같이 정의된다.

$$E = \sum_{k=1}^k \sum_{i=1}^n (\hat{y}_k(x_I, w) - y_{ki}^o)^2$$

다만, binary classification 문제를 풀기 위해서는, 다음과 같이 cross-entropy error를 이용한다.

$$E = - \sum_{k=1}^k \sum_{i=1}^n [y_{ki}^o \log \hat{y}_k(x_I, w) + (1 - y_{ki}^o) \log(1 - \hat{y}_k(x_I, w))] \quad (26)$$

Error function 식 (26)의 절대 최소값은, 는 K 개 output에 대한 n 개의 subjects에 대해 $y_{ik}^o = \hat{y}_{ik}$ 일 때 발생하며, 다음 식 (27)와 같이 표현된다.

$$E_{\min} = - \sum_{k=1}^k \sum_{i=1}^n [y_{ki}^o \log y_{ki}^o + (1 - y_{ki}^o) \log(1 - y_{ki}^o)] \quad (27)$$

2. Discrete Time Models for Survival Data

Continuous-time survival data일 때, hazard function $h(t)$ 는 다음과 같이 정의된다.

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(T < t + \Delta t)}{\Delta t}$$

다만 이 페이퍼에서는 discrete time을 고려하였다. 즉, L 개로 분할 된 $0 < t_1 < t_2 < \dots < t_L$ 을 고려하였다. 이는 continuous survival times $\ell = 1, 2, \dots, L$ 이 있을 때, $A_\ell = (t_{\ell-1}, T_\ell]$ 로 구분한 것이다. 이 때 $t_0 = 0$ 이고, ℓ_i 는 i 번째 subject에 있어 i 번째 시간의 interval 끝자락에서 식별된 것이다.

적절한 survival function을 다음과 같이 정의하고

$$S(t_\ell) = P(T > t_\ell)$$

discrete probability function을 다음과 같이 정의하면

$$f_\ell = P(T \in A_\ell) = S(t_{\ell-1}) - S(T_\ell) \quad (28)$$

discrete hazard rate h_ℓ 은 다음과 같은 conditional failure probability로 정의할 수 있다.

$$h_\ell = P(T \in A_\ell | T > t_{\ell-1}) = \frac{f_\ell}{S(t_{\ell-1})} \quad (29)$$

A_ℓ 을 무한대로 보내면, conditional failure probability h_ℓ 은 continuous hazard function $h(t)$ 에 수렴하게 된다. 식 (28)과 식 (29)로부터 다음 $S(t)$ 를 정의할 수 있다.

$$S(t) = \prod_{\ell: t_i \leq t} (1 - h_\ell)$$

i 번째 subject에 대한 가능성도 함수에 대한 기여는, 시간 간격에 대한 conditional survival probability와 관심 이벤트가 발생하는 간격 A_ℓ 에서의 conditional failure probability의 곱으로 주어진다.

우종도절단만을 고려했을 때, uncensored subjects의 set U 가 기여하는 부분은

$$P(T_i \in A_{\ell_i}) = S(t_{\ell_i}) = \prod_{\ell=1}^{\ell_i-1} (1 - h_{i\ell})$$

censored subjects C 가 기여하는 부분은 다음과 같이 표현된다.

$$P(T_i > t_{\ell_i}) = S(t_{\ell_i}) = \prod_{\ell=1}^{\ell_i} (1 - h_{i\ell})$$

Censoring indicator를 $d_{i\ell} \circ$ 라고 하고, $d_{i\ell} = 1$ 이면 uncensored subjects를, $d_{i\ell} = 0$ 이면 censored subject를 포함한다고 했을 때, 최종적으로 likelihood는 다음과 같이 표현할 수 있다.

$$L = \prod_{i=1}^n \prod_{\ell=1}^{\ell_i} h_{i\ell}^{d_{i\ell}} (1 - h_{i\ell})^{1-d_{i\ell}} = \prod_{\ell=1}^L \prod_{i \in R_i} h_{i\ell}^{d_{i\ell}} (1 - h_{i\ell})^{1-d_{i\ell}} \quad (30)$$

여기서, R_ℓ 은 the set of individuals at risk in the ℓ th interval of time. $R_\ell \circ$ Bernoulli trial이라고 한다면, 식 (30)을 다음과 같이 고쳐쓸 수 있다.

$$L = \prod_{\ell=1}^L \binom{n_\ell}{s_\ell} h_\ell^{s_\ell} (1 - h_\ell)^{n_\ell - s_\ell}$$

여기서, n_ℓ 은 the number of subjects at risk, s_ℓ 은 the number of failures in the time interval ℓ .

이제 covariates을 고려해보자. David Cox는 grouped survival times를 위해 다음과 같은 proportional odds model을 제안하였다.

$$\frac{h_\ell(x_i)}{1 - h_\ell(x_i)} = \frac{h_\ell(0)}{1 - h_\ell(0)} \exp(\beta^T x_i)$$

Baseline hazard rate를 $\theta_\ell = \log\left(\frac{h_\ell(0)}{1 - h_\ell(0)}\right)$ 로 치환하면, 위 식으로부터 $h_\ell(x_i)$ 를 다음과 같이 표현할 수 있다.

$$h_\ell(x_i) = \frac{\exp(\theta_\ell + \beta^T x_i)}{1 + \exp(\theta_\ell + \beta^T x_i)}$$

3. Partial Logistic Regression Models with ANN (PLANN)

식 (30)에 log를 써우면 다음과 같다.

$$L = - \prod_{i=1}^n \prod_{\ell=1}^{\ell_i} [d_{il} \log h_{il} + (1 - d_{il}) \log(1 - h_{il})] \quad (31)$$

이와 같은 error function을 사용한다는 것은 hidden layer가 없고, logistic activation function ϕ_o 을 사용한 neural network를 사용한다는 것과 같다.

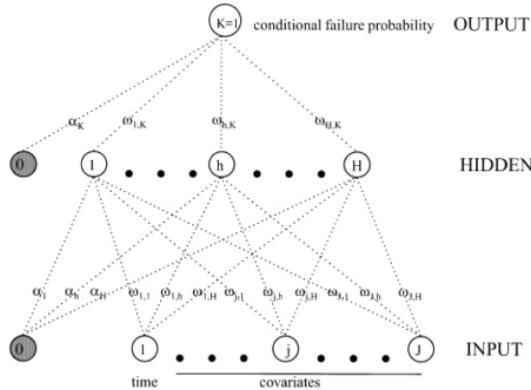


Figure 38: Feed forward neural network model for partial logistic regression(PLANN)

PLANN model은 하나의 input node j 에 설명변수 x_j 를 대입하고, 추가로 time interval a_ℓ input node를 추가한 것이다. 이는 Figure 3과 같은 그림으로 나타낼 수 있다. 결국 이 페이퍼는 Time interval node를 추가함으로써 time varying term을 control하겠다는 아이디어인 것이다.

설명변수가 categorical data라면, n_m 개 subjects와 s_m 개 events를 가진 공변량 집에 기초하여 $m = 1, 2, \dots, M$ 의 디자인 셀에 그룹화할 수 있다. 따라서 각 셀에 대한 empirical estimate \hat{h}_m 은 다음과 같이 얻을 수 있다.

$$\hat{h}_m = \frac{s_m}{n_m}$$

이러한 경우, error function을 다음과 같이 사용한다.

$$L = - \sum_{m=1}^M \left[\hat{h}_m \log \frac{h_m(x_i, a_i)}{\hat{h}_m} + (1 - \hat{h}_m) \log \frac{1 - h_m(x_i, a_i)}{1 - \hat{h}_m} \right] n_m \quad (32)$$

4. PLANN and Previously Proposed ANN Approaches for Grouped Survival Data

작성중...

23.5.4 의견

이 페이퍼에서 사용했다고 소개하고 있는 R의 nnet 패키지에 속해 있는 nnet 함수에 censored 인자가 있는데, 이것이 PLANN 모델과 무슨 관계가 있는지는 확인해볼 필요가 있다. censored 인자가 없이, time interval a_ℓ 을 그대로 nnet 함수에 넣는 방식으로 실행할 수도 있기 때문이다.

23.6 Anny Xiang et al.(2000), Comparison of the performance of neural network methods and Cox regression for censored survival data

Anny Xiang, Pablo Lapuerta, Alex Ryutov, Jonathan Buckley and Stanley Azen(2000), **Comparison of the performance of neural network methods and Cox regression for censored survival data**, *Computational Statistics & Data Analysis*, Vol. 34, 243–257.

23.6.1 데이터셋

이 페이퍼에서는 9가지 synthetic dataset를 생성하여 실험에 이용하였다. 이 데이터셋에 대한 자세한 요약은 Chatper ??를 참고하도록 하자.

Dataset 4. Anny Xiang et al.(2000)의 Synthetic Data

- Design 1: No censoring. $n = 100$, 50 replications.

$$\gamma = \exp [x_1 + 0.25x_2]$$

$$x_1 \sim \text{Bernoilli}(0.5) \quad x_2 \sim N(0, 1)$$

- Design 2: Average censoring for training and testing sets = 19% each. $n = 100$, 50 replications.

$$\gamma = \exp [x_1 + 0.25x_2]$$

$$x_1 \sim \text{Bernoilli}(0.5) \quad x_2 \sim N(0, 1)$$

- Design 3: Average censoring for training and testing sets = 20% each. $n = 100$, 50 replications.

$$\gamma = \exp [x_1 + 0.25x_2 + 0.2x_1x_2]$$

$$x_1 \sim \text{Bernoilli}(0.5) \quad x_2 \sim N(0, 1)$$

- Design 4: Average censoring for training and testing sets = 32%. $n = 100$, 50 replications.

$$\begin{aligned} \gamma = \exp[2(x_1 + x_2) + 0.5x_3 + 1.0x_4 \\ + 1.0(x_1x_2 + x_1x_3) \\ + 0.5(x_1x_4 + x_2x_3 + x_2x_4 + x_3x_4) \\ + 0.5(x_1x_2x_3 + x_1x_2x_4 + x_1x_3x_4 + x_2x_3x_4 + x_1x_2x_3x_4)] \end{aligned}$$

$$x_1 \sim \text{Bernoilli}(0.25) \quad x_2 \sim \text{Bernoilli}(0.5) \quad x_3 \sim N(0, 1) \quad x_4 \sim N(0, 1)$$

- Design 5: Same as Design 4. Average censoring for training and testing sets = 70%. $n = 100$, 50 replications.

- Design 6: Average censoring for training and testing sets = 47%. $n = 100$, 50 replications.

$$\gamma = \exp [0.5(x_1 + x_2) + 0.25(x_3 + x_4) + 3.0(x_1x_2x_3 + x_1x_2x_4 + x_2x_3x_4)]$$

$$x_1 \sim \text{Bernoilli}(0.25) \quad x_2 \sim \text{Bernoilli}(0.5) \quad x_3 \sim N(0, 1) \quad x_4 \sim N(0, 1)$$

- Design 7: Non-proportional hazard function of Design 2. $\beta_1 = 1.0$ and $\beta_2 = 0.25$, x_1 and x_2 before the time point with 70% survival probability. Therefore. $\beta_1 = \beta_2 = 0$. $n = 100$, 50 replications.

- Design 8: Same as Design 3 except $n = 200$ cases.

- Design 9: Same as Design 5 except $n = 200$ cases.

23.6.2 요약

이 페이퍼는 "Faraggi and Simon(1995)[22]의 방법", "Liestol et al.(1994)[42]의 방법", 그리고 Buckley and James 방법 (1979)[18]에 neural network를 불인 "수정된 Buckley and James 방법"을 비교하는 연구를 하였다.

1. Faraggi-Simon method Faraggi and Simon(1995)[22]의 방법은 Chapter 23.3에서 이미 정리하였다.
다시한 번 요약하자면, Faraggi and Simon(1995)는 Cox regression의 일반적인 공변량의 선형 결합 대신에, 비선형 함수를 허용하도록 일반화하였다.

2. Liestol-Andersen-Andersen method

Liestol et al.(1994)[42]의 방법은 Chapter 23.2에서 이미 이미 정리하였다.

다시한 번 요약하자면, Liestol et al.(1994)는, 생존 시간은, 위험이 일정하다는 가정 하에서, 출력 노드가 시간 간격으로 그룹화된다.

Hidden layer가 없고, 입력 노드에서 출력 노드까지 동일한 가중치 조건으로 설정된 NN은, 각 objects에 대한 conditional event probability를 training하도록 예측되며, 그 결과는 그룹화된 Cox regression과 동일하다.

각 출력 노드에 대한 가중치가 허용될 때는, 시간에 따른 일정하지 않은 위험을 모델링할 수 있다. Hidden layer를 추가하면, 비선형 covariate와 Cox-NN 하이브리드 형식의 모델이 만들어지고, linearity는 이 hidden layer의 수와 activation function에 달려 있다.

3. Modified Buckley-James method

선형 회귀 분석에 적용된 원래의 Buckley-James 방법의 경우, censored survival time은 공변량과 잔여 Kaplan-Keier로부터 추정된 회귀선에 따라 예상 값이 결정되게 된다. Residual distribution은 파라미터의 함수이기 때문에, 반복 추정을 하는 방법이 이용되며, 각 반복에서의 추정 값은, 현재의 파라미터의 추정값을 기반으로 한다.

Modified Buckley-James 방법의 경우, NN outputs은 fitted regression line 대신 사용되며, ... 작성중...

위 세 방법과 원래의 cox-regression까지 4가지 방법을 C-index를 이용하여 비교하였다.

이 논문에서의 discussion은 다음과 같이 서술하고 있다.

- Faraggi-Simon의 방법은 Design 7, Design 9에 대해서 잘 작동하는 모습을 보였다.
- Liestol et al.의 방법은 Design 7에서 잘 작동하는 모습을 보였다.
- Modified Buckley-James 방법은 Design 5, Design 9에서 잘 작동하는 모습을 보였다. 그러나 Design 4, Design 6, Design 9에서는 성능이 좋지 않은 모습을 보였다. 이 페이퍼에서는 아마도 error function을 잘못 선택했기 때문이라고 생각하고 있었다.
- 모델에 interaction이 포함되지 않았을 때는, 대체로 NN 방법을 이용하는 것이 Cox-regression보다 상대적으로 성능이 더 좋아지는 것을 볼 수 있었다. 이러한 결과는, 생존 분석을 위해 설계된 NN이, 자동으로 interaction을 수용할 수 있는 반면, Cox-regression을 사용하려면 분석가의 통찰력과 경험이 필요함을 나타낸다.
다만 modified Buckley-James 방법은 이에 해당하지 않는다고 한다. 특히 Hidden Layer 수를 늘림으로써 NN 아키텍처를 변경하니, 오히려 성능이 떨어지는 것을 관찰했다고 한다.
- Liestol et al.의 방법은 생존 시간을 얼마나 이산화했는지가, 성능에 크게 영향을 주는 것으로 보인다고 한다. 이산화를 매우 자잘하게 할 수록 성능이 좋아진다고 한다.

23.6.3 의견

생존 분석을 위해 design된 NN 모델이 interaction을 자동으로 수용할 수 있는 장점이 있음을 보여준 연구이다.

Synthetic data를 이용함으로써 discussion에 대한 이야기를 받아들일 수 있다는 것도 장점이다. Real data를 사용하지 않아서 한계가 있다고 이 페이퍼에서는 서술하고 있는데, 이는 Faraggi and Simon과 Liestol et al.의 원래 연구에서 진행되고 있으므로 커버가 된다고 생각한다.

23.7 Jared L. Katzman et al.(2016), Deep Survival: A Deep Cox Proportional Hazards Networks

Jared L. Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang and Yuval Kluger(2016), **Deep Survival: A Deep Cox Proportional Hazards Networks**, [arXiv:1606.00931v2 \[stat.ML\]](https://arxiv.org/abs/1606.00931v2) 25 Oct 2016.

23.7.1 데이터셋

여기에서는 2가지 synthetic dataset과 real dataset을

- Simulated Survival Data
 - Linear Risk Experiment(Linear)
 - Nonlinear Risk Experiment(Gaussian)
- Real Survival Data Experiments
 - Worcester Heart Attack Study(WHAS)
 - Molecular Taxonomy of Breast Cancer International Consortium(Metabric)
- Treatment Recommender System Experiments
 - Simulated Treatment Data(Treatment)
 - Hormone Treatment Recommendations for Breast Cancer(GBSG)
- 논문에는 있지만, 공개된 코드에는 있는 데이터셋
 - Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT)

23.7.2 코드

일단 CPH regression, Kaplan-Meier estimation, C-index statistics, log-rank test는 파이썬의 Lifelines 패키지를 이용하였다.

이 논문의 contribution인 DeepSurv는 Theano와 Lasagne 패키지를 이용하여 구현되었다.

이 페이퍼를 위해 별도로 구현된 코드는 다음 Github 링크를 통해 공개되어있다.

<https://github.com/jaredleekatzman/DeepSurv>

Random Survival Forest는 R의 randomForestSRC 패키지를 이용하였다.

23.7.3 요약

David Faraggi and Richard Simon(1995)[22]의 연구는 hidden layer가 없거나 하나인 모델만 사용하였다. 여기에서는 이 연구에 본격적으로 Deep Layers를 붙이기 시작하였다.

23.7.4 의견

작성중...

23.8 Travers Ching et al.(2016), Cox-nnet: an artificial neural network Cox regression for prognosis prediction

Travers Ching, Xun Zhu, and Lana Garmire(2016),
Cox-nnet: an artificial neural network Cox regression for prognosis prediction, [bioRxiv:093021](https://www.biorxiv.org/content/early/2016/09/02/093021).

23.8.1 데이터셋

이 페이지에서는 TCGA 데이터셋을 활용하였다. 출처는 다음과 같다.

Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 322 2014. Broad Institute of MIT and Harvard, doi:10.7908/C1DN43V9 (2014).

<https://gdac.broadinstitute.org/>

여기에서 RNA-Seq expression and clinical data를 활용하였다.

위 데이터셋 중 다음 10가지 데이터셋에 대해 분석을 진행하였다. 이들은 sample size가 300 이상, survival event가 50 이상인 것들로 선정한 것이라고 한다.

- [BLCA](#): Bladder Urothelia Carcinoma
- [BRCA](#): Breast invasive carcinoma
- [HNSC](#): Head and Neck squamous cell carcinoma
- [KIRC](#): Kidney renal clear cell carcinoma
- [LGG](#): Brain Lower Grade Glioma
- [LIHC](#): Liver hepatocellular carcinoma
- [LUAD](#): Lung adenocarcinoma
- [LUSC](#): Lung squamous cell carcinoma
- [OV](#): Ovarian serous cystadenocarcinoma
- [STAD](#): Stomach adenocarcinoma

또한 raw count data는 R의 DESeq2 패키지를 이용하여 normalization과 log-transformation을 진행하였다고 한다.

23.8.2 코드

이 페이지에 대한 코드는 다음 Github 링크를 통해 공개되어있다.

<https://github.com/lanagarmire/cox-nnet>

다만, Python 2.x 기반으로 작성되어 있으므로, Python 3.x 기반으로 수정할 필요가 있었다. 수정한 코드는 다음 Github 링크를 참고하도록 하자.

<https://github.com/praster1/cox-nnet>

23.8.3 요약

작성중...

23.8.4 의견

다만 이 페이지에 table이나 figure가 온전하게 첨부되어 있지는 않은 상황이 아쉽다.

23.9 Margaux Luck et al.(2017), Deep Learning for Patient-Specific Kidney Graft Survival Analysis

Margaux Luck, Tristan Sylvain, Meloise Cardinal, Andrea Lodi and Yoshua Bengio(2017),
Deep Learning for Patient-Specific Kidney Graft Survival Analysis, [arXiv:1705.10245v1 \[cs.LG\]](https://arxiv.org/abs/1705.10245v1) 29 May 2017.

23.9.1 데이터셋

작성중...

23.9.2 요약

Margaux Luck et al.[46]의 페이퍼는 앞선 Chapter 1~3의 기본개념 바탕을 그대로 하고 있다. (또한, 우종도절단만을 고려한다.) 그들의 구체적인 기여는, 3.3.3.의 부분가능도 함수를 딥러닝으로 풀이한 것으로 여겨진다.

23.9.3 의견

작성중...

24 Life Data Analysis without Neural Network

24.1 Torsten Hothorn et al.(2006), Survival Ensembles

Torsten Hothorn, Peter Buhlmann, Sandrine Dudoit, Annette Molinaro, and Mark J. van der Laan(2005), Survival Ensembles, *Biostatistics*, Vol. 7, No. 3, 355-373.

24.1.1 데이터셋

Dataset 5. T. Hothorn et al.(2006)의 Acute Myeloid Leukemia from Bullinger et al. (2004)

Bullinger et al.(2004)[19]에 의해 제공된 데이터로, 다음 URL에 접속한 후, accession number "GSE425"를 입력하여 다운로드할 수 있다.

<https://www.ncbi.nlm.nih.gov/geo/>

이 데이터셋에 대한 자세한 요약은 Chatper 25.2.20를 참고하도록 하자.

Dataset 6. Node-positive breast Cancer

German Breast Cancer Study Group에서 제공하는 유방암 데이터셋으로, 총 인스턴스 수는 686이 R의 TH.data 패키지의 GBSG2 함수로 제공이 된다.

이 데이터셋에 대한 자세한 요약은 Chatper 25.2.21를 참고하도록 하자.

24.1.2 Additional Materials

이 페이퍼에서의 실험 내용은 다음 URL의 문서를 통해 재현할 수 있다.

<https://cran.r-project.org/web/packages/mboost/vignettes/SurvivalEnsembles.pdf>

24.1.3 요약

Bagging, random forest, oosting과 같은 앙상블(Ensemble) 알고리즘 중에서, random forest, gradient boosting을 censored data의 제약을 고려하여 다시 제고한 방법이다.

1. Model

(a) Full Data World

$Y = \log(T)$, $T \in \mathbf{R}^+$ 라고 할 때, 관찰된 확률변수 $\mathbf{Z} = (Y, \mathbf{X}) \sim \mathcal{F}_{Y, \mathbf{X}}$ 이라고 하자. 이 때 $\mathbf{X} = (X_1, \dots, X_p)$ 는 sample space $\chi = \chi_1 \times \dots \times \chi_p$ 로부터 얻은 것이다.

조건부 분포 $\mathcal{F}_{Y|\mathbf{X}} = \mathcal{F}_{Y|f(\mathbf{X})}$ 는 covariates \mathbf{X} 가 주어졌을 때 Y 에 매핑하는 실함수이다. $f : \chi \rightarrow \mathbf{R}$.

이 regression function f 에서, 추정하고자 하는 파라메터의 parameter space를 Ψ 로 표시하고, loss function L 의 최소화를 다음과 같이 표현하였다.

$$E_{Y, \mathbf{X}} L(Y, f(\mathbf{X})) = \int L(Y, f(\mathbf{X})) d\mathcal{F}_{Y, \mathbf{X}} = \min_{\psi \in \Psi} \int L(Y, \psi(\mathbf{X})) d\mathcal{F}_{Y, \mathbf{X}} \quad (33)$$

(b) Observed Data World

Censoring을 고려한 환경을 이 페이퍼에서는 Observed data world라고 칭하였고, 다음과 같이 세팅하였다.

$\tilde{Y} = \log(\tilde{T})$, $\tilde{T} = \min(T, C)$, censoring indicator $\Delta = I(T \leq C)$ 이라고 할 때, 관찰된 확률변수 $\mathbf{O} = (\tilde{Y}, \Delta, \mathbf{X}) \sim \mathcal{F}_{\tilde{Y}, \Delta, \mathbf{X}}$ 라고 하자.

Conditional censoring distribution $P(C \leq c | \mathbf{Z})$ 는 covariates에 의존하므로, $P(C \leq c | \mathbf{Z}) = P(C \leq c | \mathbf{X})$ 로 다시 표시할 수 있다. 이는 곧, survival time T 와 censoring time C 가 covariates \mathbf{X} 에 대한 조건부 독립이라는 말과 동치이다. 이 가정을 "coarsening at random(CAR)" 가정이라고 부른다.

또한, conditional censoring survival function $G(c | \mathbf{X}) = P(C > c | \mathbf{X})$ 라고 할 때, $G(T | \mathbf{X})$ 는 항상 full data distribution $\mathcal{F}_{Y, \mathbf{X}}$ 에 대해, 거의 모든 부분에 있어서 0보다 크다고 가정한다.

만약 full data라면(censored data가 없다면) f 의 추정 함수 \hat{f} 는 식 (33)에서 정의한 loss function의 최소화만으로도 충분하다. 하지만, life data는 censoring을 반드시 censoring을 고려해야 한다. Laan and Robins(2003)[64]와 Laan and Dudoit(2003)[63]은 nuisance parameter η 가 주어진 loss function $L(\tilde{Y}, \psi(\mathbf{X}) | \eta)$ 을 정의하고, 식 (33)을 대신하여, 모든 후보 추정치 $\psi \in \Psi$ 하에서 다음과 같은 관계가 있다고 하였다.

$$E_{Y, \mathbf{X}} L(Y, \psi(\mathbf{X})) = \int L(Y, \psi(\mathbf{X})) d\mathcal{F}_{Y, \mathbf{X}} = \int L(\tilde{Y}, \psi(\mathbf{X} | \eta)) d\mathcal{F}_{\tilde{Y}, \Delta, \mathbf{X}} = E_{\tilde{Y}, \Delta, \mathbf{X}} L(\tilde{Y}, \psi(\mathbf{X} | \eta)) \quad (34)$$

(c) Inverse Probability of Censoring Weights

여기에서는 particular nuisance parameter η 를 결정하는 방법에 대해 이야기하고 있다.
작성중...

2. Ensemble Learning

(a) Random Forest

Algorithm 1: Random Forest for Censored Data

- 1 **Initialization** Set $m = 1$ and fix $M > 1$
- 2 **Bootstrap** Draw a random vector of case counts $\mathbf{v}_m = (v_{m1}, \dots, v_{mn})$ from the multinomial distribution with parameters n and $(\sum_{i=1}^n)^{-1} \mathbf{w}$.
- 3 and fit the base learner $\hat{\theta}_{U,X}$ s to the new 'response' U_i by weighted least squares.
- 4 **Base Learner** Construct a partition $\pi_m = (R_{m1}, \dots, R_{mK(m)})$ of the sample space χ into $K(m)$ cells via a regression tree. The tree is build using the learning sample \mathfrak{L} with case counts \mathbf{v}_m , i.e. is based on a perturbation of the learning sample \mathfrak{L} with observation i occurring v_{mi} times.
- 5 **Interaction** Increase m by one and repeat **Steps 2 and 3** until $m = M$

(b) Gradient Boosting - Full Data World

Fitting the base learner can be performed by minimizing any loss function, for example solving the least-squares problem

$$\hat{\theta}_{U,X} = \operatorname{argmin}_{\theta} \sum_{i=1}^n (U_i - h(X_i|\theta))^2$$

Algorithm 2: Generic Gradient Boosting for Uncensored Data

- 1 **Initialization** Define $U_i = Y_i$ ($i = 1 \dots, n$), set $m = 0$, and $\hat{f}_0(\cdot) = h(\cdot|\hat{\theta})_{U,X}$. Fix $M > 1$.
 - 2 **Gradient** Compute the residuals
- $$U_i = -\frac{\partial L(Y_i, \psi)}{\partial \psi} \Big|_{\psi=\hat{f}_m(X_i)}$$
- 3 **Update** Update $\hat{f}_{m+1}(\cdot) = \hat{f}_m(\cdot) + \nu h(\cdot|\hat{\theta}_{U,X})$ with step size $0 < \nu \leq 1$, for example, $\nu = 0.1$.
 - 4 **Iteration** Increase m by one and repeat **Steps 2 and 3** until $m = M$.

(c) Gradient Boosting - Observed Data World

Fitting the base learner can be performed by minimizing any loss function, for example solving the least-squares problem

$$\hat{\theta}_{\tilde{U},X} = \operatorname{argmin}_{\theta} w_i \sum_{i=1}^n (U_i - h(X_i|\theta))^2 \text{ with pseudo-responses } \tilde{U}_i = -\frac{\partial L(\tilde{Y}_i, \psi)}{\partial \psi} \Big|_{\psi=\hat{f}_m(X_i)}$$

Algorithm 3: Generic Gradient Boosting for Censored Data

- 1 **Initialization** Define $\tilde{U}_i = \tilde{Y}_i$ ($i = 1 \dots, n$), set $m = 0$, and $\hat{f}_0(\cdot) = h(\cdot|\hat{\theta})_{U,X}$. Fix $M > 1$.
 - 2 **Gradient** Compute the residuals
- $$\tilde{U}_i = -\frac{\partial L(\tilde{Y}_i, \psi)}{\partial \psi} \Big|_{\psi=\hat{f}_m(X_i)}$$
- 3 and fit the base learner $\hat{\theta}_{\tilde{U},X}$ s to the new 'response' \tilde{U}_i by weighted least squares.
 - 4 **Update** Update $\hat{f}_{m+1}(\cdot) = \hat{f}_m(\cdot) + \nu h(\cdot|\hat{\theta}_{\tilde{U},X})$ with step size $0 < \nu \leq 1$, for example, $\nu = 0.1$.
 - 5 **Iteration** Increase m by one and repeat **Steps 2 and 3** until $m = M$.

(d) Gradient Boosting - Choice of Base Learners and Stop Criterion

24.1.4 의견

작성중...

24.2 Hemant Ishwaran et al.(2008), Random Survival Forests

Hemant Ishwaran, Udaya B. Kogalua, Eugene H. Blackstone and Michael S. Lauer(2008), **Random Survival Forests**, The Annals of Applied Statistics, Vol. 2, No. 3, 841-860.

24.2.1 데이터셋

여기에서는 다음 8가지 데이터셋을 이용하였다.

- Node-positive breast cancer data studied in Hothorn et al.(2006)[36]

이 데이터는 Torsten Hothorn et al.(2006), "Survival Ensembles"에서 다른었던 데이터이다.

Torsten Hothorn et al.(2006)에 대한 자세한 내용은 Chapter 24.1를 참고하자.

R의 TH.data 패키지의 GBSG2 함수로 제공이 되며, 총 인스턴스 수는 686이지만, Hemant Ishwaran et al.은 이 중 10, 50, 100개를 임의로 선별하여 각각 breast10, breast50, breast100으로 명명하여 활용하였다.

- Veteran's administration lung cancer data from Kalbfleisch and Prentice (1980)[41] (veteran)

R의 randomSurvivalForest 패키지의 veteran 함수로 제공이 된다.

Additional Materials인 Ishwaran and Kogular(2007)[38]에서 자세한 활용 예시를 소개하였다.

- Primary biliary cirrhosis data from Fleming and Harrington(1991)[25] (pbc)

R의 randomSurvivalForest 패키지의 pbc 함수로 제공이 된다.

Additional Materials인 Ishwaran and Kogular(2007)[38]에서 자세한 활용 예시를 소개하였다.

- Burn patient data from Kalbfleisch and Prentice(1980)[41] (burn)

R의 randomSurvivalForest 패키지의 burn 함수로 제공이 된다.

그러나 randomSurvivalForest 패키지가 randomForestSRC 패키지로 바뀌면서, 이 데이터는 포함되지 않게 되었다. 따라서 아직 구하지 못하였다.

- Recidivism data from Rossi, Berk and Lenihan(1980)[55] (recid)

R의 randomSurvivalForest 패키지의 recid 함수로 제공이 된다.

그러나 randomSurvivalForest 패키지가 randomForestSRC 패키지로 바뀌면서, 이 데이터는 포함되지 않게 되었다. 따라서 아직 구하지 못하였다.

- A prostate dataset described in Kattan(2003)[40] (prostate)

이 데이터는 아직 구하지 못하였다.

- A dataset comprising patients listed for heart transplant at Cleveland Clinic in Ishwaran et al.(2004)[37] (transplant)

이 데이터는 아직 구하지 못하였다.

- Early stage esophageal cancer data considered in Ishwaran et al.(2004)[37] (esophagus)

이 데이터는 아직 구하지 못하였다.

24.2.2 Additional Materials

Ishwaran H. and Kogalur U. B.(2007), **Random survival forests for R**, Rnews, Vol. 7, No. 2, 25-31.

24.2.3 요약

24.2.4 의견

Xinliang Zhu, Jiawen Yao and Xin Luo(2016),

Lung cancer survival prediction from pathological images and genetic data - An integration study, 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI).

작성중...

Part X

Survey and Research Process

25 데이터셋 요약

Chapter IX에서 리뷰한 페이퍼들(+ α)로부터 사용된 데이터셋들을 요약 정리하였다.

• Synthetic Data

- Chapter: 25.1.1 ← 생존 시간의 생성 by Bender et al.(2005)
 - * 리뷰한 페이퍼에서 사용되지 않음.
- Chapter: 25.1.2 ← Anny Xiang et al.(2000)의 intersection을 고려한 데이터
 - * Chapter 23.6: Anny Xiang et al.(2000), Comparison of the performance of neural network methods and Cox regression for censored survival data.
- Chapter: 25.1.3 ← Linear Risk Experiment(Linear) by Jared L. Katzman et al.(2016)
 - * Chapter 23.7: Jared L. Katzman et al.(2016), Deep Survival: A Deep Cox Proportional Hazards Networks
- Chapter: 25.1.4 ← Nonlinear Risk Experiment(Gaussian) by Jared L. Katzman et al.(2016)
 - * Chapter 23.7: Jared L. Katzman et al.(2016), Deep Survival: A Deep Cox Proportional Hazards Networks

• Real Data

- Chapter: 25.2.1 ← Drzewiecki and Andersen(1982)의 피부암 데이터(Melanoma)
 - * Chapter 23.2: Knut Liestol et al.(1994), Survival analysis and neural nets.
- Chapter: 25.2.2 ← Byar and Green(1980)의 전립선암 데이터(Byar)
 - * Chapter 23.3: David Faraggi and Richard Simon(1995), A Neural Network Model for Survival Data.
- Chapter: 25.2.3 ← Head and Neck cancer trial from Efron B.(1988)
 - * Chapter 23.5: Elia Biganzoli et al.(1998), Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach
- Chapter: 25.2.4 ← Veteran's administration lung cancer trial from Kalbfleisch and Prentice (1980) (veteran)
 - * Chapter 23.5: Elia Biganzoli et al.(1998), Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach
 - * Chapter 24.2: Hemant Ishwaran et al.(2008), Random Survival Forests.
- Chapter: 25.2.5 ← Worcester Heart Attack Study(WHAS)
 - * Chapter 23.7: Jared L. Katzman et al.(2016), Deep Survival: A Deep Cox Proportional Hazards Networks
- Chapter: 25.2.6 ← Molecular Taxonomy of Breast Cancer International Consortium(Metabric)
 - * Chapter 23.7: Jared L. Katzman et al.(2016), Deep Survival: A Deep Cox Proportional Hazards Networks
- Chapter: 25.2.7 ← Simulated Treatment Data(Treatment)
 - * Chapter 23.7: Jared L. Katzman et al.(2016), Deep Survival: A Deep Cox Proportional Hazards Networks
- Chapter: 25.2.8 ← Hormone Treatment Recommendations for Breast Cancer(GBSG)
 - * Chapter 23.7: Jared L. Katzman et al.(2016), Deep Survival: A Deep Cox Proportional Hazards Networks
- Chapter: 25.2.9 ← Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT)
 - * Chapter 23.7: Jared L. Katzman et al.(2016), Deep Survival: A Deep Cox Proportional Hazards Networks
- Chapter: 25.2.10 ← BLCA: Bladder Urothelia Carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)
 - * Chapter 23.8: Travers Ching et al.(2016), Cox-nnet: an artificial neural network Cox regression for prognosis prediction.
- Chapter: 25.2.11 ← BRCA: Breast invasive carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)
 - * Chapter 23.8: Travers Ching et al.(2016), Cox-nnet: an artificial neural network Cox regression for prognosis prediction.
- Chapter: 25.2.12 ← HNSC: Head and Neck squamous cell carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)
 - * Chapter 23.8: Travers Ching et al.(2016), Cox-nnet: an artificial neural network Cox regression for prognosis prediction.
- Chapter: 25.2.13 ← KIRC: Kidney renal clear cell carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)
 - * Chapter 23.8: Travers Ching et al.(2016), Cox-nnet: an artificial neural network Cox regression for prognosis prediction.
- Chapter: 25.2.14 ← LGG: Brain Lower Grade Glioma from Broad Institute TCGA Genome Data Analysis Center (2014)
 - * Chapter 23.8: Travers Ching et al.(2016), Cox-nnet: an artificial neural network Cox regression for prognosis prediction.
- Chapter: 25.2.15 ← LIHC: Liver hepatocellular carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)
 - * Chapter 23.8: Travers Ching et al.(2016), Cox-nnet: an artificial neural network Cox regression for prognosis prediction.
- Chapter: 25.2.16 ← LUAD: Lung adenocarcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)
 - * Chapter 23.8: Travers Ching et al.(2016), Cox-nnet: an artificial neural network Cox regression for prognosis prediction.
- Chapter: 25.2.17 ← LUSC: Lung squamous cell carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)
 - * Chapter 23.8: Travers Ching et al.(2016), Cox-nnet: an artificial neural network Cox regression for prognosis prediction.
- Chapter: 25.2.18 ← OV: Ovarian serous cystadenocarcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)
 - * Chapter 23.8: Travers Ching et al.(2016), Cox-nnet: an artificial neural network Cox regression for prognosis prediction.
- Chapter: 25.2.19 ← STAD: Stomach adenocarcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)
 - * Chapter 23.8: Travers Ching et al.(2016), Cox-nnet: an artificial neural network Cox regression for prognosis prediction.
- Chapter: 25.2.20 ← Acute Myeloid Leukemia from Bullinger et al. (2004)
 - * Chapter 24.1: Torsten Hothorn et al.(2006), Survival Ensembles.
- Chapter: 25.2.21 ← Node-positive breast Cancer from German Breast Cancer Study Group (GBCS2)
 - * Chapter 24.1: Torsten Hothorn et al.(2006), Survival Ensembles.
 - * Chapter 24.2: Hemant Ishwaran et al.(2008), Random Survival Forests.
- Chapter: 25.2.22 ← Primary biliary cirrhosis data from Fleming and Harrington(1991) (pbc)
 - * Chapter 24.2: Hemant Ishwaran et al.(2008), Random Survival Forests.

25.1 Synthetic Data

25.1.1 생존 시간의 생성 by Bender et al.(2005)

Bender et al.(2005)[14]는 지수 분포를 따르는 Cox 비례위험모형에서, 생존 시간을 생성하기 위해 $S_0(t) = e^{-\lambda t}$ 를 가정하였다. i 번째 대상의 생존 시간은 다음과 같이 표현된다.

$$T_i = -\frac{\log(U_i)}{\lambda e^{\beta' x}}$$

생존 시간을 생성할 때, 독립변수 x 는 평균이 0이고 공분산이 0.1인 다변량 정규분포를 가정하였다. 독립변수의 개수는 2개이며, 2번째 독립변수를 새로운 예측모형에 포함된 인자로 가정하였다.

여기서 $U_i \sim Unif(0, 1)$ 를 바탕으로 생성된 임의의 실수이며, $\lambda = 0.5$ 를 사용하여 생존 시간을 생성하였다. 회귀계수 β 는 다음과 같이 총 3가지 상황을 가정하였다.

- $\beta = (0, 1)$

기존의 독립변수는 생존 시간에 영향을 미치며, 새로 추가된 인자는 생존 시간에 전혀 무관한 독립변수임을 가정하였다.

- $\beta = (1, 0.3)$

새로 추가된 인자의 영향력이 다소 향상된 경우이다. 이 경우 Cox 비례위험모형에서의 회귀계수에 대한 유의확률은, 유의수준 0.05보다 크게 나타난다.

- $\beta = (1, 0.5)$

새로 추가된 인자의 영향력을 보다 향상시킨 경우이다. 이 경우 Cox 비례위험모형에서의 새로 추가된 변수에 대한 회귀 계수의 유의확률은, 유의수준 0.05보다 굉장히 작게 나타난다.

25.1.2 Anny Xiang et al.(2000)의 intersection을 고려한 데이터

이 데이터셋은 Anny Xiang et al.(2000), "Comparison of the performance of neural network methods and Cox regression for censored survival data"에서 정의 및 활용하였다. 이 페이퍼는 Chapter 23.6를 참고하자.

- Design 1: No censoring. $n = 100$, 50 replications.

$$\gamma = \exp[x_1 + 0.25x_2]$$

$$x_1 \sim \text{Bernoilli}(0.5) \quad x_2 \sim N(0, 1)$$

- Design 2: Average censoring for training and testing sets = 19% each. $n = 100$, 50 replications.

$$\gamma = \exp[x_1 + 0.25x_2]$$

$$x_1 \sim \text{Bernoilli}(0.5) \quad x_2 \sim N(0, 1)$$

- Design 3: Average censoring for training and testing sets = 20% each. $n = 100$, 50 replications.

$$\gamma = \exp[x_1 + 0.25x_2 + 0.2x_1x_2]$$

$$x_1 \sim \text{Bernoilli}(0.5) \quad x_2 \sim N(0, 1)$$

- Design 4: Average censoring for training and testing sets = 32%. $n = 100$, 50 replications.

$$\begin{aligned} \gamma = & \exp[2(x_1 + x_2) + 0.5x_3 + 1.0x_4 \\ & + 1.0(x_1x_2 + x_1x_3) \\ & + 0.5(x_1x_4 + x_2x_3 + x_2x_4 + x_3x_4) \\ & + 0.5(x_1x_2x_3 + x_1x_2x_4 + x_1x_3x_4 + x_2x_3x_4 + x_1x_2x_3x_4)] \end{aligned}$$

$$x_1 \sim \text{Bernoilli}(0.25) \quad x_2 \sim \text{Bernoilli}(0.5) \quad x_3 \sim N(0, 1) \quad x_4 \sim N(0, 1)$$

- Design 5: Same as Design 4. Average censoring for training and testing sets = 70%. $n = 100$, 50 replications.

- Design 6: Average censoring for training and testing sets = 47%. $n = 100$, 50 replications.

$$\gamma = \exp[0.5(x_1 + x_2) + 0.25(x_3 + x_4) + 3.0(x_1x_2x_3 + x_1x_2x_4 + x_2x_3x_4)]$$

$$x_1 \sim \text{Bernoilli}(0.25) \quad x_2 \sim \text{Bernoilli}(0.5) \quad x_3 \sim N(0, 1) \quad x_4 \sim N(0, 1)$$

- Design 7: Non-proportional hazard function of Design 2. $\beta_1 = 1.0$ and $\beta_2 = 0.25$, x_1 and x_2 before the time point with 70% survival probability. Therefore. $\beta_1 = \beta_2 = 0$. $n = 100$, 50 replications.

- Design 8: Same as Design 3 except $n = 200$ cases.

- Design 9: Same as Design 5 except $n = 200$ cases.

25.1.3 Linear Risk Experiment(Linear) by Jared L. Katzman et al.(2016)

25.1.4 Nonlinear Risk Experiment(Gaussian) by Jared L. Katzman et al.(2016)

25.2 Real Data

25.2.1 Drzewiecki and Andersen(1982)의 피부암 데이터(Melanoma)

Drzewiecki and Andersen(1982)[23]의 연구를 통해 소개된 피부암 데이터이다. 205명의 환자 데이터이며, R의 'timereg' 패키지에 내장되어 있다. 다음과 같이 불러올 수 있다.

Code 17. Drzewiecki and Andersen(1982)의 피부암 데이터 불러오기(R 코드)

```
1 library("timereg")
2 data("melanoma")
```

25.2.1.1 변수 요약

Instance 수는 205명이며, 총 6개의 변수가 있다. 다만 `no` 변수는 환자를 구분하기 위한 용도 외에는 큰 의미가 없다.

- `no`, a numeric vector. Patient code.
- `status`, a numeric vector code. Survival status.
 - 1: dead from melanoma
 - 2: alive
 - 3: dead from other cause
- `days`, a numeric vector. Survival time.
- `ulc`, a numeric vector code. Ulceration.
 - 1: present
 - 0: absent
- `thick`, a numeric vector. Tumour thickness (1/100 mm).
- `sex`, a numeric vector code.
 - 0: female
 - 1: male

	no	status	days	ulc	thick	sex
All	Not meaningful	1: 57 2: 134 3: 14	Min. 10.000 1st Qu. 1525.000 Median 2005.000 Mean 2153.000 3rd Qu. 3042.000 Max. 5565.000 Std.Dev. 1122.061	0: 115 1: 90	Min. 10.000 1st Qu. 97.000 Median 194.000 Mean 292.000 3rd Qu. 356.000 Max. 1742.000 Std.Dev. 295.943	0: 126 1: 79
Status=1	Not meaningful	1: 57 2: 0 3: 0	Min. 185.000 1st Qu. 718.000 Median 1062.000 Mean 1253.000 3rd Qu. 1667.000 Max. 3338.000 Std.Dev. 758.998	0: 16 1: 41	Min. 32.000 1st Qu. 224.000 Median 354.000 Mean 431.100 3rd Qu. 484.000 Max. 1742.000 Std.Dev. 357.381	0: 28 1: 29
Status=2	Not meaningful	1: 0 2: 134 3: 0	Min. 35.000 1st Qu. 1858.000 Median 2374.000 Mean 2621.000 3rd Qu. 3330.000 Max. 5565.000 Std.Dev. 948.138	0: 92 1: 42	Min. 10.000 1st Qu. 81.000 Median 135.500 Mean 224.500 3rd Qu. 286.000 Max. 1288.000 Std.Dev. 232.617	0: 91 1: 43
Status=3	Not meaningful	1: 0 2: 0 3: 14	Min. 10.000 1st Qu. 262.800 Median 1126.500 Mean 1338.300 3rd Qu. 2028.800 Max. 3458.000 Std.Dev. 1247.805	0: 7 1: 7	Min. 16.000 1st Qu. 129.000 Median 226.000 Mean 371.800 3rd Qu. 580.000 Max. 1256.000 Std.Dev. 363.162	0: 7 1: 7
ulc=0	Not meaningful	1: 16 2: 92 3: 7	Min. 30.000 1st Qu. 1682.000 Median 2103.000 Mean 2415.000 3rd Qu. 3181.000 Max. 5565.000 Std.Dev. 1034.016	0: 115 1: 0	Min. 10.000 1st Qu. 65.000 Median 129.000 Mean 181.100 3rd Qu. 194.000 Max. 1466.000 Std.Dev. 218.586	0: 79 1: 36
ulc=1	Not meaningful	1: 41 2: 42 3: 7	Min. 10.000 1st Qu. 827.800 Median 1799.500 Mean 1817.800 3rd Qu. 2486.500 Max. 4492.000 Std.Dev. 1146.313	0: 0 1: 90	Min. 16.000 1st Qu. 224.500 Median 354.000 Mean 433.600 3rd Qu. 516.000 Max. 1742.000 Std.Dev. 321.5295	0: 47 1: 43
sex=0	Not meaningful	1: 28 2: 91 3: 7	Min. 99.000 1st Qu. 1636.000 Median 2059.000 Mean 2283.000 3rd Qu. 3131.000 Max. 5565.000 Std.Dev. 1089.818	0: 79 1: 47	Min. 10.000 1st Qu. 97.000 Median 162.000 Mean 248.600 3rd Qu. 306.000 Max. 1742.000 Std.Dev. 275.462	0: 126 1: 0
sex=1	Not meaningful	1: 29 2: 43 3: 7	Min. 10.000 1st Qu. 1052.000 Median 1860.000 Mean 1946.000 3rd Qu. 2784.000 Max. 4492.000 Std.Dev. 1148.382	0: 36 1: 43	Min. 16.000 1st Qu. 105.000 Median 258.000 Mean 361.100 3rd Qu. 484.000 Max. 1466.000 Std.Dev. 315.572	0: 0 1: 79

Table 77: Drzewiecki and Andersen(1982)의 피부암 데이터 요약

25.2.2 Byar and Green(1980)의 전립선암 데이터(Byar)

Byar and Green(1980)[20]의 연구를 통해 소개된 전립선암 데이터이다. 506명의 환자 데이터이며, R의 'clustMD' 패키지에 내장되어 있으나, clustMD 패키지에서는 475명의 환자 데이터를 제공하여 준다.⁴² 다음과 같이 불러올 수 있다.

Code 18. Byar and Green(1980)의 전립선암 데이터 불러오기(R 코드)

```
1 library("clustMD")
2 data("Byar")
```

25.2.2.1 변수 요약

Instance 수는 475명이며, 총 15개의 변수가 있다. 다만 **Observation** 변수는 환자를 구분하기 위한 용도 외에는 큰 의미가 없다.

- **Age**, a numeric vector indicating the age of the patient.
- **Weight**, a numeric vector indicating the weight of the patient.
- **Performance.rating**, an ordinal variable indicating how active the patient is
 - 0: normal activity
 - 1: in bed less than 50% of daytime
 - 2: in bed more than 50% of daytime
 - 3: confined to bed.
- **Cardiovascular.disease.history**, a binary variable indicating if the patient has a history of cardiovascular disease
 - 0: no
 - 1: yes
- **Systolic.Blood.pressure**, a numeric vector indicating the systolic blood pressure of the patient in units of ten.
- **Diastolic.blood.pressure**, a numeric vector indicating the diastolic blood pressure of the patient in units of ten.
- **Electrocardiogram.code**, a nominal variable indicating the electorcardiogram code
 - 0: normal
 - 1: benign
 - 2: rythmic disturbances and electrolyte changes
 - 3: heart blocks or conduction defects
 - 4: heart strain
 - 5: old myocardial infarct
 - 6: recent myocardial infarct
- **Serum.haemoglobin**, a numeric vector indicating the serum haemoglobin levels of the patient measured in g/100ml.
- **Size.of.primary.tumour**, a numeric vector indicating the estimated size of the patient's primary tumour in centimeters squared.
- **Index.of.tumour.stage.and.histologic.grade**, a numeric vector indicating the combined index of tumour stage and histologic grade of the patient.
- **Serum.prostatic.acid.phosphatase**, a numeric vector indicating the serum prostatic acid phosphatase levels of the patient in King-Armstrong units.
- **Bone.metastases**, a binary vector indicating the presence of bone metastasis:
 - 0: no
 - 1: yes
- **Stage**, the stage of the patient's prostate cancer.
- **Observation**, a patient ID number.
- **SurvStat**, the post trial survival status of the patient:
 - 0: alive
 - 1: dead from prostatic cancer
 - 2: dead from heart or vascular disease
 - 3: dead from cerebrovascular accident
 - 4: dead form pulmonary embolus
 - 5: dead from other cancer
 - 6: dead from respiratory disease
 - 7: dead from other specific non-cancer cause
 - 8: dead from other unspecified non-cancer cause
 - 9: dead from unknown cause

⁴²이 데이터셋은 R뿐만 아니라 다른 링크에서도 구할 수 있었지만, 506명의 환자 데이터를 제공하는 곳을 찾지는 못했다. clustMD 패키지와 같은 데이터를 다음 링크의 07번 항목에서도 제공하고 있다.

<http://portal.uni-freiburg.de/imbi/Royston-Sauerbrei-book>

	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
All	Min. 48.000 1st Qu. 70.000 Median 73.000 Mean 71.560 3rd Qu. 76.000 Max. 89.000 Std.Dev. 6.920	Min. 69.000 1st Qu. 90.000 Median 98.000 Mean 99.010 3rd Qu. 107.000 Max. 152.000 Std.Dev. 13.341	0: 428 1: 32 2: 13 3: 2	0: 268 1: 207	8: 1 17: 33 9: 3 18: 16 10: 13 19: 12 11: 26 20: 2 12: 60 21: 2 13: 70 22: 3 14: 90 23: 1 15: 70 24: 1 16: 71 30: 1	4: 4 10: 62 5: 5 11: 9 6: 40 12: 5 7: 99 13: 1 8: 155 14: 1 9: 93 18: 1	0: 161 1: 23 2: 50 3: 25 4: 145 5: 70 6: 1	Min. 59.000 1st Qu. 122.500 Median 137.000 Mean 134.200 3rd Qu. 147.000 Max. 182.000 Std.Dev. 19.382
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	Surv.Stat	
Performance.rating ==0	Min. 0.000 1st Qu. 5.000 Median 10.000 Mean 14.290 3rd Qu. 21.000 Max. 69.000 Std.Dev. 12.236	5: 3 11: 110 6: 8 12: 26 7: 6 13: 70 8: 66 14: 5 9: 132 15: 16 10: 33	Min. 1.000 1st Qu. 5.000 Median 7.000 Mean 125.700 3rd Qu. 29.500 Max. 9999.000 Std.Dev. 638.486	0: 398 1: 77	3: 273 4: 202	Not Meaningful	0: 137 5: 24 1: 121 6: 16 2: 93 7: 27 3: 31 8: 6 4: 14 9: 6	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Performance.rating ==1	Min. 49.000 1st Qu. 70.000 Median 73.000 Mean 71.460 3rd Qu. 76.000 Max. 89.000 Std.Dev. 6.802	Min. 69.000 1st Qu. 90.000 Median 99.000 Mean 99.480 3rd Qu. 108.000 Max. 152.000 Std.Dev. 13.094	0: 428 1: 0 2: 0 3: 0	0: 249 1: 179	8: 1 17: 30 9: 3 18: 14 10: 13 19: 12 11: 24 20: 2 12: 56 21: 1 13: 64 22: 1 14: 75 23: 1 15: 66 24: 1 16: 65 30: 1	4: 4 10: 59 5: 3 11: 8 6: 38 12: 2 7: 88 13: 1 8: 136 14: 1 9: 87 18: 1	0: 151 1: 20 2: 46 3: 23 4: 128 5: 59 6: 1	Min. 59.000 1st Qu. 123.000 Median 137.000 Mean 135.300 3rd Qu. 147.000 Max. 182.000 Std.Dev. 18.964
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	Surv.Stat	
Performance.rating ==1	Min. 0.000 1st Qu. 5.000 Median 10.000 Mean 13.830 3rd Qu. 20.000 Max. 69.000 Std.Dev. 11.838	5: 2 11: 97 6: 8 12: 23 7: 6 13: 59 8: 62 14: 3 9: 125 15: 13 10: 30	Min. 1.000 1st Qu. 5.000 Median 7.000 Mean 89.100 3rd Qu. 24.250 Max. 5960.000 Std.Dev. 418.059	0: 371 1: 57	3: 255 4: 173	Not Meaningful	0: 133 5: 24 1: 101 6: 16 2: 84 7: 23 3: 25 8: 4 4: 12 9: 6	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Performance.rating ==2	Min. 48.000 1st Qu. 69.750 Median 73.000 Mean 72.190 3rd Qu. 78.000 Max. 87.000 Std.Dev. 8.731	Min. 71.000 1st Qu. 87.750 Median 96.000 Mean 95.340 3rd Qu. 99.250 Max. 36.000 Std.Dev. 14.098	0: 0 1: 32 2: 0 3: 0	0: 10 1: 22	8: 17: 3 9: 18: 1 10: 19: 1 11: 2 20: 1 12: 2 21: 1 13: 3 22: 2 14: 12 23: 1 15: 2 24: 1 16: 3 30: 1	4: 4 10: 3 5: 1 11: 1 6: 2 12: 3 7: 7 13: 1 8: 13 14: 1 9: 2 18: 1	0: 7 1: 2 2: 3 3: 1 4: 12 5: 7 6:	Min. 91.000 1st Qu. 117.000 Median 128.500 Mean 127.400 3rd Qu. 140.000 Max. 176.000 Std.Dev. 17.816
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	Surv.Stat	
Performance.rating ==2	Min. 1.000 1st Qu. 4.000 Median 15.000 Mean 18.000 3rd Qu. 25.000 Max. 61.000 Std.Dev. 15.149	5: 1 11: 8 6: 12: 2 7: 13: 5 8: 3 14: 2 9: 6 15: 3 10: 3	Min. 2.000 1st Qu. 6.000 Median 9.000 Mean 80.090 3rd Qu. 23.500 Max. 1278.000 Std.Dev. 237.014	0: 23 1: 9	3: 15 4: 17	Not Meaningful	0: 2 5: 1: 9 6: 2: 8 7: 3 3: 6 8: 2 4: 2 9:	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Performance.rating ==2	Min. 60.000 1st Qu. 72.000 Median 74.000 Mean 73.920 3rd Qu. 78.000 Max. 80.000 Std.Dev. 5.235	Min. 73.000 1st Qu. 79.000 Median 93.000 Mean 94.777 3rd Qu. 105.000 Max. 134.000 Std.Dev. 17.331	0: 0 1: 0 2: 13 3: 0	0: 7 1: 6 2: 21: 1 3: 2 22: 1 4: 2 23: 1 5: 2 24: 1 6: 3 30: 1	8: 17: 9: 18: 2 10: 19: 11: 20: 12: 2 21: 13: 2 22: 14: 2 23: 15: 2 24: 16: 3 30:	4: 4 10: 5: 1 11: 6: 2 12: 7: 7 13: 8: 13 14: 9: 4 18:	0: 2 5: 1: 1 1: 2: 1 2: 3: 3 4: 4: 5 4: 5: 4 6:	Min. 72.000 1st Qu. 101.000 Median 123.000 Mean 116.500 3rd Qu. 135.000 Max. 148.000 Std.Dev. 25.877
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	Surv.Stat	
Performance.rating ==3	Min. 1.000 1st Qu. 11.000 Median 17.000 Mean 20.920 3rd Qu. 26.000 Max. 54.000 Std.Dev. 15.141	5: 1 11: 4 6: 12: 1 7: 13: 5 8: 1 14: 9: 1 15: 10: 1	Min. 3.000 1st Qu. 12.000 Median 204.000 Mean 691.500 3rd Qu. 430.000 Max. 3160.000 Std.Dev. 1101.967	0: 4 1: 9	3: 3 4: 10	Not Meaningful	0: 2 5: 1: 9 6: 2: 1 7: 1 3: 3 8: 4: 9	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Performance.rating ==3	Min. 59.000 1st Qu. 62.750 Median 66.500 Mean 66.500 3rd Qu. 70.250 Max. 74.000 Std.Dev. 10.607	Min. 74.000 1st Qu. 80.250 Median 86.500 Mean 86.500 3rd Qu. 92.750 Max. 99.000 Std.Dev. 17.678	0: 0 1: 0 2: 0 3: 2	0: 2 1: 0	8: 17: 9: 18: 10: 19: 11: 20: 12: 21: 13: 1 22: 14: 1 23: 15: 2 24: 16: 3 30:	4: 4 10: 5: 1 11: 6: 2 12: 7: 7 13: 8: 14 14: 9: 18:	0: 1 1: 1 2: 2 3: 3: 3 4: 4: 6:	Min. 112.000 1st Qu. 115.800 Median 119.500 Mean 119.500 3rd Qu. 123.200 Max. 127.000 Std.Dev. 10.607
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	Surv.Stat	
Performance.rating ==3	Min. 2.000 1st Qu. 5.750 Median 9.500 Mean 9.500 3rd Qu. 13.250 Max. 17.000 Std.Dev. 10.607	5: 1 11: 1 6: 12: 1 7: 13: 1 8: 14: 9: 15: 10: 1	Min. 37.000 1st Qu. 2528.000 Median 5018.000 Mean 5018.000 3rd Qu. 7508.000 Max. 9999.000 Std.Dev. 7044.198	0: 1: 2	3: 4: 2	Not Meaningful	0: 2 5: 1: 2 6: 2: 7 7: 3: 8 8: 4: 9 9:	

Table 78: Byar and Green(1980)의 전립선암 데이터 요약

	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Cardiovascular.disease.history == 0	Min. 49.000 1st Qu. 68.000 Median 72.000 Mean 70.520 3rd Qu. 75.000 Max. 87.000 Std.Dev. 7.201	Min. 69.000 1st Qu. 90.000 Median 98.000 Mean 98.320 3rd Qu. 106.250 Max. 145.000 Std.Dev. 13.178	0: 249 1: 10 2: 7 3: 2	0: 268 1: 1	8: 1 17: 17 9: 2 18: 6 10: 4 19: 10 11: 16 20: 1 12: 43 21: 1 13: 39 22: 1 14: 53 23: 1 15: 44 24: 1 16: 34 30: 1	4: 2 10: 35 5: 2 11: 5 6: 24 12: 2 7: 58 13: 1 8: 83 14: 1 9: 57 18: 1	0: 113 1: 15 2: 28 3: 15 4: 72 5: 25 6:	Min. 59.000 1st Qu. 122.800 Median 137.000 Mean 133.900 3rd Qu. 147.000 Max. 175.000 Std.Dev. 19.799
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	SurvStat	
Cardiovascular.disease.history == 1	Min. 5.000 1st Qu. 6.000 Median 7.000 Mean 15.260 3rd Qu. 23.000 Max. 69.000 Std.Dev. 12.692	5: 1 11: 66 6: 4 12: 16 7: 3 13: 48 8: 29 14: 2 9: 65 15: 11 10: 23	Min. 1.000 1st Qu. 5.000 Median 8.000 Mean 144.950 3rd Qu. 38.250 Max. 9999.000 Std.Dev. 784.587	0: 218 1: 50	3: 143 4: 125	Not Meaningful	0: 99 5: 14 1: 79 6: 12 2: 29 7: 14 3: 9 8: 4 4: 7 9: 1	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Cardiovascular.disease.history == 1	Min. 48.000 1st Qu. 71.000 Median 74.000 Mean 72.900 3rd Qu. 76.000 Max. 89.000 Std.Dev. 6.306	Min. 71.000 1st Qu. 90.500 Median 98.000 Mean 99.920 3rd Qu. 107.500 Max. 152.000 Std.Dev. 13.528	0: 179 1: 22 2: 6 3:	0: 207 1: 1	8: 17 16 9: 1 18: 10 10: 9 19: 5 11: 10 20: 1 12: 17 21: 1 13: 31 22: 3 14: 37 23: 1 15: 26 24: 1 16: 37 30: 1	4: 2 10: 27 5: 3 11: 4 6: 16 12: 3 7: 41 13: 1 8: 72 14: 1 9: 36 18: 1	0: 48 1: 8 2: 22 3: 10 4: 73 5: 45 6: 1	Min. 82.000 1st Qu. 122.500 Median 136.000 Mean 134.600 3rd Qu. 147.000 Max. 182.000 Std.Dev. 18.869
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	SurvStat	
Cardiovascular.disease.history == 1	Min. 0.000 1st Qu. 5.000 Median 9.000 Mean 13.020 3rd Qu. 18.000 Max. 62.000 Std.Dev. 11.525	5: 2 11: 44 6: 4 12: 10 7: 3 13: 22 8: 37 14: 3 9: 67 15: 5 10: 10	Min. 1.000 1st Qu. 4.000 Median 7.000 Mean 100.800 3rd Qu. 18.500 Max. 353.5000 Std.Dev. 372.909	0: 180 1: 27	3: 130 4: 77	Not Meaningful	0: 38 5: 10 1: 42 6: 4 2: 64 7: 13 3: 22 8: 2 4: 7 9: 5	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Bone.metastases == 0	Min. 48.000 1st Qu. 70.000 Median 73.000 Mean 71.777 3rd Qu. 76.000 Max. 89.000 Std.Dev. 6.656	Min. 72.000 1st Qu. 91.000 Median 99.000 Mean 100.100 3rd Qu. 108.000 Max. 152.000 Std.Dev. 13.203	0: 371 1: 23 2: 4 3:	0: 218 1: 180	8: 17 27 9: 3 18: 13 10: 10 19: 11 11: 23 20: 1 12: 47 21: 2 13: 62 22: 2 14: 68 23: 1 15: 63 24: 1 16: 63 30: 1	4: 3 10: 51 5: 3 11: 9 6: 27 12: 4 7: 86 13: 1 8: 133 14: 1 9: 79 18: 1	0: 142 1: 17 2: 37 3: 22 4: 118 5: 61 6: 1	Min. 59.000 1st Qu. 125.200 Median 138.000 Mean 136.800 3rd Qu. 148.000 Max. 182.000 Std.Dev. 17.858
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	SurvStat	
Bone.metastases == 0	Min. 0.000 1st Qu. 4.000 Median 9.000 Mean 12.920 3rd Qu. 18.750 Max. 69.000 Std.Dev. 11.455	5: 2 11: 81 6: 8 12: 20 7: 6 13: 44 8: 65 14: 3 9: 132 15: 8 10: 29	Min. 1.000 1st Qu. 5.000 Median 7.000 Mean 39.850 3rd Qu. 14.500 Max. 353.5000 Std.Dev. 199.774	0: 398 1:	3: 272 4: 126	Not Meaningful	0: 127 5: 23 1: 77 6: 16 2: 81 7: 24 3: 26 8: 5 4: 13 9: 6	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Bone.metastases == 1	Min. 49.000 1st Qu. 68.000 Median 72.000 Mean 70.430 3rd Qu. 76.000 Max. 87.000 Std.Dev. 8.105	Min. 69.000 1st Qu. 84.000 Median 93.000 Mean 93.480 3rd Qu. 102.000 Max. 123.000 Std.Dev. 12.746	0: 57 1: 9 2: 9 3:	0: 50 1: 27	8: 1 17: 6 9: 2 18: 3 10: 3 19: 1 11: 3 20: 1 12: 13 21: 2 13: 8 22: 1 14: 22 23: 1 15: 7 24: 1 16: 8 30: 1	4: 1 10: 11 5: 2 11: 1 6: 13 12: 1 7: 13 13: 1 8: 22 14: 1 9: 14 18: 1	0: 19 1: 6 2: 13 3: 3 4: 6 1	Min. 70.000 1st Qu. 105.000 Median 123.000 Mean 121.000 3rd Qu. 137.000 Max. 160.000 Std.Dev. 21.579
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	SurvStat	
Bone.metastases == 1	Min. 2.000 1st Qu. 10.000 Median 19.000 Mean 21.340 3rd Qu. 30.000 Max. 62.000 Std.Dev. 13.716	5: 1 11: 29 6: 8 12: 6 7: 13: 26 8: 1 14: 2 9: 15: 8 10: 4	Min. 4.000 1st Qu. 20.000 Median 93.000 Mean 569.600 3rd Qu. 385.000 Max. 9999.000 Std.Dev. 1447.698	0: 398 1: 77	3: 1 4: 76	Not Meaningful	0: 10 5: 1 1: 44 6: 1 2: 12 7: 3 3: 5 8: 1 4: 1 9:	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Stage == 3	Min. 49.000 1st Qu. 70.000 Median 73.000 Mean 71.900 3rd Qu. 76.000 Max. 89.000 Std.Dev. 6.490	Min. 72.000 1st Qu. 91.000 Median 99.000 Mean 100.200 3rd Qu. 109.000 Max. 152.000 Std.Dev. 13.094	0: 255 1: 15 2: 3 3:	0: 143 1: 130	8: 1 17: 19 9: 2 18: 9 10: 6 19: 9 11: 31 21: 1 12: 51 22: 2 13: 45 23: 1 14: 38 24: 1 15: 40 30: 1	4: 3 10: 38 5: 2 11: 6 6: 17 12: 3 7: 63 13: 1 8: 90 14: 1 9: 48 18: 1	0: 95 1: 14 2: 21 3: 16 4: 81 5: 45 6: 1	Min. 59.000 1st Qu. 125.000 Median 138.000 Mean 137.200 3rd Qu. 149.000 Max. 182.000 Std.Dev. 18.519
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	SurvStat	
Stage == 3	Min. 0.000 1st Qu. 4.000 Median 8.000 Mean 11.420 3rd Qu. 16.000 Max. 69.000 Std.Dev. 10.294	5: 2 11: 34 6: 8 12: 6 7: 6 13: 6 8: 65 14: 5 9: 130 15: 16 10: 16	Min. 1.000 1st Qu. 4.000 Median 5.000 Mean 6.689 3rd Qu. 7.000 Max. 297.000 Std.Dev. 17.974	0: 272 1:	3: 273 4:	Not Meaningful	0: 91 5: 19 1: 31 6: 12 2: 63 7: 18 3: 21 8: 2 4: 10 9: 6	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Stage == 4	Min. 48.000 1st Qu. 70.000 Median 73.000 Mean 71.900 3rd Qu. 76.000 Max. 87.000 Std.Dev. 7.454	Min. 69.000 1st Qu. 89.000 Median 97.000 Mean 97.480 3rd Qu. 105.000 Max. 150.000 Std.Dev. 13.548	0: 173 1: 17 2: 10 3:	0: 125 1: 77	8: 1 17: 14 9: 1 18: 7 10: 7 19: 3 11: 10 20: 1 12: 29 21: 1 13: 19 22: 1 14: 45 23: 1 15: 32 24: 1 16: 31 30: 1	4: 1 10: 24 5: 3 11: 3 6: 23 22: 2 7: 36 13: 1 8: 65 14: 1 9: 45 18: 1	0: 66 1: 9 2: 29 3: 9 4: 64 5: 25 6:	Min. 70.000 1st Qu. 118.000 Median 134.000 Mean 130.100 3rd Qu. 144.000 Max. 168.000 Std.Dev. 19.820
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	SurvStat	
Stage == 4	Min. 0.000 1st Qu. 6.250 Median 16.000 Mean 18.160 3rd Qu. 26.000 Max. 62.000 Std.Dev. 13.544	5: 1 11: 76 6: 12: 20 7: 13: 64 8: 1 14: 5 9: 2 15: 16 10: 17	Min. 2.000 1st Qu. 16.000 Median 39.500 Mean 286.600 3rd Qu. 200.000 Max. 9999.000 Std.Dev. 956.902	0: 126 1: 76	3: 4: 202	Not Meaningful	0: 46 5: 5 1: 90 6: 4 2: 30 7: 9 3: 10 8: 4 4: 4 9:	

Table 79: Byar and Green(1980)의 전립선암 데이터 요약(cont'd)

25.2.3 Head and Neck cancer trial from Efron B.(1988)

이 데이터는 Efron B.(1988)[24]에서 처음 소개가 된 데이터로, 페이퍼를 통해 데이터를 그대로 유추해낼 수 있다.
설명 변수가 없기 때문에, Cox-regression과 같은 모델에는 적합하지 않다.

Dataset 7. Raw data for Arm A (Table 1) from Efron (1988)

7	34	42	63	64	74+	83	84	91	108
112	129	133	133	139	140	140	146	149	154
157	160	160	165	173	176	185+	218	225	241
248	273	277	279+	297	319+	405	417	420	440
523	523+	583	594	1101	1116+	1146	226+	1349+	1412+
1417									

Table 80: Raw data for Arm A (Table 1) from Efron (1988)

Code 19. Raw data for Arm A (Table 1) from Efron (1988) 불러오기(R 코드)

```

1 # Raw data for Arm A (Table 1)
2 Adays <- c( 7, 34, 42, 63, 64, 74, 83, 84, 91, 108,
3   112, 129, 133, 133, 139, 140, 140, 146, 149, 154,
4   157, 160, 160, 165, 173, 176, 185, 218, 225, 241,
5   248, 273, 277, 279, 297, 319, 405, 417, 420, 440,
6   523, 523, 583, 594, 1101, 1116, 1146, 1226, 1349, 1412,
7   1417)
8
9 Astatus <- rep(1,51)
10 Astatus[c(6,27,34,36,42,46,48,49,50)] <-0
11 Aobj <- Surv(time = Adays, Astatus==1)
12 Aobj

```

Dataset 8. Raw data for Arm A (Table 1) from Efron (1988)

37	84	92	94	110	112	119	127	130	133
140	146	155	159	169+	173	179	194	195	209
249	281	319	339	432	469	519	528+	547+	613+
633	725	759+	817	1092+	1245+	1331+	1557	1642+	1771+
1776	1897+	2023+	2146+	2297+					

Table 81: Raw data for Arm B (Table 2) from Efron (1988)

Code 20. Raw data for Arm B (Table 2) from Efron (1988) 불러오기(R 코드)

```

1 # Raw data for Arm B (Table 2)
2 Bdys <- c(37, 84, 92, 94, 110, 112, 119, 127, 130, 133, 140, 146, 155, 159,
3   169, 173, 179, 194, 195, 209, 249, 281, 319, 339,
4   432, 469, 519, 528, 547, 613, 633, 725, 759, 817,
5   1092, 1245, 1331, 1557, 1642, 1771,
6   1776, 1897, 2023, 2146, 2297)
7
8 Bstatus <- rep(1,45)
9 Bstatus[c(15,28,29,30,33,35,36,37,39,40,42,43,44,45)] <- 0
10 Bobj <- Surv(Bdys, Bstatus == 1)
11 Bobj

```

25.2.4 Veteran's administration lung cancer data from Kalbfleisch and Prentice (1980)

Kalbfleisch J. and Prentice R(1980)[41]에서 소개된 데이터이다.

R의 randomSurvivalForest⁴³ 패키지의 veteran 함수, 또는 survival 패키지에 내장되어 있다. 둘 다 같은 데이터이니 어느 것을 써도 무방하다.

Code 21. Veteran's administration lung cancer data from Kalbfleisch and Prentice (1980) 데이터 불러오기(R 코드)

```
1 require(randomForestSRC)
2 data("veteran")
```

25.2.4.1 변수 요약

Instance 수는 137명이며, 총 8개의 변수가 있다.

- **trt**, Treatment
 - 0: standard
 - 1: test
- **celltype**, Histological type of tumor
 - 1: squamous
 - 2: smallcell
 - 3: adeno
 - 4: large cell
- **time**, survival time
- **status**, censoring status
 - 0
 - 1
- **karno**, A measure at randomization, of the patient's performance status (Karnofsky rating) (100=good)
 - 10 ~ 30: completely hospitalized
 - 40 ~ 60: partial confinement
 - 70 ~ 90: able to care for self
- **diagtime**, Time in months from diagnosis to randomization
- **age**, Age in years
- **prior**, Prior therapy
 - 0: no
 - 10: yes

⁴³2017/12/29일 기준으로. 이 패키지의 원래 이름은 randomSurvivalForest였으나, 바뀌었다.

	trt	celltype	time	status	karno	diagtime	age	prior
All	1: 69 2: 68	1: 35 2: 48 3: 27 4: 27	Min. 1.000 1st Qu. 25.000 Median 80.000 Mean 121.600 3rd Qu. 144.000 Max. 999.000 Std.Dev. 157.817	0: 9 1: 128	10: 1 70: 23 20: 7 75: 2 30: 14 80: 24 40: 16 85: 1 50: 14 90: 7 60: 27 99: 1	Min. 1.000 1st Qu. 3.000 Median 5.000 Mean 8.774 3rd Qu. 11.000 Max. 87.000 Std.Dev. 10.612	Min. 34.000 1st Qu. 51.000 Median 62.000 Mean 58.310 3rd Qu. 66.000 Max. 81.000 Std.Dev. 10.542	0: 97 10: 40
trt==1	1: 69 2: 0	1: 15 2: 30 3: 9 4: 15	Min. 3.000 1st Qu. 25.000 Median 97.000 Mean 115.100 3rd Qu. 153.000 Max. 553.000 Std.Dev. 112.740	0: 5 1: 64	10: 3 70: 10 20: 6 75: 1 30: 9 80: 16 40: 7 85: 2 50: 15 99: 99	Min. 1.000 1st Qu. 4.000 Median 5.000 Mean 8.652 3rd Qu. 11.000 Max. 58.000 Std.Dev. 8.761	Min. 34.000 1st Qu. 49.000 Median 62.000 Mean 57.510 3rd Qu. 66.000 Max. 81.000 Std.Dev. 10.811	0: 48 10: 21
trt==2	1: 0 2: 68	1: 20 2: 18 3: 18 4: 12	Min. 1.000 1st Qu. 24.750 Median 52.500 Mean 128.210 3rd Qu. 117.250 Max. 999.000 Std.Dev. 0.237	0: 4 1: 64	10: 1 70: 13 20: 4 75: 1 30: 8 80: 8 40: 7 85: 1 50: 7 90: 5 60: 12 99: 1	Min. 1.000 1st Qu. 3.000 Median 4.500 Mean 8.897 3rd Qu. 11.000 Max. 87.000 Std.Dev. 12.274	Min. 35.000 1st Qu. 52.000 Median 62.000 Mean 59.120 3rd Qu. 66.000 Max. 81.000 Std.Dev. 10.278	0: 49 10: 19
celltype=="squamous"	1: 15 2: 20	1: 35 2: 3: 4:	Min. 1.000 1st Qu. 31.500 Median 111.000 Mean 200.200 3rd Qu. 262.500 Max. 999.000 Std.Dev. 248.232	0: 4 1: 31	10: 3 70: 8 20: 2 75: 5 30: 2 80: 5 40: 2 85: 5 50: 5 90: 4 60: 6 99: 99	Min. 1.000 1st Qu. 4.000 Median 7.000 Mean 11.030 3rd Qu. 12.500 Max. 58.000 Std.Dev. 11.529	Min. 35.000 1st Qu. 51.500 Median 62.000 Mean 58.460 3rd Qu. 64.500 Max. 81.000 Std.Dev. 10.373	0: 21 10: 14
celltype=="smallcell"	1: 30 2: 18	1: 2: 48 3: 4:	Min. 2.000 1st Qu. 20.000 Median 51.000 Mean 71.670 3rd Qu. 97.500 Max. 392.000 Std.Dev. 85.775	0: 3 1: 45	10: 3 70: 7 20: 3 75: 1 30: 8 80: 6 40: 7 85: 1 50: 4 90: 1 60: 11 99: 99	Min. 1.000 1st Qu. 2.000 Median 4.000 Mean 9.250 3rd Qu. 11.000 Max. 87.000 Std.Dev. 13.909	Min. 35.000 1st Qu. 54.750 Median 62.500 Mean 59.880 3rd Qu. 67.000 Max. 72.000 Std.Dev. 9.920	0: 37 10: 11
celltype=="adeno"	1: 9 2: 18	1: 2: 3: 27 4:	Min. 3.000 1st Qu. 21.500 Median 51.000 Mean 64.110 3rd Qu. 91.000 Max. 186.000 Std.Dev. 50.591	0: 1 1: 26	10: 1 70: 3 20: 1 75: 6 30: 1 80: 6 40: 6 85: 1 50: 3 90: 1 60: 4 99: 1	Min. 2.000 1st Qu. 3.000 Median 4.000 Mean 5.630 3rd Qu. 5.000 Max. 22.000 Std.Dev. 4.765	Min. 34.000 1st Qu. 50.000 Median 61.000 Mean 57.410 3rd Qu. 63.000 Max. 81.000 Std.Dev. 11.318	0: 22 10: 5
celltype=="large"	1: 15 2: 12	1: 2: 3: 4: 27	Min. 12.000 1st Qu. 76.500 Median 1556.000 Mean 166.100 3rd Qu. 223.500 Max. 553.000 Std.Dev. 124.221	0: 1 1: 26	10: 3 70: 5 20: 3 75: 1 30: 3 80: 7 40: 1 85: 1 50: 2 90: 2 60: 6 99: 99	Min. 1.000 1st Qu. 4.000 Median 8.000 Mean 8.148 3rd Qu. 12.000 Max. 18.000 Std.Dev. 4.990	Min. 37.000 1st Qu. 46.000 Median 62.000 Mean 56.220 3rd Qu. 65.500 Max. 68.000 Std.Dev. 11.164	0: 17 10: 10
status==0	1: 5 2: 4	1: 4 2: 3 3: 1 4: 1	Min. 25.000 1st Qu. 87.000 Median 100.000 Mean 114.600 3rd Qu. 123.000 Max. 231.000 Std.Dev. 59.800	0: 9 1:	10: 2 70: 2 20: 5 75: 2 30: 8 80: 2 40: 1 85: 1 50: 1 90: 1 60: 1 99: 1	Min. 2.000 1st Qu. 3.000 Median 5.000 Mean 6.778 3rd Qu. 8.000 Max. 22.000 Std.Dev. 6.200	Min. 36.000 1st Qu. 52.000 Median 55.000 Mean 55.440 3rd Qu. 62.000 Max. 70.000 Std.Dev. 10.297	0: 6 10: 3
status==1	1: 64 2: 64	1: 31 2: 45 3: 26 4: 26	Min. 1.000 1st Qu. 23.500 Median 62.000 Mean 122.100 3rd Qu. 145.800 Max. 999.000 Std.Dev. 162.610	0: 1: 128	10: 1 70: 21 20: 7 75: 2 30: 14 80: 22 40: 15 85: 1 50: 13 90: 6 60: 26 99: 99	Min. 1.000 1st Qu. 3.000 Median 5.000 Mean 8.914 3rd Qu. 11.000 Max. 87.000 Std.Dev. 10.857	Min. 34.000 1st Qu. 50.750 Median 62.000 Mean 58.510 3rd Qu. 66.000 Max. 81.000 Std.Dev. 10.569	0: 91 10: 37
prior=0	1: 48 2: 49	1: 21 2: 37 3: 22 4: 17	Min. 1.000 1st Qu. 29.000 Median 80.000 Mean 112.200 3rd Qu. 139.000 Max. 587.000 Std.Dev. 121.944	0: 6 1: 91	10: 1 70: 15 20: 5 75: 2 30: 10 80: 17 40: 9 85: 1 50: 10 90: 6 60: 20 99: 1	Min. 1.000 1st Qu. 3.000 Median 4.000 Mean 5.938 3rd Qu. 7.000 Max. 36.000 Std.Dev. 5.866	Min. 34.000 1st Qu. 52.000 Median 62.000 Mean 58.98 3rd Qu. 66.00 Max. 81.00 Std.Dev. 10.708	0: 97 10:
prior=10	1: 21 2: 19	1: 14 2: 11 3: 5 4: 10	Min. 1.000 1st Qu. 17.500 Median 69.000 Mean 144.600 3rd Qu. 182.800 Max. 999.000 Std.Dev. 222.449	0: 3 1: 37	10: 2 70: 8 20: 4 75: 7 30: 4 80: 7 40: 7 85: 1 50: 4 90: 1 60: 7 99: 99	Min. 2.000 1st Qu. 8.000 Median 12.000 Mean 15.650 3rd Qu. 18.250 Max. 87.000 Std.Dev. 15.476	Min. 36.000 1st Qu. 50.000 Median 60.000 Mean 56.670 3rd Qu. 65.000 Max. 70.000 Std.Dev. 10.070	0: 40 10:

Table 82: Veteran's administration lung cancer data from Kalbfleisch and Prentice (1980) 테이터 요약

25.2.4.2 분석

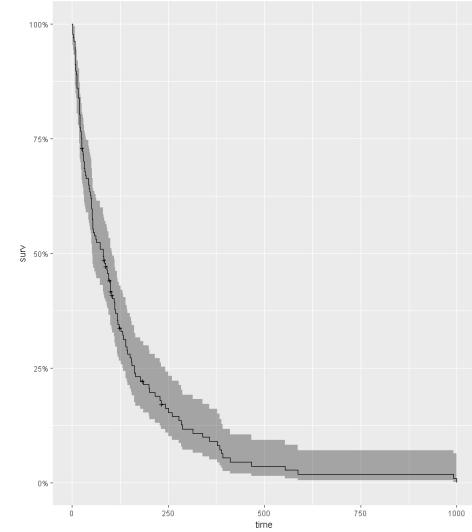
1. Kaplan-Meier Analysis

시간 경과에 따른 생존 확률에 대한 Kaplan-Meier 추정치와 생존 곡선을 다음과 같이 구하였다.

```
1 km_fit <- survfit(Surv(time, status) ~ 1, data=veteran)
2 summary(km_fit, times = c(1,30,60,90*(1:10)))
3 autoplot(km_fit)
```

```
> km_fit <- survfit(Surv(time, status) ~ 1, data=veteran)
> summary(km_fit, times = c(1,30,60,90*(1:10)))
Call: survfit(formula = Surv(time, status) ~ 1, data = veteran)

time n.risk n.event survival std.err lower 95% CI upper 95% CI
  1    137      2    0.985  0.0102   0.96552  1.0000
  30    97      39    0.700  0.0392   0.62774  0.7816
  60    73      22    0.538  0.0427   0.46070  0.6288
  90    62      10    0.464  0.0428   0.38731  0.5560
 180    27      30    0.222  0.0369   0.16066  0.3079
 270    16      9    0.144  0.0319   0.09338  0.2223
 360    10      6    0.090  0.0265   0.05061  0.1602
 450     5      5    0.045  0.0194   0.01931  0.1049
 540     4      1    0.036  0.0175   0.01389  0.0934
 630     2      2    0.018  0.0126   0.00459  0.0707
 720     2      0    0.018  0.0126   0.00459  0.0707
 810     2      0    0.018  0.0126   0.00459  0.0707
 900     2      0    0.018  0.0126   0.00459  0.0707
```

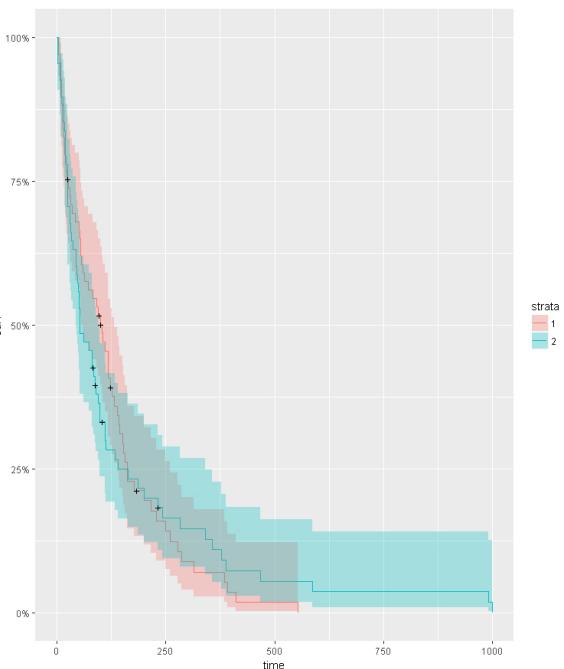


다음으로, treatment에 대한 생존 곡선을 다음과 같이 구하였다.

```
1 km_fit <- survfit(Surv(time, status) ~ trt, data=veteran)
2 summary(km_fit, times = c(1,30,60,90*(1:10)))
```

```
> km_trt_fit <- survfit(Surv(time, status) ~ trt, data=veteran)
> summary(km_trt_fit)
Call: survfit(formula = Surv(time, status) ~ trt, data = veteran)

trt=1          trt=2
time n.risk n.event survival std.err lower 95% CI upper 95% CI
  3    69      1    0.9584  0.0414   0.90000  1.0000
  4    68      1    0.9710  0.0242   0.93223  1.0000
  7    67      1    0.9565  0.0246   0.90959  1.0000
  9    66      1    0.9565  0.0246   0.90959  1.0000
 10    64      2    0.8988  0.0583   0.83000  0.973
 12    62      2    0.8751  0.0424   0.81792  0.942
 13    61      2    0.8751  0.0424   0.81792  0.942
 15    59      1    0.8445  0.0441   0.75849  0.912
 18    57      2    0.7971  0.0484   0.70764  0.898
 20    55      1    0.7971  0.0484   0.70764  0.898
 21    54      1    0.7681  0.0508   0.67472  0.874
 22    53      1    0.7536  0.0519   0.65811  0.882
 23    52      1    0.7536  0.0519   0.65811  0.882
 30    50      1    0.7241  0.0539   0.62158  0.818
 31    49      1    0.7241  0.0539   0.62158  0.818
 34    48      1    0.6943  0.0551   0.59368  0.852
 35    48      1    0.6943  0.0551   0.59368  0.852
 42    47      1    0.6797  0.0563   0.57785  0.800
 50    46      1    0.6797  0.0563   0.57785  0.800
 52    45      1    0.6502  0.0578   0.54645  0.777
 54    44      1    0.6502  0.0578   0.54645  0.777
 56    42      1    0.6059  0.0595   0.50040  0.734
 59    41      1    0.5911  0.0595   0.48526  0.720
 60    40      1    0.5911  0.0595   0.48526  0.720
 72    39      1    0.5613  0.0603   0.45159  0.693
 73    38      1    0.5613  0.0603   0.45159  0.693
 74    37      1    0.5320  0.0604   0.42577  0.665
 75    36      1    0.5172  0.0605   0.41116  0.651
 76    34      1    0.5172  0.0605   0.41116  0.651
 103    32      1    0.4983  0.0607   0.38070  0.621
 104    31      1    0.4983  0.0607   0.38070  0.621
 105    30      1    0.4549  0.0607   0.35028  0.591
 110    30      1    0.4239  0.0607   0.35028  0.591
 118    27      1    0.3979  0.0607   0.30357  0.545
 122    26      1    0.3922  0.0598   0.29069  0.529
 124    25      1    0.3922  0.0598   0.29069  0.529
 132    23      1    0.3595  0.0590   0.26031  0.496
 139    21      1    0.3483  0.0589   0.25020  0.486
 143    21      1    0.3483  0.0587   0.23057  0.483
 144    20      1    0.3103  0.0575   0.21193  0.446
 152    19      1    0.3103  0.0575   0.21193  0.446
 153    18      1    0.2778  0.0559   0.18725  0.412
 154    18      1    0.2778  0.0559   0.18725  0.412
 162    16      2    0.2388  0.0527   0.14563  0.359
 162    16      2    0.2388  0.0527   0.14563  0.359
 177    14      1    0.2124  0.0514   0.13213  0.341
 178    14      1    0.2124  0.0514   0.13213  0.341
 209    13      1    0.1770  0.0486   0.10340  0.303
 215    13      1    0.1770  0.0486   0.10340  0.303
 219    12      1    0.1346  0.0448   0.07614  0.263
 219    12      1    0.1346  0.0448   0.07614  0.263
 260    8      1    0.1239  0.0426   0.06318  0.243
 279    7      1    0.1239  0.0426   0.06318  0.243
 287    6      1    0.0885  0.0373   0.03895  0.201
 314    5      1    0.0531  0.0295   0.01788  0.158
 384    4      1    0.0531  0.0295   0.01788  0.158
 392    3      1    0.0354  0.0244   0.00017  0.137
 433    3      1    0.0354  0.0244   0.00017  0.137
 553    1      1    0.0000  NA      NA      NA
```



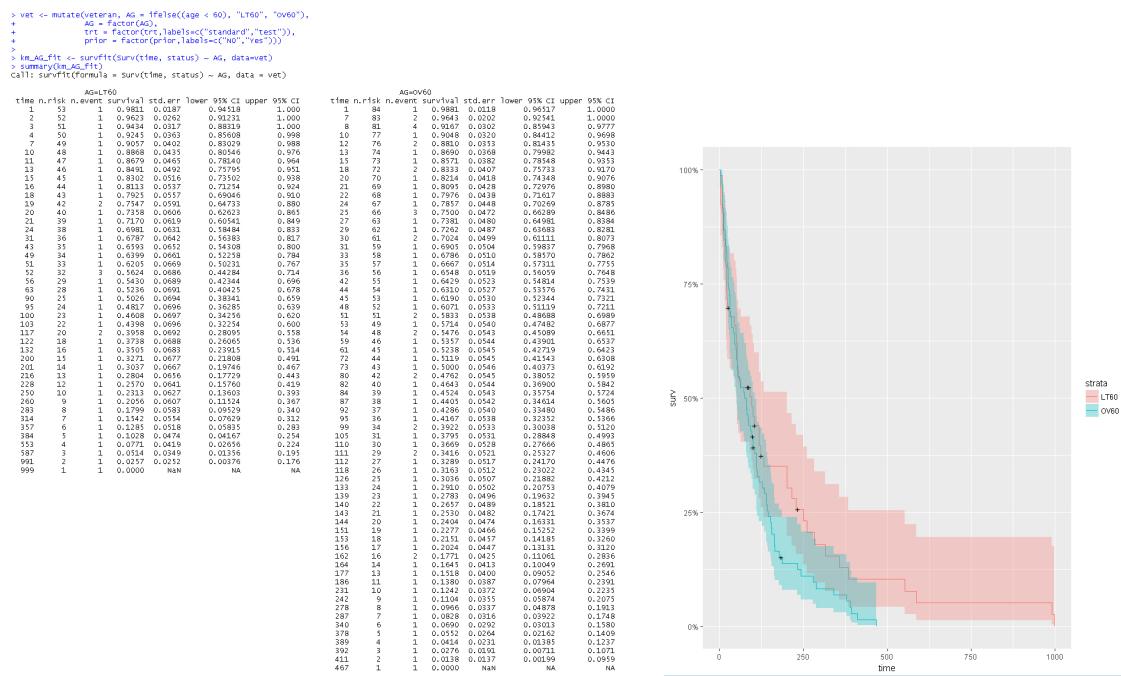
또 다른 하나는, 나이에 따른 생존 곡선을 구하였다. 이 때, age가 범주형 데이터는 아니기 때문에, 60세 미만을 LT60, 60세 이상을 GT60으로 표기하는 새 벡터 AG를 만들었다. 또한 trt와 prior를 요인 변수로 담은 새 데이터 프레임을 만들었다.

두 가지 곡선이 처음 50일 동안에는 많이 겹쳐보이지만, 젊은 환자일수록 1년 이상 생존할 가능성이 더 높아졌다.

```

1  vet <- mutate(veteran, AG = ifelse((age < 60), "LT60", "OV60"),
2    AG = factor(AG),
3    trt = factor(trt, labels=c("standard", "test")),
4    prior = factor(prior, labels=c("No", "Yes")))
5
6  km_AG_fit <- survfit(Surv(time, status) ~ AG, data=vet)
7  summary(km_AG_fit)
8  autoplot(km_AG_fit)

```



2. Cox-regression

모든 공변량을 사용하는 Cox-regression을 적합하여 보았다.

```

1 cox <- coxph(Surv(time, status) ~ trt + celltype + karno + diagtime + age + prior , data = vet)
2 summary(cox)
3
4 cox_fit <- survfit(cox)
5 autoplot(cox_fit)

```

```

> cox <- coxph(Surv(time, status) ~ trt + celltype + karno + diagtime + age + prior , data = vet)
> summary(cox)
Call:
coxph(formula = Surv(time, status) ~ trt + celltype + karno +
    diagtime + age + prior, data = vet)

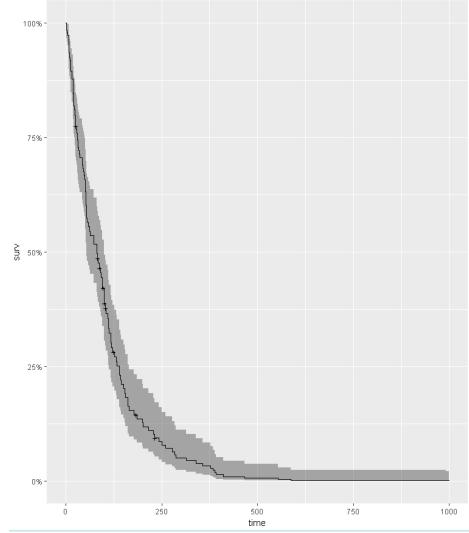
n= 137, number of events= 128

            coef  exp(coef)  se(coef)      z Pr(>|z|)
trttest     2.946e-01 1.343e+00 2.075e-01 1.419  0.15577
celltypesmallcell 8.616e-01 2.367e+00 2.753e-01 3.130  0.00275 ** 
celltypeadeno   1.196e+00 3.307e+00 3.009e-01 3.975 7.05e-05 *** 
celltypelarge    4.013e-01 1.494e+00 2.827e-01 1.420  0.15574
karno        -3.282e-02 9.677e-03 5.508e-03 -5.958 2.55e-09 *** 
diagtime      8.132e-05 1.000e+00 9.136e-03 0.009  0.99290
age          -8.706e-03 9.913e-03 9.300e-03 -0.936  0.34920
priorYes     7.159e-02 1.074e+00 2.323e-01 0.308  0.75794
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

exp(coef) exp(-coef) lower .95 upper .95
trttest      1.3426  0.7448  0.8939  2.0166
celltypesmallcell 2.3669  0.4225  1.3799  4.0597
celltypeadeno   3.2071  0.2024  1.8236  5.9647
celltypelarge    1.4938  0.6695  0.8583  2.5996
karno         0.9677  1.0334  0.9573  0.9782
diagtime      1.0001  0.9999  0.9823  1.0182
age           0.9913  1.0087  0.9734  1.0096
priorYes      1.0742  0.9309  0.6813  1.6937

Concordance= 0.736  (se = 0.03 )
Rsquare= 0.364  (max possible= 0.999 )
Likelihood ratio test= 62.1  on 8 df,  p=1.799e-10
Wald test       = 62.37  on 8 df,  p=1.596e-10
Score (logrank) test = 66.74  on 8 df,  p=2.186e-11

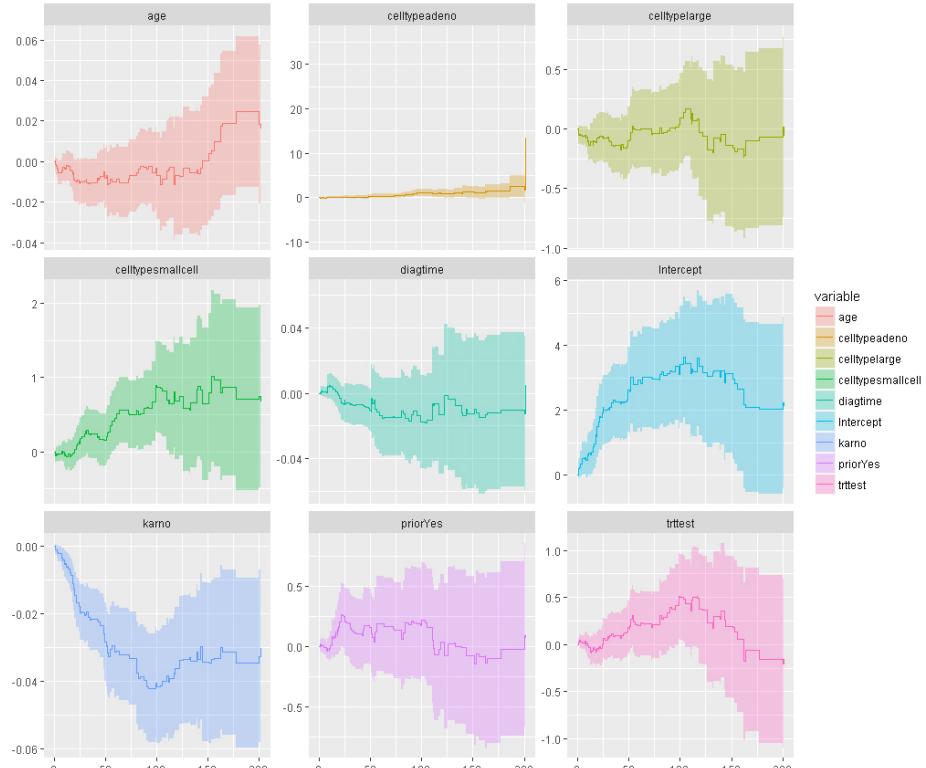
```



celltypesmallcell, celltypeadeno, karno에 대한 계수가 유의한 것으로 표시되었다.

다만, Cox Proportional Hazard 모델은 robust한 것으로 알려져있지만, 공변량이 시간에 따라 변하지 않는다고 가정하고 있으며, Terry Therneau et al.(2014)??에 따르면 Karnofsky는 실제로는 시간에 의존적이므로 Cox 모델에 대한 가정은 충족되지 않는다고 한다. 다음 그림을 통해, 공변량의 효과가 시간에 따라 어떻게 변하는지를 보여준다. Karnofsky의 경사면이 매우 가파른 모습을 볼 수 있다.

```
1 autoplot(aa_fit)
```



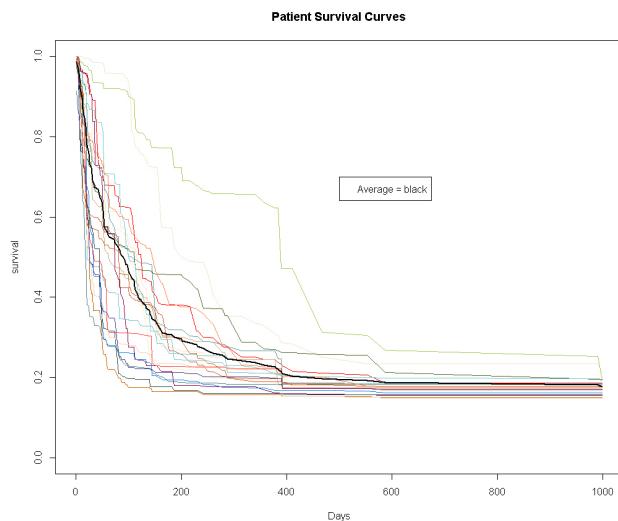
3. Random Forest

Cox-regression에서와 동일한 변수를 사용하였고, 모든 한자의 전체 평균 생존률을 나타내는 곡선과, 무작위로 선택된 환자 20명의 생존률을 그린다. Random Forest는 변수의 수가 많을 때 분석하기가 용이하다.

```

1 # ranger model
2 r_fit <- ranger(Surv(time, status) ~ trt + celltype +
3   karno + diagtime + age + prior,
4   data = vet,
5   mtry = 4,
6   importance = "permutation",
7   splitrule = "extratrees",
8   verbose = TRUE)
9
10 # Average the survival models
11 death_times <- r_fit$unique.death.times
12 surv_prob <- data.frame(r_fit$survival)
13 avg_prob <- sapply(surv_prob, mean)
14
15 # Plot the survival models for each patient
16 plot(r_fit$unique.death.times, r_fit$survival[,],
17   type = "l",
18   ylim = c(0,1),
19   col = "red",
20   xlab = "Days",
21   ylab = "survival",
22   main = "Patient Survival Curves")
23
24 #
25 cols <- colors()
26 for (n in sample(c(2:dim(vet)[1]), 20)){
27   lines(r_fit$unique.death.times, r_fit$survival[n,], type = "l", col = cols[n])
28 }
29 lines(death_times, avg_prob, lwd = 2)
30 legend(500, 0.7, legend = c('Average = black'))
31

```



다음은 변수의 중요도와 C-index를 계산하여 출력한 결과이다. karno, celltype이 비교적 중요한 두 가지 flag로 표시되었다.

```

1 vi <- data.frame(sort(round(r_fit$variable.importance, 4), decreasing = TRUE))
2 names(vi) <- "importance"
3 head(vi)
4
5 cat("Prediction Error = 1 - Harrell's c-index = ", r_fit$prediction.error)

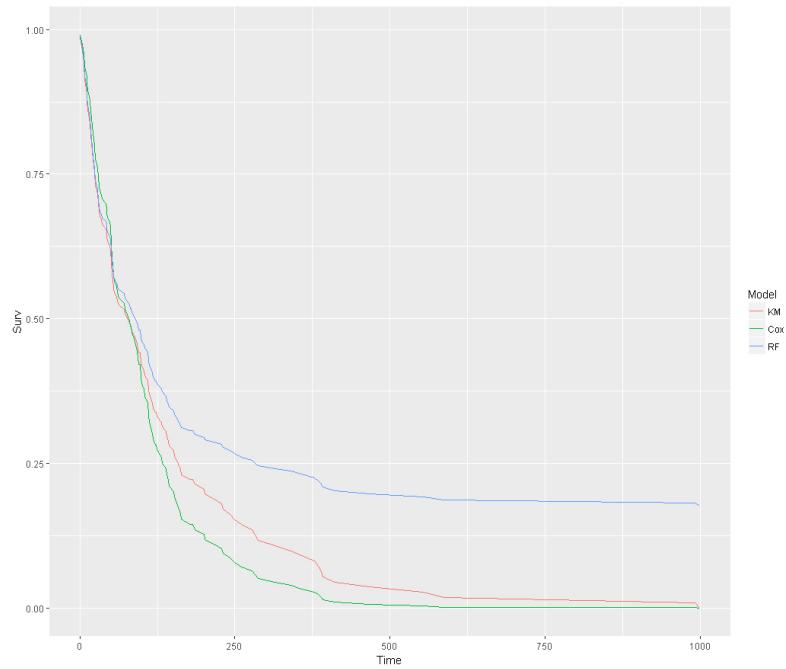
```

```

> vi <- data.frame(sort(round(r_fit$variable.importance, 4), decreasing = TRUE))
> names(vi) <- "importance"
> head(vi)
  importance
karno      0.0738
celltype   0.0313
diagtime   0.0027
trt        0.0022
prior      -0.0004
age        -0.0026
> cat("Prediction Error = 1 - Harrell's c-index = ", r_fit$prediction.error)
Prediction Error = 1 - Harrell's c-index =  0.3038392>

```

다음은 앞서 분석에 사용한 3가지 모델을 통해 구한 생존률을 비교하기 위한 그래프이다.



25.2.5 Worcester Heart Attack Study(WHAS)

데이터 소개 작성중...

25.2.5.1 변수 요약

Instance 수는 328개 있으며, 총 8개의 변수가 있다. 각 변수의 정확한 description은 아직 파악하지 못하였다.

- **t**, a numeric vector

- **e**, a binary vector

– 0

– 1

- **x1**, a binary vector

– 0

– 1

- **x2**, a numeric vector

- **x3**, a binary vector

– 0

– 1

- **x4**, a numeric vector

- **x5**, a binary vector

– 0

– 1

- **x6**, a binary vector

– 0

– 1

	t	e	x1	x2	x3	x4	x5	x6	
All	Min. 0.6608 1st Qu. 510.5929 Median 1175.0000 Mean 1057.0279 3rd Qu. 1559.2130 Max. 1993.0000 Std.Dev. 634.7183		0: 190 1: 138	0: 266 1: 62	Min. 36.0000 1st Qu. 58.7500 Median 72.0000 Mean 69.0900 3rd Qu. 80.0000 Max. 104.0000 Std.Dev. 14.3586	0: 196 1: 132	Min. 13.0500 1st Qu. 23.6900 Median 26.4400 Mean 27.1200 3rd Qu. 30.3200 Max. 49.4200 Std.Dev. 5.5588	0: 231 1: 97	0: 219 1: 109
e==0	Min. 371.0000 1st Qu. 1108.0000 Median 1337.0000 Mean 1314.0000 3rd Qu. 1874.0000 Max. 1993.0000 Std.Dev. 529.8522		0: 190 1:	0: 190 1:	Min. 36.0000 1st Qu. 53.0000 Median 64.0000 Mean 63.4400 3rd Qu. 74.0000 Max. 89.0000 Std.Dev. 13.1674	0: 125 1: 65	Min. 15.8700 1st Qu. 24.6700 Median 27.3900 Mean 28.3600 3rd Qu. 31.6400 Max. 47.3000 Std.Dev. 4.9793	0: 164 1: 26	0: 139 1: 51
e==1	Min. 0.6608 1st Qu. 171.0753 Median 599.2713 Mean 702.6750 3rd Qu. 1120.5550 Max. 1965.9110 Std.Dev. 596.6045		0: 1: 138	0: 76 1: 62	Min. 42.0000 1st Qu. 72.0000 Median 78.0000 Mean 76.8700 3rd Qu. 84.0000 Max. 104.0000 Std.Dev. 12.1534	0: 71 1: 67	Min. 13.0500 1st Qu. 21.9500 Median 24.5400 Mean 25.4200 3rd Qu. 28.5000 Max. 49.4200 Std.Dev. 5.8747	0: 67 1: 71	0: 80 1: 58
x1==0	Min. 3.3020 1st Qu. 604.2500 Median 1249.0000 Mean 1182.0690 3rd Qu. 1836.0000 Max. 1993.0000 Std.Dev. 587.2176		0: 190 1: 76	0: 266 1:	Min. 36.0000 1st Qu. 57.0000 Median 68.0000 Mean 67.1500 3rd Qu. 77.7500 Max. 104.0000 Std.Dev. 14.2237	0: 164 1: 102	Min. 13.0500 1st Qu. 24.0000 Median 26.6000 Mean 27.5600 3rd Qu. 30.7000 Max. 49.4200 Std.Dev. 5.2728	0: 198 1: 68	0: 183 1: 83
x1==1	Min. 0.6608 1st Qu. 33.4698 Median 368.4169 Mean 520.5595 3rd Qu. 728.7963 Max. 1965.9110 Std.Dev. 548.5695		0: 1: 62	0: 1: 62	Min. 50.0000 1st Qu. 72.0000 Median 78.0000 Mean 77.4400 3rd Qu. 85.7500 Max. 99.0000 Std.Dev. 11.7914	0: 32 1: 30	Min. 13.0500 1st Qu. 20.6900 Median 24.4800 Mean 25.2500 3rd Qu. 28.5300 Max. 49.4200 Std.Dev. 6.3599	0: 33 1: 29	0: 36 1: 26
x3==0	Min. 0.6608 1st Qu. 543.0000 Median 1212.0835 Mean 1084.3396 3rd Qu. 1531.3782 Max. 1993.0000 Std.Dev. 626.5321		0: 125 1: 71	0: 164 1: 32	Min. 36.0000 1st Qu. 53.0000 Median 65.0000 Mean 64.5900 3rd Qu. 76.0000 Max. 98.0000 Std.Dev. 13.7157	0: 196 1:	Min. 18.8700 1st Qu. 24.9400 Median 27.3900 Mean 28.1500 3rd Qu. 31.3600 Max. 47.3000 Std.Dev. 4.5170	0: 150 1: 46	0: 129 1: 67
x3==1	Min. 1.8180 1st Qu. 472.3100 Median 1134.8090 Mean 1016.4740 3rd Qu. 1621.4800 Max. 1980.0000 Std.Dev. 646.9456		0: 65 1: 67	0: 102 1: 30	Min. 43.0000 1st Qu. 69.5000 Median 77.0000 Mean 75.7700 3rd Qu. 84.0000 Max. 104.0000 Std.Dev. 12.6229	0: 132	Min. 13.0500 1st Qu. 21.2400 Median 24.4100 Mean 25.6000 3rd Qu. 28.5000 Max. 49.4200 Std.Dev. 6.5437	0: 81 1: 51	0: 90 1: 42
x5==0	Min. 0.6608 1st Qu. 549.5000 Median 1266.0000 Mean 1150.2817 3rd Qu. 1836.0000 Max. 1993.0000 Std.Dev. 629.1501		0: 164 1: 67	0: 198 1: 33	Min. 36.0000 1st Qu. 56.0000 Median 68.0000 Mean 66.9200 3rd Qu. 78.0000 Max. 99.0000 Std.Dev. 14.1034	0: 150 1: 81	Min. 23.9234 1st Qu. 25.7689 Median 26.1757 Mean 20.0287 3rd Qu. 21.9511 Max. 22.1429 Std.Dev. 5.2158	0: 231 1:	0: 168 1: 63
x5==1	Min. 2.3690 1st Qu. 341.9300 Median 722.7460 Mean 834.9490 3rd Qu. 1234.0000 Max. 1965.9110 Std.Dev. 594.2783		0: 26 1: 71	0: 68 1: 29	Min. 39.0000 1st Qu. 65.0000 Median 76.0000 Mean 74.2600 3rd Qu. 83.0000 Max. 104.0000 Std.Dev. 13.6902	0: 46 1: 51	Min. 13.1400 1st Qu. 22.8500 Median 27.4700 Mean 27.2400 3rd Qu. 30.5400 Max. 49.4200 Std.Dev. 6.3292	0: 97 1:	0: 51 1: 46
x6==0	Min. 0.6608 1st Qu. 539.9210 Median 1231.0000 Mean 1105.4107 3rd Qu. 1774.5565 Max. 1993.0000 Std.Dev. 640.6466		0: 139 1: 80	0: 183 1: 36	Min. 37.0000 1st Qu. 58.0000 Median 71.0000 Mean 68.8300 3rd Qu. 80.0000 Max. 104.0000 Std.Dev. 14.8294	0: 129 1: 90	Min. 13.0500 1st Qu. 23.3700 Median 25.97000 Mean 26.8900 3rd Qu. 29.5800 Max. 49.4200 Std.Dev. 5.6702	0: 168 1: 51	0: 219 1:
x6==1	Min. 0.6608 1st Qu. 171.0753 Median 599.2713 Mean 702.6750 3rd Qu. 1120.5550 Max. 1965.9110 Std.Dev. 596.6045		0: 1: 138	0: 76 1: 62	Min. 42.0000 1st Qu. 72.0000 Median 78.0000 Mean 76.8700 3rd Qu. 84.0000 Max. 104.0000 Std.Dev. 12.1534	0: 71 1: 67	Min. 13.0500 1st Qu. 21.9500 Median 24.5400 Mean 25.4200 3rd Qu. 28.5000 Max. 49.4200 Std.Dev. 5.8747	0: 67 1: 71	0: 80 1: 58

Table 83: WHAS 테이터 요약(cont'd)

25.2.6 Molecular Taxonomy of Breast Cancer International Consortium(Metabric)

데이터 소개 작성중...

25.2.6.1 변수 요약

Instance 수는 381개 있으며, 총 11개의 변수가 있다. 각 변수의 정확한 description은 아직 파악하지 못하였다.

- `t`, a numeric vector

- `e`, a binary vector

– 0

– 1

- `x1`, a numeric vector

- `x2`, a numeric vector

- `x3`, a numeric vector

- `x4`, a numeric vector

- `x5`, a binary vector

– 0

– 1

- `x6`, a binary vector

– 0

– 1

- `x7`, a binary vector

– 0

– 1

- `x8`, a binary vector

– 0

– 1

- `x9`, a binary vector

– 0

– 1

	t	e	x1	x2	x3	x4	x5	x6	x7	x8	x9		
All			Min. 0.000 1st Qu. 63.200 Median 122.800 Mean 131.200 3rd Qu. 190.200 Max. 337.000 Std.Dev. 77.697	0: 165 1: 216	Min. 5.288 1st Qu. 5.690 Median 5.910 Mean 6.216 3rd Qu. 6.457 Max. 13.120 Std.Dev. 0.888	Min. 5.064 1st Qu. 5.380 Median 5.380 Mean 5.840 3rd Qu. 6.197 Max. 9.928 Std.Dev. 1.022	Min. 6.919 1st Qu. 9.936 Median 10.515 Mean 10.786 3rd Qu. 11.240 Max. 14.464 Std.Dev. 0.345	Min. 5.204 1st Qu. 5.645 Median 5.846 Mean 5.897 3rd Qu. 6.088 Max. 7.206 Std.Dev. 0.345	0: 128 1: 253	0: 153 1: 228	0: 307 1: 74	0: 88 1: 293	Min. 28.370 1st Qu. 51.420 Median 63.020 Mean 61.700 3rd Qu. 70.750 Max. 90.020 Std.Dev. 12.714
c==0			Min. 0.000 1st Qu. 112.900 Median 165.700 Mean 161.100 3rd Qu. 213.400 Max. 337.000 Std.Dev. 71.552	0: 165 1: 216	Min. 5.290 1st Qu. 5.743 Median 6.016 Mean 6.221 3rd Qu. 6.477 Max. 10.118 Std.Dev. 0.723	Min. 5.092 1st Qu. 5.388 Median 5.936 Mean 6.270 3rd Qu. 6.816 Max. 9.928 Std.Dev. 1.432	Min. 7.224 1st Qu. 9.940 Median 10.442 Mean 10.819 3rd Qu. 11.279 Max. 14.464 Std.Dev. 1.432	Min. 5.268 1st Qu. 5.588 Median 5.762 Mean 5.863 3rd Qu. 6.047 Max. 7.206 Std.Dev. 0.369	0: 51 1: 114	0: 61 1: 104	0: 128 1: 37	0: 38 1: 127	Min. 29.98 1st Qu. 47.92 Median 55.95 Mean 56.66 3rd Qu. 65.58 Max. 83.99 Std.Dev. 11.129
e==1			Min. 5.833 1st Qu. 46.733 Median 94.833 Mean 108.385 3rd Qu. 152.500 Max. 300.700 Std.Dev. 74.547	0: 216 1: 216	Min. 5.288 1st Qu. 5.655 Median 5.844 Mean 6.211 3rd Qu. 6.327 Max. 13.120 Std.Dev. 0.997	Min. 2.064 1st Qu. 5.369 Median 5.705 Mean 6.142 3rd Qu. 6.600 Max. 9.522 Std.Dev. 0.997	Min. 6.919 1st Qu. 9.935 Median 10.565 Mean 10.760 3rd Qu. 11.221 Max. 14.464 Std.Dev. 1.291	Min. 5.204 1st Qu. 5.696 Median 5.882 Mean 5.922 3rd Qu. 6.115 Max. 7.152 Std.Dev. 0.480	0: 77 1: 139	0: 92 1: 124	0: 179 1: 37	0: 50 1: 166	Min. 28.370 1st Qu. 58.110 Median 67.820 Mean 65.540 3rd Qu. 74.440 Max. 90.020 Std.Dev. 12.527
x5==0			Min. 0.000 1st Qu. 45.330 Median 124.730 Mean 138.130 3rd Qu. 224.520 Max. 337.030 Std.Dev. 93.479	0: 51 1: 77	Min. 5.288 1st Qu. 5.785 Median 6.200 Mean 6.542 3rd Qu. 7.130 Max. 10.118 Std.Dev. 0.954	Min. 5.092 1st Qu. 5.329 Median 5.480 Mean 5.984 3rd Qu. 6.537 Max. 8.633 Std.Dev. 0.929	Min. 8.014 1st Qu. 9.909 Median 10.636 Mean 11.053 3rd Qu. 11.661 Max. 14.464 Std.Dev. 1.588	Min. 5.268 1st Qu. 5.709 Median 5.918 Mean 5.978 3rd Qu. 6.177 Max. 7.206 Std.Dev. 0.382	0: 128 1: 62	0: 66 1: 37	0: 91 1: 60	0: 68 1: 60	Min. 28.370 1st Qu. 47.730 Median 55.510 Mean 57.230 3rd Qu. 66.180 Max. 84.870 Std.Dev. 13.201
x5==1			Min. 2.000 1st Qu. 71.800 Median 122.500 Mean 127.700 3rd Qu. 177.300 Max. 318.200 Std.Dev. 68.282	0: 114 1: 139	Min. 5.290 1st Qu. 5.654 Median 5.828 Mean 6.050 3rd Qu. 6.165 Max. 13.120 Std.Dev. 0.804	Min. 5.064 1st Qu. 5.418 Median 5.984 Mean 6.305 3rd Qu. 6.889 Max. 9.928 Std.Dev. 1.051	Min. 6.919 1st Qu. 9.940 Median 10.473 Mean 10.650 3rd Qu. 11.115 Max. 14.464 Std.Dev. 1.197	Min. 5.204 1st Qu. 5.602 Median 5.805 Mean 5.855 3rd Qu. 6.037 Max. 7.152 Std.Dev. 0.318	0: 253 1: 166	0: 87 1: 37	0: 216 1: 233	0: 20 1: 233	Min. 33.800 1st Qu. 55.830 Median 65.040 Mean 63.960 3rd Qu. 72.560 Max. 90.020 Std.Dev. 11.860
x6==0			Min. 0.000 1st Qu. 63.200 Median 123.700 Mean 134.700 3rd Qu. 198.400 Max. 337.000 Std.Dev. 80.509	0: 61 1: 92	Min. 5.288 1st Qu. 5.677 Median 5.965 Mean 6.221 3rd Qu. 6.396 Max. 12.123 Std.Dev. 0.883	Min. 5.086 1st Qu. 5.377 Median 5.909 Mean 6.228 3rd Qu. 6.768 Max. 9.928 Std.Dev. 1.046	Min. 6.919 1st Qu. 10.103 Median 10.637 Mean 10.845 3rd Qu. 11.241 Max. 14.464 Std.Dev. 1.310	Min. 5.204 1st Qu. 5.632 Median 5.823 Mean 5.862 3rd Qu. 6.055 Max. 7.104 Std.Dev. 0.328	0: 66 1: 87	0: 153 1: 16	0: 137 1: 117	0: 36 1: 117	Min. 29.980 1st Qu. 51.740 Median 64.040 Mean 62.850 3rd Qu. 72.250 Max. 90.000 Std.Dev. 12.909
x6==1			Min. 2.000 1st Qu. 63.710 Median 120.300 Mean 128.840 3rd Qu. 184.890 Max. 318.200 Std.Dev. 75.839	0: 104 1: 124	Min. 5.290 1st Qu. 5.696 Median 5.889 Mean 6.212 3rd Qu. 6.469 Max. 13.120 Std.Dev. 0.893	Min. 5.064 1st Qu. 5.383 Median 5.5786 Mean 6.177 3rd Qu. 6.756 Max. 9.517 Std.Dev. 1.007	Min. 7.224 1st Qu. 9.852 Median 10.459 Mean 10.746 3rd Qu. 11.221 Max. 14.464 Std.Dev. 1.382	Min. 5.368 1st Qu. 5.660 Median 5.857 Mean 5.862 3rd Qu. 6.055 Max. 7.206 Std.Dev. 0.355	0: 62 1: 166	0: 228 1: 58	0: 170 1: 176	0: 52 1: 176	Min. 28.370 1st Qu. 51.380 Median 61.230 Mean 60.920 3rd Qu. 70.070 Max. 90.020 Std.Dev. 12.551
x7==0			Min. 0.000 1st Qu. 69.930 Median 128.100 Mean 137.790 3rd Qu. 197.350 Max. 337.030 Std.Dev. 77.648	0: 128 1: 179	Min. 5.288 1st Qu. 5.658 Median 5.860 Mean 6.097 3rd Qu. 6.241 Max. 12.123 Std.Dev. 0.770	Min. 5.064 1st Qu. 5.415 Median 5.976 Mean 6.293 3rd Qu. 6.905 Max. 9.928 Std.Dev. 1.049	Min. 6.919 1st Qu. 9.944 Median 10.515 Mean 10.694 3rd Qu. 11.194 Max. 14.464 Std.Dev. 1.204	Min. 5.268 1st Qu. 5.615 Median 5.818 Mean 5.867 3rd Qu. 6.051 Max. 7.206 Std.Dev. 0.322	0: 91 1: 216	0: 137 1: 170	0: 307 1: 262	0: 45 1: 262	Min. 29.980 1st Qu. 55.810 Median 65.580 Mean 64.370 3rd Qu. 72.670 Max. 90.020 Std.Dev. 11.756
x7==1			Min. 5.833 1st Qu. 44.475 Median 87.867 Mean 103.826 3rd Qu. 143.442 Max. 278.267 Std.Dev. 72.172	0: 37 1: 37	Min. 5.429 1st Qu. 5.926 Median 6.460 Mean 6.707 3rd Qu. 7.198 Max. 13.120 Std.Dev. 1.148	Min. 5.130 1st Qu. 5.311 Median 5.408 Mean 5.800 3rd Qu. 6.023 Max. 8.461 Std.Dev. 0.791	Min. 7.224 1st Qu. 9.874 Median 10.485 Mean 11.168 3rd Qu. 12.335 Max. 14.464 Std.Dev. 1.807	Min. 5.204 1st Qu. 5.755 Median 5.952 Mean 6.018 3rd Qu. 6.226 Max. 7.152 Std.Dev. 0.409	0: 37 1: 37	0: 16 1: 58	0: 74 1: 74	0: 43 1: 31	Min. 28.370 1st Qu. 44.480 Median 50.210 Mean 50.600 3rd Qu. 56.550 Max. 73.740 Std.Dev. 10.363
x8==0			Min. 0.000 1st Qu. 33.680 Median 102.070 Mean 116.970 3rd Qu. 187.820 Max. 295.330 Std.Dev. 88.936	0: 38 1: 50	Min. 5.397 1st Qu. 6.215 Median 6.935 Mean 7.056 3rd Qu. 7.615 Max. 13.120 Std.Dev. 1.176	Min. 5.092 1st Qu. 5.277 Median 5.373 Mean 5.479 3rd Qu. 5.477 Max. 8.206 Std.Dev. 0.467	Min. 8.014 1st Qu. 9.627 Median 10.365 Mean 11.111 3rd Qu. 12.393 Max. 14.464 Std.Dev. 1.904	Min. 5.204 1st Qu. 5.802 Median 6.068 Mean 6.122 3rd Qu. 6.424 Max. 7.206 Std.Dev. 0.433	0: 68 1: 20	0: 36 1: 52	0: 45 1: 43	0: 88 1: 88	Min. 28.37 1st Qu. 47.18 Median 55.37 Mean 56.01 3rd Qu. 63.77 Max. 84.76 Std.Dev. 12.239
x8==1			Min. 2.000 1st Qu. 72.670 Median 125.700 Mean 135.470 3rd Qu. 192.200 Max. 337.030 Std.Dev. 73.621	0: 127 1: 166	Min. 5.288 1st Qu. 5.638 Median 5.825 Mean 5.963 3rd Qu. 6.099 Max. 12.123 Std.Dev. 0.580	Min. 5.064 1st Qu. 5.482 Median 5.538 Mean 6.413 3rd Qu. 7.072 Max. 9.928 Std.Dev. 1.045	Min. 6.919 1st Qu. 10.033 Median 10.538 Mean 10.688 3rd Qu. 11.169 Max. 14.464 Std.Dev. 1.122	Min. 5.268 1st Qu. 5.601 Median 5.785 Mean 5.829 3rd Qu. 6.020 Max. 6.698 Std.Dev. 0.281	0: 60 1: 233	0: 117 1: 176	0: 262 1: 32	0: 293	Min. 29.980 1st Qu. 54.290 Median 64.990 Mean 63.400 3rd Qu. 72.280 Max. 90.020 Std.Dev. 12.371

Table 84: Metabric 페인트 요약

25.2.7 Simulated Treatment Data(Treatment)

데이터 소개 작성중...

25.2.7.1 변수 요약

Instance 수는 1000개 있으며, 총 14개의 변수가 있다. 각 변수의 정확한 description은 아직 파악하지 못하였다.

- `t`, a numeric vector

- `e`, a binary vector

– 0

– 1

- `hr`, a numeric vector

- `x1`, a numeric vector

- `x2`, a numeric vector

- `x3`, a numeric vector

- `x4`, a numeric vector

- `x5`, a numeric vector

- `x6`, a numeric vector

- `x7`, a numeric vector

- `x8`, a numeric vector

- `x9`, a numeric vector

- `x10`, a numeric vector

- `x11`, a binary vector

– 0

– 1

	t	e	hr	x1	x2	x3	x4
All	Min. 0.0122 1st Qu. 1.5043 Median 3.3492 Mean 4.9124 3rd Qu. 7.3970 Max. 15.0000 Std.Dev. 4.3726	0: 60 1: 940	Min. -0.7851 1st Qu. -0.4878 Median -1.1882 Mean 0.0000 3rd Qu. 0.3760 Max. 1.4631 Std.Dev. 0.6019	Min. -0.9963 1st Qu. -0.4839 Median -0.0100 Mean 0.0070 3rd Qu. 0.5155 Max. 0.9985 Std.Dev. 0.5702	Min. -0.9995 1st Qu. -0.4998 Median 0.0079 Mean -0.0028 3rd Qu. 0.4884 Max. 0.9972 Std.Dev. 0.5930	Min. -0.9996 1st Qu. -0.4949 Median 0.0250 Mean 0.0151 3rd Qu. 0.5264 Max. 0.9997 Std.Dev. 0.5877	Min. -0.9991 1st Qu. -0.5113 Median 0.0178 Mean 0.0028 3rd Qu. 0.4996 Max. 0.9993 Std.Dev. 0.5731
	x5	x6	x7	x8	x9	x10	x11
e==0	15: 60	0: 60 1:	Min. -0.7709 1st Qu. -0.6662 Median -0.4435 Mean -0.3120 3rd Qu. 0.5010 Max. 1.4298 Std.Dev. 0.5172	Min. -0.9747 1st Qu. -0.7045 Median -0.2078 Mean -0.1201 3rd Qu. 0.4361 Max. 0.9947 Std.Dev. 0.6377	Min. -0.9976 1st Qu. -0.6513 Median 0.0481 Mean 0.0391 3rd Qu. 0.7936 Max. 0.9530 Std.Dev. 0.7212	Min. -0.9685 1st Qu. -0.3268 Median 0.1849 Mean 0.1330 3rd Qu. 0.5259 Max. 0.9996 Std.Dev. 0.5881	Min. -0.9982 1st Qu. -0.5233 Median -0.0408 Mean -0.0131 3rd Qu. 0.5259 Max. 0.9996 Std.Dev. 0.5881
	x5	x6	x7	x8	x9	x10	x11
e==1	Min. -0.9940 1st Qu. -0.3286 Median 0.0479 Mean 0.0636 3rd Qu. 0.5302 Max. 0.9990 Std.Dev. 0.5297	Min. -0.9677 1st Qu. -0.2794 Median 0.1268 Mean 0.1448 3rd Qu. 0.5799 Max. 0.9895 Std.Dev. 0.5470	Min. -0.9204 1st Qu. -0.5082 Median -0.0741 Mean -0.0186 3rd Qu. 0.4406 Max. 0.9943 Std.Dev. 0.5715	Min. -0.9664 1st Qu. -0.6461 Median -0.0742 Mean -0.0535 3rd Qu. 0.5744 Max. 0.9849 Std.Dev. 0.6244	Min. -0.9751 1st Qu. -0.5268 Median 0.0169 Mean -0.0079 3rd Qu. 0.5008 Max. 0.9616 Std.Dev. 0.5678	Min. -0.9967 1st Qu. -0.5176 Median -0.0123 Mean -0.0317 3rd Qu. 0.4972 Max. 0.9295 Std.Dev. 0.5833	Min. -0.9858 1st Qu. -0.3714 Median 0.0431 Mean 0.0361 3rd Qu. 0.4332 Max. 0.9796 Std.Dev. 0.5612
	x5	x6	x7	x8	x9	x10	x11
x11==0	Min. 0.0122 1st Qu. 1.4176 Median 3.0821 Mean 4.2685 3rd Qu. 6.2966 Max. 14.9230 Std.Dev. 3.6639	0: 1: 940	Min. -0.7851 1st Qu. -0.4746 Median -0.1597 Mean 0.02000 3rd Qu. 0.4074 Max. 1.4631 Std.Dev. 0.6017	Min. -0.9995 1st Qu. -0.4850 Median 0.0071 Mean -0.0055 3rd Qu. 0.4707 Max. 0.9972 Std.Dev. 0.5626	Min. -0.9996 1st Qu. -0.5107 Median 0.0126 Mean 0.0076 3rd Qu. 0.5217 Max. 0.9993 Std.Dev. 0.5650	Min. -0.9991 1st Qu. -0.5138 Median 0.0171 Mean 0.006 3rd Qu. 0.5028 Max. 0.9993 Std.Dev. 0.5650	Min. -0.9991 1st Qu. -0.5138 Median 0.0171 Mean 0.006 3rd Qu. 0.5028 Max. 0.9993 Std.Dev. 0.5650
	x5	x6	x7	x8	x9	x10	x11
x11==1	Min. -0.9999 1st Qu. -0.4494 Median 0.0095 Mean 0.0083 3rd Qu. 0.4892 Max. 0.9991 Std.Dev. 0.5650	Min. -0.9993 1st Qu. -0.4904 Median -0.0323 Mean -0.0087 3rd Qu. 0.4925 Max. 0.9975 Std.Dev. 0.5730	Min. -0.9969 1st Qu. -0.4719 Median 0.0320 Mean 0.0146 3rd Qu. 0.5116 Max. 0.9998 Std.Dev. 0.5730	Min. -0.9981 1st Qu. -0.5326 Median -0.0364 Mean -0.0230 3rd Qu. 0.4924 Max. 0.9954 Std.Dev. 0.5772	Min. -0.9959 1st Qu. -0.4733 Median 0.0061 Mean 0.0052 3rd Qu. 0.5121 Max. 0.9985 Std.Dev. 0.5602	Min. -0.9983 1st Qu. -0.5234 Median -0.0423 Mean -0.0119 3rd Qu. 0.5269 Max. 0.9996 Std.Dev. 0.5887	Min. -0.9983 1st Qu. -0.5234 Median -0.0423 Mean -0.0119 3rd Qu. 0.5269 Max. 0.9996 Std.Dev. 0.5887
	t	e	hr	x1	x2	x3	x4
x11==0	Min. 0.0122 1st Qu. 1.6952 Median 3.5128 Mean 4.8911 3rd Qu. 7.3026 Max. 15.0000 Std.Dev. 4.2269	0: 24 1: 467	Min. -0.7794 1st Qu. -0.5045 Median -0.1918 Mean -0.0214 3rd Qu. 0.3455 Max. 1.4601 Std.Dev. 0.5973	Min. -0.9899 1st Qu. -0.4759 Median 0.0160 Mean 0.0245 3rd Qu. 0.5378 Max. 0.9984 Std.Dev. 0.5739	Min. -0.9982 1st Qu. -0.5199 Median 0.0121 Mean -0.0017 3rd Qu. 0.0525 Max. 0.9950 Std.Dev. 0.5806	Min. -0.9996 1st Qu. -0.5187 Median 0.0120 Mean 0.0001 3rd Qu. 0.5017 Max. 0.9997 Std.Dev. 0.5918	Min. -0.9996 1st Qu. -0.5064 Median 0.0109 Mean -0.0092 3rd Qu. 0.4926 Max. 0.9967 Std.Dev. 0.5622
	x5	x6	x7	x8	x9	x10	x11
x11==1	Min. -0.9990 1st Qu. -0.4447 Median 0.0101 Mean -0.0003 3rd Qu. 0.4649 Max. 0.9990 Std.Dev. 0.5631	Min. -0.9963 1st Qu. -0.4652 Median 0.0545 Mean 0.0252 3rd Qu. 0.5106 Max. 0.9870 Std.Dev. 0.5680	Min. -0.9946 1st Qu. -0.4482 Median 0.0876 Mean 0.0362 3rd Qu. 0.5347 Max. 0.9904 Std.Dev. 0.5760	Min. -0.9982 1st Qu. -0.5293 Median -0.0344 Mean -0.0178 3rd Qu. 0.4876 Max. 0.9949 Std.Dev. 0.5726	Min. -0.9959 1st Qu. -0.4770 Median 0.0393 Mean 0.0045 3rd Qu. 0.5124 Max. 0.9947 Std.Dev. 0.5608	Min. -0.9967 1st Qu. -0.5305 Median -0.0083 Mean 0.0071 3rd Qu. 0.5390 Max. 0.9996 Std.Dev. 0.6023	Min. -0.9967 1st Qu. -0.5305 Median -0.0083 Mean 0.0071 3rd Qu. 0.5390 Max. 0.9996 Std.Dev. 0.6023
	t	e	hr	x1	x2	x3	x4
x11==1	Min. 0.0315 1st Qu. 1.4123 Median 3.2296 Mean 4.9330 3rd Qu. 7.5196 Max. 15.0000 Std.Dev. 4.5128	0: 36 1: 473	Min. -0.7851 1st Qu. -0.4636 Median -0.1857 Mean 0.0207 3rd Qu. 0.4417 Max. 1.4493 Std.Dev. 0.6062	Min. -0.9963 1st Qu. -0.5103 Median 0.0181 Mean -0.0099 3rd Qu. 0.4513 Max. 0.9981 Std.Dev. 0.5666	Min. -0.9995 1st Qu. -0.4755 Median 0.0423 Mean -0.0039 3rd Qu. 0.4563 Max. 0.9972 Std.Dev. 0.5661	Min. -0.99551 1st Qu. -0.44000 Median 0.04266 Mean 0.02953 3rd Qu. 0.54078 Max. 0.9993 Std.Dev. 0.5840	Min. -0.9991 1st Qu. -0.5118 Median 0.0496 Mean 0.0143 3rd Qu. 0.5035 Max. 0.9993 Std.Dev. 0.5628
	x5	x6	x7	x8	x9	x10	x11
x11==1	Min. -0.9999 1st Qu. -0.4353 Median 0.0102 Mean 0.0232 3rd Qu. 0.5112 Max. 0.9991 Std.Dev. 0.5628	Min. -0.9993 1st Qu. -0.5000 Median -0.0660 Mean -0.0234 3rd Qu. 0.4898 Max. 0.9975 Std.Dev. 0.5692	Min. -0.9969 1st Qu. -0.5121 Median -0.0392 Mean -0.0102 3rd Qu. 0.4577 Max. 0.9998 Std.Dev. 0.5691	Min. -0.9912 1st Qu. -0.5678 Median -0.0384 Mean -0.0317 3rd Qu. 0.5181 Max. 0.9954 Std.Dev. 0.5873	Min. -0.9939 1st Qu. -0.4752 Median -0.0026 Mean 0.0043 3rd Qu. 0.5057 Max. 0.9985 Std.Dev. 0.5604	Min. -0.9983 1st Qu. -0.5117 Median -0.0619 Mean 0.0043 3rd Qu. 0.5114 Max. 0.9899 Std.Dev. 0.5740	Min. -0.9983 1st Qu. -0.5117 Median -0.0619 Mean 0.0043 3rd Qu. 0.5114 Max. 0.9899 Std.Dev. 0.5740

Table 85: Treatment 테이터 요약

25.2.8 Hormone Treatment Recommendations for Breast Cancer(GBSG)

데이터 소개 작성중...

25.2.8.1 변수 요약

Instance 수는 686개 있으며, 총 9개의 변수가 있다. 각 변수의 정확한 description은 아직 파악하지 못하였다.

- **t**, a numeric vector
 - 0
 - 1
- **e**, a binary vector
 - 0
 - 1
- **x1**, a binary vector
 - 0
 - 1
 - 2
- **x2**, a categorical vector
 - 0
 - 1
 - 2
- **x3**, a binary vector
 - 0
 - 1
- **x4**, a numeric vector
- **x5**, a numeric vector
- **x6**, a numeric vector
- **x7**, a numeric vector

	t	e	x1	x2	x3	x4	x5	x6	x7
All	Min. 0.2628 1st Qu. 18.6530 Median 35.6140 Mean 36.9442 3rd Qu. 55.3511 Max. 87.3593 Std.Dev. 21.1184	0: 387 1: 299	0: 440 1: 246	0: 180 1: 453 2: 53	0: 290 1: 396	Min. 21.0000 1st Qu. 46.0000 Median 53.0000 Mean 53.0500 3rd Qu. 61.0000 Max. 81.0000 Std.Dev. 10.1207	Min. 1.0000 1st Qu. 1.0000 Median 3.0000 Mean 5.0100 3rd Qu. 7.0000 Max. 51.0000 Std.Dev. 5.4755	Min. 0.0000 1st Qu. 7.0000 Median 32.5000 Mean 110.0000 3rd Qu. 131.8000 Max. 2380.0000 Std.Dev. 202.3316	Min. 0.0000 1st Qu. 8.0000 Median 36.0000 Mean 96.2500 3rd Qu. 114.0000 Max. 1144.0000 Std.Dev. 153.0840
e==0	Min. 0.2628 1st Qu. 29.2402 Median 47.4086 Mean 45.2376 3rd Qu. 60.7639 Max. 87.3593 Std.Dev. 20.6001	0: 387 1: 152	0: 235 1: 115 2: 23	0: 115 1: 249 2: 216	0: 171 1: 216	Min. 31.0000 1st Qu. 46.0000 Median 52.0000 Mean 53.0900 3rd Qu. 61.0000 Max. 80.0000 Std.Dev. 9.5231	Min. 1.0000 1st Qu. 1.0000 Median 2.0000 Mean 3.8450 3rd Qu. 5.0000 Max. 51.0000 Std.Dev. 4.5839	Min. 0.0000 1st Qu. 11.5000 Median 56.0000 Mean 140.5000 3rd Qu. 167.0000 Max. 2380.0000 Std.Dev. 242.8588	Min. 0.0000 1st Qu. 11.0000 Median 42.0000 Mean 104.5000 3rd Qu. 129.0000 Max. 1144.0000 Std.Dev. 153.0770
e==1	Min. 2.3650 1st Qu. 13.9960 Median 21.2240 Mean 26.2100 3rd Qu. 36.1230 Max. 80.6900 Std.Dev. 16.4488	0: 299 1: 299	0: 205 1: 94	0: 65 1: 204 2: 30	0: 119 1: 180	Min. 21.0000 1st Qu. 46.0000 Median 54.0000 Mean 53.0000 3rd Qu. 61.5000 Max. 80.0000 Std.Dev. 10.8617	Min. 1.0000 1st Qu. 2.0000 Median 5.0000 Mean 6.5180 3rd Qu. 9.0000 Max. 36.0000 Std.Dev. 6.1362	Min. 0.0000 1st Qu. 2.5000 Median 19.0000 Mean 70.53000 3rd Qu. 80.5000 Max. 912.0000 Std.Dev. 122.2060	Min. 0.0000 1st Qu. 4.0000 Median 28.0000 Mean 85.5400 3rd Qu. 109.0000 Max. 1091.0000 Std.Dev. 152.6825
x1==0	Min. 0.2628 1st Qu. 17.9959 Median 31.7700 Mean 34.8166 3rd Qu. 51.6797 Max. 84.2053 Std.Dev. 20.3288	0: 235 1: 205	0: 440 1: 113	0: 113 1: 288 2: 39	0: 231 1: 209	Min. 21.0000 1st Qu. 45.0000 Median 50.0000 Mean 51.0600 3rd Qu. 59.0000 Max. 80.0000 Std.Dev. 5.5581	Min. 1.0000 1st Qu. 1.0000 Median 3.0000 Mean 4.9430 3rd Qu. 7.0000 Max. 51.0000 Std.Dev. 6.1362	Min. 0.0000 1st Qu. 7.0000 Median 32.0000 Mean 102.0000 3rd Qu. 130.0000 Max. 1600.0000 Std.Dev. 170.0113	Min. 0.0000 1st Qu. 8.0000 Median 32.0000 Mean 79.7200 3rd Qu. 92.2500 Max. 898.0000 Std.Dev. 124.2161
x1==1	Min. 0.4928 1st Qu. 22.8583 Median 40.0986 Mean 40.7498 3rd Qu. 59.7289 Max. 87.3593 Std.Dev. 21.9949	0: 152 1: 94	0: 246 1: 246	0: 67 1: 165 2: 14	0: 59 1: 187	Min. 32.0000 1st Qu. 50.0000 Median 58.0000 Mean 56.6200 3rd Qu. 63.0000 Max. 80.0000 Std.Dev. 5.3337	Min. 1.0000 1st Qu. 1.0000 Median 3.0000 Mean 5.1300 3rd Qu. 7.0000 Max. 36.0000 Std.Dev. 5.3337	Min. 0.0000 1st Qu. 7.2500 Median 35.0000 Mean 124.2900 3rd Qu. 133.0000 Max. 2380.0000 Std.Dev. 249.6968	Min. 0.0000 1st Qu. 9.0000 Median 46.0000 Mean 125.8000 3rd Qu. 182.5000 Max. 1144.0000 Std.Dev. 191.0648
x2==0	Min. 0.5585 1st Qu. 21.5441 Median 43.8275 Mean 41.3655 3rd Qu. 61.1581 Max. 80.4600 Std.Dev. 22.6660	0: 115 1: 65	0: 113 1: 67	0: 180 1: 78 2: 102	0: 78 1: 102	Min. 21.0000 1st Qu. 46.0000 Median 53.0000 Mean 53.1500 3rd Qu. 62.0000 Max. 80.0000 Std.Dev. 10.8820	Min. 1.0000 1st Qu. 1.0000 Median 2.00000 Mean 3.4890 3rd Qu. 5.0000 Max. 33.0000 Std.Dev. 3.6973	Min. 0.0000 1st Qu. 10.0000 Median 55.5000 Mean 120.7000 3rd Qu. 137.5000 Max. 2380.0000 Std.Dev. 236.5264	Min. 0.0000 1st Qu. 8.0000 Median 35.0000 Mean 101.0000 3rd Qu. 120.5000 Max. 1060.0000 Std.Dev. 165.3625
x2==1	Min. 0.2628 1st Qu. 18.7269 Median 34.2669 Mean 36.0345 3rd Qu. 53.1253 Max. 87.3593 Std.Dev. 20.3435	0: 249 1: 204	0: 288 1: 165	0: 453 1: 269	0: 184 1: 269	Min. 25.0000 1st Qu. 46.0000 Median 53.0000 Mean 53.3300 3rd Qu. 61.0000 Max. 80.0000 Std.Dev. 9.8854	Min. 1.0000 1st Qu. 2.0000 Median 3.0000 Mean 5.0004 3rd Qu. 7.0000 Max. 38.0000 Std.Dev. 5.0825	Min. 0.0000 1st Qu. 7.0000 Median 31.0000 Mean 107.2000 3rd Qu. 131.0000 Max. 1600.0000 Std.Dev. 190.5852	Min. 0.0000 1st Qu. 8.0000 Median 38.0000 Mean 99.3100 3rd Qu. 121.0000 Max. 1144.0000 Std.Dev. 153.5548
x2==2	Min. 0.4928 1st Qu. 14.7187 Median 25.4949 Mean 29.7039 3rd Qu. 44.1232 Max. 76.2218 Std.Dev. 19.5160	0: 23 1: 30	0: 39 1: 14	0: 28 1: 25 2: 53	0: 28 1: 25	Min. 31.0000 1st Qu. 44.0000 Median 49.0000 Mean 50.3400 3rd Qu. 57.0000 Max. 70.0000 Std.Dev. 9.1671	Min. 1.0000 1st Qu. 4.0000 Median 8.0000 Mean 10.2300 3rd Qu. 15.0000 Max. 51.0000 Std.Dev. 9.3596	Min. 0.0000 1st Qu. 3.0000 Median 22.0000 Mean 98.0600 3rd Qu. 84.0000 Max. 796.0000 Std.Dev. 174.2503	Min. 0.0000 1st Qu. 5.0000 Median 28.0000 Mean 54.0200 3rd Qu. 76.0000 Max. 569.0000 Std.Dev. 87.4316
x3==0	Min. 0.2628 1st Qu. 17.6427 Median 35.5647 Mean 36.3348 3rd Qu. 56.3860 Max. 84.2053 Std.Dev. 21.8817	0: 171 1: 119	0: 231 1: 59	0: 78 1: 184 2: 28	0: 290 1:	Min. 21.0000 1st Qu. 40.2500 Median 45.0000 Mean 43.9500 3rd Qu. 48.0000 Max. 60.0000 Std.Dev. 6.1967	Min. 1.0000 1st Qu. 1.0000 Median 3.0000 Mean 4.8620 3rd Qu. 6.7500 Max. 38.0000 Std.Dev. 5.3184	Min. 0.0000 1st Qu. 8.2500 Median 38.0000 Mean 105.5900 3rd Qu. 137.7500 Max. 1600.0000 Std.Dev. 172.9402	Min. 0.0000 1st Qu. 5.0000 Median 25.0000 Mean 49.9800 3rd Qu. 62.0000 Max. 628.0000 Std.Dev. 80.0824
x3==1	Min. 0.4928 1st Qu. 20.4025 Median 35.6304 Mean 37.3905 3rd Qu. 55.0308 Max. 87.3593 Std.Dev. 20.5580	0: 216 1: 180	0: 209 1: 187	0: 102 1: 269 2: 25	0: 396 1:	Min. 42.0000 1st Qu. 55.0000 Median 60.0000 Mean 59.7200 3rd Qu. 64.0000 Max. 80.0000 Std.Dev. 6.6482	Min. 1.0000 1st Qu. 1.0000 Median 3.0000 Mean 5.1190 3rd Qu. 7.0000 Max. 51.0000 Std.Dev. 5.5919	Min. 0.0000 1st Qu. 6.0000 Median 29.5000 Mean 113.2000 3rd Qu. 120.5000 Max. 1144.0000 Std.Dev. 221.5556	Min. 0.0000 1st Qu. 9.0000 Median 58.5000 Mean 130.1000 3rd Qu. 191.0000 Max. 1144.0000 Std.Dev. 182.2759

Table 86: Metabric 데이터 요약

25.2.9 Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUP-PORT)

데이터 소개 작성중...

25.2.9.1 변수 요약

Instance 수는 1775개 있으며, 총 14개의 변수가 있다. 각 변수의 정확한 description은 아직 파악하지 못하였다.

- `t`, a numeric vector
 - 0
 - 1
- `hr`, a numeric vector
- `x1`, a numeric vector
- `x2`, a binary vector
 - 0
 - 1
- `x3`, a categorical vector
 - 0
 - 1
 - 2
 - 3
 - 4
 - 5
 - 6
 - 7
- `x4`, a categorical vector
 - 0
 - 1
 - 2
 - 3
 - 4
 - 5
- `x5`, a binary vector
 - 0
 - 1
- `x6`, a binary vector
 - 0
 - 1
- `x7`, a categorical vector
 - 0
 - 1
 - 2
- `x8`, a numeric vector
- `x9`, a numeric vector
- `x10`, a numeric vector
- `x11`, a numeric vector
- `x12`, a numeric vector
- `x13`, a numeric vector
- `x14`, a numeric vector

	t	e	x1	x2	x3	x4	x5	x6
All	Min.	3.0000		Min.	18.1900		0:	
	1st Qu.	25.0000		1st Qu.	52.5100		1:	230
	Median	237.0000	0: 562	Median	64.7200	0: 1004	0:	18
	Mean	486.1000	1: 1213	Mean	62.1400	1: 771	1:	564
	3rd Qu.	790.5000		3rd Qu.	73.7000		2:	493
	Max.	2029.0000		Max.	101.8500		3:	1381
	Std.Dev.	569.4312		Std.Dev.	15.8495		4:	283
							5:	53
							6:	19
							7:	7
	x7	x8	x9	x10	x11	x12	x13	x14
		Min. 0.0000	Min. 0.0000	Min. 0.0000	Min. 31.7000	Min. 115.0000	Min. 0.0000	Min. 0.2000
	0: 383	1st Qu. 64.0000	1st Qu. 72.0000	1st Qu. 18.0000	1st Qu. 36.2000	1st Qu. 134.0000	1st Qu. 7.1000	1st Qu. 0.8999
	1: 1142	Median 78.0000	Median 100.0000	Median 24.0000	Median 36.8000	Median 137.0000	Median 10.6000	Median 1.2000
	2: 250	Mean 85.0900	Mean 97.0100	Mean 23.4100	Mean 37.1100	Mean 137.5000	Mean 12.0600	Mean 1.7311
		3rd Qu. 108.0000	3rd Qu. 118.5000	3rd Qu. 28.0000	3rd Qu. 38.2000	3rd Qu. 141.0000	3rd Qu. 14.8000	3rd Qu. 1.7998
		Max. 193.0000	Max. 250.0000	Max. 90.0000	Max. 41.2000	Max. 164.0000	Max. 10.0000	Max. 14.6992
		Std.Dev. 27.6333	Std.Dev. 31.3924	Std.Dev. 9.7759	Std.Dev. 1.2401	Std.Dev. 5.7964	Std.Dev. 100.0000	Std.Dev. 1.6241
	x7	x8	x9	x10	x11	x12	x13	x14
e==0	Min. 348.0000		Min. 18.1900		0: 133			
	1st Qu. 637.2000		1st Qu. 44.4300		1:	143	0:	
	Median 948.5000	0: 562	Median 59.4000	0: 308	2:	146	1:	421
	Mean 1086.5000	1:	Mean 57.5600	1: 254	3:	85	2:	112
	3rd Qu. 1551.8000		3rd Qu. 71.2200		4:	37	3:	21
	Max. 2029.0000		Max. 101.8500		5:	14	4:	4
	Std.Dev. 520.0974		Std.Dev. 17.2902		6:	4	5:	1
					7:			
	x7	x8	x9	x10	x11	x12	x13	x14
		Min. 0.0000	Min. 0.0000	Min. 0.0000	Min. 31.7000	Min. 117.0000	Min. 0.0900	Min. 0.3000
e==1	0: 40	1st Qu. 67.0000	1st Qu. 70.0000	1st Qu. 18.0000	1st Qu. 36.2000	1st Qu. 134.0000	1st Qu. 7.5250	1st Qu. 0.7999
	1: 457	Median 80.0000	Median 100.0000	Median 24.0000	Median 36.8000	Median 137.0000	Median 11.1000	Median 1.2000
	2: 65	Mean 87.0000	Mean 97.0100	Mean 23.3000	Mean 37.2300	Mean 137.4000	Mean 12.1780	Mean 1.6335
		3rd Qu. 108.0000	3rd Qu. 118.7500	3rd Qu. 28.0000	3rd Qu. 38.3000	3rd Qu. 141.0000	3rd Qu. 15.2490	3rd Qu. 1.7000
		Max. 163.0000	Max. 250.0000	Max. 67.0000	Max. 40.3000	Max. 155.0000	Max. 80.0000	Max. 14.6992
		Std.Dev. 26.8324	Std.Dev. 31.4478	Std.Dev. 9.3832	Std.Dev. 1.2434	Std.Dev. 5.4777	Std.Dev. 7.1805	Std.Dev. 1.5227
	x7	x8	x9	x10	x11	x12	x13	x14
		Min. 3.0000	Min. 19.4400		0: 97			
	1st Qu. 13.0000	1st Qu. 55.2900		1:	421	0:	15	
x2==0	Median 60.0000	0: 562	Median 66.5700	0: 696	2:	347	1:	960
	Mean 207.8000	1: 1213	Mean 64.2700	1: 517	3:	198	2:	186
	3rd Qu. 253.0000		3rd Qu. 74.8600		4:	101	3:	32
	Max. 1886.0000		Max. 95.7100		5:	32	4:	14
	Std.Dev. 323.6213		Std.Dev. 14.6646		6:	15	5:	6
					7:	2		
	x7	x8	x9	x10	x11	x12	x13	x14
		Min. 0.0000	Min. 0.0000	Min. 0.0000	Min. 32.3000	Min. 115.0000	Min. 0.0000	Min. 0.2000
	0: 343	1st Qu. 63.0000	1st Qu. 72.0000	1st Qu. 18.0000	1st Qu. 36.0900	1st Qu. 134.0000	1st Qu. 6.8990	1st Qu. 0.8999
	1: 685	Median 77.0000	Median 100.0000	Median 24.0000	Median 36.7000	Median 137.0000	Median 10.3980	Median 1.2000
x2==1	2: 185	Mean 84.0000	Mean 97.0100	Mean 23.4600	Mean 37.0500	Mean 137.6000	Mean 12.0110	Mean 1.7763
		3rd Qu. 108.0000	3rd Qu. 118.0000	3rd Qu. 28.0000	3rd Qu. 38.0900	3rd Qu. 141.0000	3rd Qu. 14.6990	3rd Qu. 1.8999
		Max. 232.0000	Max. 232.0000	Max. 90.0000	Max. 41.2000	Max. 164.0000	Max. 100.0000	Max. 13.7988
		Std.Dev. 27.9400	Std.Dev. 31.3797	Std.Dev. 9.9561	Std.Dev. 1.2344	Std.Dev. 5.9398	Std.Dev. 8.4042	Std.Dev. 1.6677
	x7	x8	x9	x10	x11	x12	x13	x14
		Min. 3.0000	Min. 18.8900		0: 127			
	1st Qu. 24.7500	1st Qu. 51.7800		1:	300	0:	10	
	Median 249.0000	0: 308	Median 64.1300	0: 1004	2:	291	1:	806
	Mean 486.0600	1: 696	Mean 61.6200	1:	3:	161	2:	149
x2==1	3rd Qu. 778.5000		3rd Qu. 73.1200		4:	79	3:	22
	Max. 2024.0000		Max. 95.5800		5:	29	4:	12
	Std.Dev. 567.0851		Std.Dev. 15.5687		6:	15	5:	5
					7:	2		
	x7	x8	x9	x10	x11	x12	x13	x14
		Min. 0.0000	Min. 0.0000	Min. 0.0000	Min. 33.4000	Min. 117.0000	Min. 0.0000	Min. 0.3000
	0: 217	1st Qu. 64.0000	1st Qu. 72.0000	1st Qu. 18.0000	1st Qu. 36.0900	1st Qu. 134.0000	1st Qu. 7.0000	1st Qu. 0.8999
	1: 645	Median 78.0000	Median 100.0000	Median 24.0000	Median 36.7000	Median 137.0000	Median 10.7000	Median 1.2000
	2: 142	Mean 85.1500	Mean 97.0800	Mean 23.5000	Mean 37.0800	Mean 137.4000	Mean 11.9800	Mean 1.7369
		3rd Qu. 108.0000	3rd Qu. 118.0000	3rd Qu. 28.0000	3rd Qu. 38.2000	3rd Qu. 141.0000	3rd Qu. 14.7000	3rd Qu. 1.7998
x2==1		Max. 163.0000	Max. 232.0000	Max. 67.0000	Max. 41.2000	Max. 160.0000	Max. 100.0000	Max. 14.6992
		Std.Dev. 27.1223	Std.Dev. 30.9628	Std.Dev. 9.4493	Std.Dev. 1.2386	Std.Dev. 5.7010	Std.Dev. 8.2063	Std.Dev. 1.6074
	x7	x8	x9	x10	x11	x12	x13	x14
		Min. 3.0000	Min. 18.1900		0: 103			
	1st Qu. 25.5000	1st Qu. 53.0900		1:	264	0:	8	
	Median 223.0000	0: 254	Median 65.7200	0:	202	1:	575	
	Mean 486.1000	1: 517	Mean 62.8300	1: 771	3:	122	2:	149
	3rd Qu. 802.0000		3rd Qu. 74.5000		4:	59	3:	31
	Max. 2029.0000		Max. 101.8500		5:	17	4:	6
	Std.Dev. 572.8405		Std.Dev. 16.1923		6:	4	5:	2
					7:			
	x7	x8	x9	x10	x11	x12	x13	x14
		Min. 0.0000	Min. 0.0000	Min. 0.0000	Min. 31.7000	Min. 115.0000	Min. 0.0500	Min. 0.2000
	0: 166	1st Qu. 63.0000	1st Qu. 71.0000	1st Qu. 16.0000	1st Qu. 36.2000	1st Qu. 134.0000	1st Qu. 7.1000	1st Qu. 0.7999
	1: 497	Median 77.0000	Median 100.0000	Median 24.0000	Median 36.8000	Median 137.0000	Median 10.5000	Median 1.0999
	2: 108	Mean 85.0200	Mean 96.9100	Mean 23.2900	Mean 37.1400	Mean 137.8000	Mean 12.1800	Mean 1.7235
		3rd Qu. 108.0000	3rd Qu. 120.0000	3rd Qu. 28.0000	3rd Qu. 38.2000	3rd Qu. 141.0000	3rd Qu. 15.0000	3rd Qu. 1.7998
		Max. 193.0000	Max. 250.0000	Max. 90.0000	Max. 40.8000	Max. 164.0000	Max. 67.3000	Max. 12.8984
		Std.Dev. 28.3024	Std.Dev. 31.9632	Std.Dev. 10.1905	Std.Dev. 1.2420	Std.Dev. 5.9141	Std.Dev. 7.8109	Std.Dev. 1.6466

Table 87: Support 테이터 요약

	t	e	x1	x2	x3	x4	x5	x6
x5==0	Min.	3.0000		Min.	18.1900		0:	
	1st Qu.	25.5000		1st Qu.	53.0900		1:	264
	Median	223.0000	0: 254	Median	65.7200	0:	2:	202
	Mean	486.1000	1: 517	Mean	62.8300	1: 771	1:	575
	3rd Qu.	802.0000		3rd Qu.	74.5000	3:	122	
	Max.	2029.0000		Max.	101.8500	4:	59	0: 586
	Std.Dev.	572.8405		Std.Dev.	16.1923	5:	17	1: 742
						6:	4	0: 29
						7:	2	
	x7	x8	x9	x10	x11	x12	x13	x14
x5==1								
	Min.	0.0000		Min.	0.0000		Min.	0.0500
	1st Qu.	63.0000		1st Qu.	71.0000		1st Qu.	0.2000
	Median	77.0000	0: 166	Median	100.0000	1st Qu.	0.7999	
	Mean	85.0200	1: 497	Mean	96.9100	Median	137.0000	
	3rd Qu.	108.0000	2: 108	3rd Qu.	120.0000	Mean	10.5000	
	Max.	193.0000		Max.	250.0000	3rd Qu.	12.1800	
	Std.Dev.	28.3024		Std.Dev.	50.0000	3rd Qu.	14.0000	
				Std.Dev.	10.1905	3rd Qu.	15.0000	
					Std.Dev.	1.2420	Max.	16.4000
						Std.Dev.	5.9141	Max.
x6==0								
	Min.	3.0000		Min.	18.1900		Min.	
	1st Qu.	24.0000		1st Qu.	50.7900	0:		
	Median	215.0000	0: 455	Median	63.7500	1:	228	
	Mean	484.7000	1: 959	Mean	61.3800	1:	518	
	3rd Qu.	785.0000		3rd Qu.	73.3400	2:	387	
	Max.	2029.0000		Max.	101.8500	3:	178	
	Std.Dev.	578.3497		Std.Dev.	16.2897	4:	73	
						5:	41	0: 1414
						6:	11	1: 1372
x6==1								
	Min.	0.0000		Min.	0.0000		Min.	
	1st Qu.	64.0000		1st Qu.	72.0000		1st Qu.	
	Median	78.0000	0: 340	Median	100.0000		Median	
	Mean	85.4000	1: 876	Mean	98.2700		Mean	
	3rd Qu.	108.0000	2: 198	3rd Qu.	120.0000		3rd Qu.	
	Max.	193.0000		Max.	232.0000		Max.	
	Std.Dev.	27.3415		Std.Dev.	31.2211		Std.Dev.	
						Std.Dev.	9.6655	Std.Dev.
							1.2459	5.8005
x6==2								
	Min.	3.0000		Min.	18.1900		Min.	
	1st Qu.	27.0000		1st Qu.	52.0200	0:		
	Median	245.0000	0: 552	Median	64.1100	1:	230	
	Mean	492.5000	1: 1165	Mean	61.5900	2:	556	
	3rd Qu.	803.0000		3rd Qu.	73.0500	3:	476	
	Max.	2029.0000		Max.	101.8500	4:	1336	
	Std.Dev.	571.5203		Std.Dev.	15.7135	5:		
						6:	270	0: 1414
						7:	53	1: 1372
x6==3								
	Min.	0.0000		Min.	0.0000		Min.	
	1st Qu.	64.0000		1st Qu.	72.0000		1st Qu.	
	Median	78.0000	0: 382	Median	100.0000		Median	
	Mean	85.1500	1: 1099	Mean	96.9999		Mean	
	3rd Qu.	108.0000	2: 236	3rd Qu.	119.0000		3rd Qu.	
	Max.	193.0000		Max.	250.0000		Max.	
	Std.Dev.	27.4641		Std.Dev.	31.3393		Std.Dev.	
						Std.Dev.	9.7356	Std.Dev.
							1.2408	5.8104
x6==4								
	Min.	3.0000		Min.	47.1100		Min.	
	1st Qu.	10.2500		1st Qu.	72.3400	0:		
	Median	34.5000	0: 10	Median	80.5700	1:	8	
	Mean	295.4300	1: 48	Mean	78.3800	2:	45	
	3rd Qu.	405.2500		3rd Qu.	85.0000	3:	17	
	Max.	1848.0000		Max.	95.7100	4:	12	
	Std.Dev.	468.9968		Std.Dev.	10.3610	5:	2	
						6:	1	0: 16
						7:	5	1: 58
x6==5								
	Min.	0.0000		Min.	36.0000		Min.	
	1st Qu.	57.7500		1st Qu.	60.0000		1st Qu.	
	Median	72.5000	0: 1	Median	68.5000		Median	
	Mean	83.4500	1: 43	Mean	97.4700		Mean	
	3rd Qu.	113.0000	2: 14	3rd Qu.	118.0000		3rd Qu.	
	Max.	163.0000		Max.	160.0000		Max.	
	Std.Dev.	32.4773		Std.Dev.	33.2087		Std.Dev.	
						Std.Dev.	10.9859	Std.Dev.
							1.2012	5.1202
x7==0								
	Min.	3.0000		Min.	20.7700		Min.	
	1st Qu.	23.0000		1st Qu.	54.8200	0:		
	Median	118.0000	0: 40	Median	64.5700	1:	229	
	Mean	262.0000	1: 343	Mean	62.3700	2:	94	
	3rd Qu.	336.0000		3rd Qu.	71.0100	3:	318	
	Max.	1988.0000		Max.	86.1900	4:		
	Std.Dev.	382.7119		Std.Dev.	11.6476	5:	38	
						6:	4	0: 340
						7:	2	1: 382
x7==1								
	Min.	0.0000		Min.	36.0000		Min.	
	1st Qu.	67.0000		1st Qu.	72.0000		1st Qu.	
	Median	83.0000	0: 383	Median	96.0100		Median	
	Mean	87.6800	1: 1142	Mean	22.8400		Mean	
	3rd Qu.	108.0000	2: 1142	3rd Qu.	116.0000		3rd Qu.	
	Max.	163.0000		Max.	181.0000		Max.	
	Std.Dev.	26.8341		Std.Dev.	29.4621		Std.Dev.	
						Std.Dev.	8.6631	Std.Dev.
							1.1307	5.4054
x7==2								
	Min.	3.0000		Min.	18.1900		Min.	
	1st Qu.	29.2500		1st Qu.	50.2000	0:		
	Median	403.5000	0: 457	Median	64.0400	1:	317	
	Mean	577.4000	1: 685	Mean	61.4900	2:	855	
	3rd Qu.	916.7500		3rd Qu.	74.5100	3:	223	
	Max.	2029.0000		Max.	101.8500	4:	37	
	Std.Dev.	604.7352		Std.Dev.	16.8386	5:	42	
						6:	12	0: 876
						7:	5	1: 1099
x7==3								
	Min.	0.0000		Min.	0.0000		Min.	
	1st Qu.	70.0000		1st Qu.	70.0000		1st Qu.	
	Median	100.0000	0: 65	Median	100.0000		Median	
	Mean	96.6400	1: 185	Mean	23.4400		Mean	
	3rd Qu.	118.0000	2: 250	3rd Qu.	118.0000		3rd Qu.	
	Max.	250.0000		Max.	90.0000		Max.	
	Std.Dev.	27.6198		Std.Dev.	31.6238		Std.Dev.	
						Std.Dev.	9.8839	Std.Dev.
							1.2579	5.8834
x7==4								
	Min.	3.0000		Min.	19.4400		Min.	
	1st Qu.	17.0000		1st Qu.	56.5400	0:		
	Median	101.5000	0: 65	Median	68.0800	1:	142	
	Mean	411.7000	1: 185	Mean	64.7600	2:	108	
	3rd Qu.	700.5000		3rd Qu.	76.7300	3:	45	
	Max.	2026.0000		Max.	95.6100	4:	33	
	Std.Dev.	537.3628		Std.Dev.	16.5345	5:	10	
						6:	7	0: 198
						7:	1	1: 236
x7==5								
	Min.	0.0000		Min.	0.0000		Min.	
	1st Qu.	63.0000		1st Qu.	72.0000		1st Qu.	
	Median	78.0000	0: 250	Median	105.0000		Median	
	Mean	85.3300	1: 250	Mean	100.2000		Mean	
	3rd Qu.	109.0000		3rd Qu.	120.0000		3rd Qu.	
	Max.	160.0000		Max.	210.0000		Max.	
	Std.Dev.	28.7349		Std.Dev.	33.0779		Std.Dev.	
						Std.Dev.	10.8187	Std.Dev.
							1.3132	5.9908
								Std.Dev.

Table 88: Support 245] \Rightarrow 요약 (Cont'd)

25.2.10 BLCA: Bladder Urothelia Carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)

Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 322 2014. Broad Institute of MIT and Harvard, doi:10.7908/C1DN43V9 (2014).

<https://gdac.broadinstitute.org/>

여기에서 RNA-Seq expression and clinical data를 활용하였다.

- Clinical Data: Clinical_Pick_Tier1 (MD5)
- mRNASeq: illuminahisseq_rnaseqv2-RSEM_genes_normalized (MD5)
- miRSeq: illuminahisseq_mirnaseq-miR_gene_expression (MD5)

25.2.10.1 변수 요약

Instance 수는 412명이며, 각 데이터는 다음과 같이 구성되어 있다. 분석할 때는 이들을 머지하여야 한다.

- clinical data는 20개의 변수
- mRNASeq data는 428명의 환자에 대한 20533개의 변수
- miRSeq data는 1288명의 환자에 대한 1048개의 변수

mRNASeq와 miRSeq의 환자 수가 clinical data와 차이가 있으며, repeated measured case도 있음을 유의할 필요가 있다. 또한 데이터에 missing value도 다수 존재하는 편이다.

- Composite
 - Value: (모든 값이 Value이므로 의미가 없다.)
- years_to_birth, a numeric vector
- vital_status
 - 0
 - 1
- days_to_death, a numeric vector
- days_to_last_followup, a numeric vector
- tumor_tissue_site
 - bladder: (모든 값이 bladder이다.)
- pathologic_stage
 - stage i
 - stage ii
 - stage iii
 - stage iv
- pathology_T_stage
 - t0
 - t1
 - t2
 - t2a
 - t2a
 - t3
 - t3a
 - t3b
 - t4
 - t4a
 - t4b
 - tx
- pathology_N_stage
 - n0
 - n1
 - n2
 - n3
 - nx

- pathology_M_stage
 - m0
 - m1
 - mx
- gender
 - female
 - male
- date_of_initial_pathologic_diagnosis, a numeric vector
- days_to_last_known_alive
 - NA: (모든 값이 NA이다.)
- radiation_therapy
 - no
 - yes
- karnofsky_performance_score, a numeric vector
- histological_type
 - muscle invasive urothelial carcinoma (pt2 or above)
- number_pack_years_smoked, a numeric vector
- number_of_lymph_nodes, a numeric vector
- race
 - asian
 - black or african american
 - white
- ethnicity
 - hispanic or latino
 - not hispanic or latino

Composite Element REF		years_to_birth	vital_status	days_to_death	days_to_last_followup	
All	Value: 412	Min. 34.0000 1st Qu. 60.0000 Median 69.0000 Mean 68.0800 3rd Qu. 76.0000 Max. 90.0000 Std.Dev. 10.6025 NA 1	0: 230 1: 182	Min. 19.0000 1st Qu. 239 Median 47 Mean 76 3rd Qu. 8 Max. 36 Std.Dev. NA NA 182	Min. -64.0000 1st Qu. 398.2000 Median 639.0000 Mean 1016.7000 3rd Qu. 1458.8000 Max. 5050.0000 Std.Dev. 531.9471 NA 182	
		tumor_tissue_site	pathologic_stage	pathology_T_stage	pathology_N_stage	pathology_M_stage
	bladder: 412	stage i: 2 stage ii: 131 stage iii: 141 stage iv: 136 NA 2	t0: 1 t1: 3 t2: 38 t2a: 26 t2b: 56 t3: 43 t3a: 71	t3b: 82 t4: 11 t4a: 43 t4b: 5 tx: 1 NA 32	n0: 239 n1: 47 n2: 76 n3: 8 nx: 36 NA 6	m0: 196 m1: 11 mx: 202 NA 3
gender	date_of_initial_pathologic_diagnosis	days_to_last_known_alive	radiation_therapy	karnofsky_performance_score		
female: 108 male: 304	Min. 1999.0000 1st Qu. 2009.0000 Median 2011.0000 Mean 2010.0000 3rd Qu. 2012.0000 Max. 2013.0000 Std.Dev. 961.5995 NA 18	Min. 1st Qu. Median Mean 3rd Qu. Max. Std.Dev. NA 412	no: 366 yes: 20 NA 26	Min. 30.0000 1st Qu. 80.0000 Median 90.0000 Mean 83.0900 3rd Qu. 90.0000 Max. 100.0000 Std.Dev. 13.6905 NA 276		
histological_type	number_pack_years_smoked	number_of_lymph_nodes	race	ethnicity		
muscle invasive urothelial carcinoma (pt2 or above): 409 NA 3	Min. 0.1500 1st Qu. 20.0000 Median 30.0000 Mean 39.0400 3rd Qu. 50.0000 Max. 730.0000 Std.Dev. 52.9262 NA 188	Min. 0.0000 1st Qu. 0.0000 Median 0.0000 Mean 2.0880 3rd Qu. 2.0000 Max. 97.0000 Std.Dev. 7.0062 NA 115	asian: 44 black or african american : 23 white: 527 NA 18	hispanic or latino 9 not hispanic or latino : 371 NA 32		

Table 89: BLCA 데이터의 clinical part 요약

25.2.11 BRCA: Breast invasive carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)

Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 322 2014. Broad Institute of MIT and Harvard, doi:10.7908/C1DN43V9 (2014).

<https://gdac.broadinstitute.org/>

여기에서 RNA-Seq expression and clinical data를 활용하였다.

- Clinical Data: Clinical_Pick_Tier1 (MD5)
- mRNASeq: illuminahisq_rnaseqv2-RSEM_isoforms_normalized (MD5)
- miRSeq: illuminahisq_mirnaseq-miR_gene_expression (MD5)

25.2.11.1 변수 요약

Instance 수는 1097명이며, 각 데이터는 다음과 같이 구성되어 있다. 분석할 때는 이들을 머지하여여 한다.

- clinical data는 18개의 변수
- mRNASeq data는 1213명의 환자에 대한 73601개의 변수
- miRSeq data는 1287명의 환자에 대한 1048개의 변수

mRNASeq와 miRSeq의 환자 수가 clinical data와 차이가 있으며, repeated measured case도 있음을 유의할 필요가 있다. 또한 데이터에 missing value도 다수 존재하는 편이다.

- Composite
 - Value: (모든 값이 Value이므로 의미가 없다.)

- years_to_birth, a numeric vector

- vital_status
 - 0
 - 1

- days_to_death, a numeric vector

- days_to_last_followup, a numeric vector

- tumor_tissue_site
 - breast: (모든 값이 breast이다.)

- pathologic_stage
 - stage i
 - stage ia
 - stage ib
 - stage ii
 - stage iia
 - stage iib
 - stage iii
 - stage iiia
 - stage iiib
 - stage iiic
 - stage iv
 - stage x

- pathology_T_stage
 - t1
 - t1a
 - t1b
 - t1c
 - t2
 - t2a
 - t2a
 - t3
 - t3a
 - t4
 - t4b
 - t4d
 - tx

- pathology_N_stage
 - n0
 - n1
 - n2
 - n3
 - nx
- pathology_M_stage
 - m0
 - m1
 - mx
- gender
 - female
 - male
- date_of_initial_pathologic_diagnosis, a numeric vector
- days_to_last_known_alive, a numeric vector
- radiation_therapy
 - no
 - yes
- histological_type
 - infiltrating carcinoma nos
 - infiltrating ductal carcinoma
 - infiltrating lobular carcinoma
 - medullary carcinoma
 - metaplastic carcinoma
 - mixed histology (please specify)
 - mucinous carcinoma
 - other, specify
- number_of_lymph_nodes, a numeric vector
- race
 - american indian or alaska native
 - asian
 - black or african american
 - white
- ethnicity
 - hispanic or latino
 - not hispanic or latino

	Composite Element REF	years_to_birth	vital_status	days_to_death	days_to_last_followup	
All	Value: 1097	Min. 26.0000 1st Qu. 49.0000 Median 59.0000 Mean 58.6000 3rd Qu. 68.0000 Max. 90.0000 Std.Dev. 13.1938 NA 15	0: 945 1: 152	Min. 116.0000 1st Qu. 700.5000 Median 1272.0000 Mean 1644.7000 3rd Qu. 2367.0000 Max. 7455.0000 Std.Dev. 1315.7558 NA 946	Min. -7.0000 1st Qu. 440.0000 Median 761.0000 Mean 1183.0000 3rd Qu. 1572.0000 Max. 8605.0000 Std.Dev. 1159.8806 NA 152	
	tumor_tissue_site	pathologic_stage	pathology_T_stage	pathology_N_stage	pathology_M_stage	
	breast: 1097	stage i: 90 stage ia: 86 stage ib: 7 stage ii: 6 stage iii: 358 stage iib: 258	stage iii: 2 stage iia: 156 stage iib: 27 stage iii: 65 stage iv: 20 stage x: 14 NA 8	t1: 41 t3: 137 t1a: 1 t3a: 1 t1b: 16 t4: 9 t1c: 223 t4b: 28 t2: 633 t4d: 3 t2a: 1 tx: 3 t2b: 1 t2c: 1	n0: 333 n1m1: 37 n0 (i-): 154 n2: 56 n0 (i+): 28 n2a: 64 n0 (mol+): 1 n3: 26 n1: 126 n3a: 47 n1a: 167 n3b: 3 n1b: 32 n3c: 1 n1c: 2 nx: 20	cm0 (i+): 6 m0: 906 m1: 22 mx: 163
	gender	date_of_initial_pathologic_diagnosis	days_to_last_known_alive	radiation_therapy	histological_type	
	female: 1085 male: 12	Min. 1988.0000 1st Qu. 2007.0000 Median 2009.0000 Mean 2008.0000 3rd Qu. 2010.0000 Max. 2013.0000 Std.Dev. 1159.8806 NA 2	Min. 735.0000 1st Qu. 1195.0000 Median 1656.0000 Mean 1656.0000 3rd Qu. 2116.0000 Max. 2576.0000 Std.Dev. 1301.7836 NA 1095	no: 446 yes: 556 NA 95	infiltrating carcinoma nos 1 infiltrating ductal carcinoma 784 infiltrating lobular carcinoma 203 medullary carcinoma 6 metaplastic carcinoma 9 mixed histology (please specify) 30 mucinous carcinoma 17 other, specify 46 NA 1	
	number_of_lymph_nodes	race	ethnicity			
	Min. 0.0000 1st Qu. 0.0000 Median 1.0000 Mean 2.3630 3rd Qu. 2.0000 Max. 35.0000 Std.Dev. 4.6339 NA 168	american indian or alaska native 1 asian: 61 black or african american : 183 white: 757 NA 95	hispanic or latino 39 not hispanic or latino : 884 NA 174			

Table 90: BRCA 데이터의 clinical part 요약

25.2.12 HNSC: Head and Neck squamous cell carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)

Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 322 2014. Broad Institute of MIT and Harvard, doi:10.7908/C1DN43V9 (2014).

<https://gdac.broadinstitute.org/>

여기에서 RNA-Seq expression and clinical data를 활용하였다.

- Clinical Data: Clinical_Pick_Tier1 (MD5)
- mRNASeq: illuminahisq_rnaseqv2-RSEM_genes_normalized (MD5)
- miRSeq: illumina_mirnaseq-miR_gene_expression (MD5)

25.2.12.1 변수 요약

Instance 수는 528명이며, 각 데이터는 다음과 같이 구성되어 있다. 분석할 때는 이들을 머지하여여 한다.

- clinical data는 20개의 변수
- mRNASeq data는 567명의 환자에 대한 20533개의 변수
- miRSeq data는 1287명의 환자에 대한 1048개의 변수

mRNASeq와 miRSeq의 환자 수가 clinical data와 차이가 있으며, repeated measured case도 있음을 유의할 필요가 있다. 또한 데이터에 missing value도 다수 존재하는 편이다.

- Composite
 - Value: (모든 값이 Value이므로 의미가 없다.)
- years_to_birth, a numeric vector
- vital_status
 - 0
 - 1
- days_to_death, a numeric vector
- days_to_last_followup, a numeric vector
- tumor_tissue_site
 - breast: (모든 값이 head and neck이다.)
- pathologic_stage
 - stage i
 - stage ii
 - stage iii
 - stage iva
 - stage ivb
 - stage ivc
- pathology_T_stage
 - t0
 - t1
 - t2
 - t3
 - t4
 - t4a
 - t4b
 - tx
- pathology_N_stage
 - n0
 - n1
 - n2
 - n2a
 - n2b
 - n2c
 - n3
 - nx
- pathology_M_stage
 - m0
 - m1
 - mx

- gender
 - female
 - male
- date_of_initial_pathologic_diagnosis, a numeric vector
- days_to_last_known_alive
- radiation_therapy
 - no
 - yes
- histological_type
 - head & neck squamous cell carcinoma
 - head & neck squamous cell carcinoma basaloid type
 - head & neck squamous cell carcinoma, spindle cell variant
- number_pack_years_smoked, a numeric vector
- year_of_tobacco_smoking_onset, a numeric vector
- number_of_lymph_nodes, a numeric vector
- race
 - american indian or alaska native
 - asian
 - black or african american
 - white
- ethnicity
 - hispanic or latino
 - not hispanic or latino

	Composite Element REF	years_to_birth	vital_status	days_to_death	days_to_last_followup	
All	Value: 528	Min. 19.0000 1st Qu. 53.0000 Median 61.0000 Mean 60.9100 3rd Qu. 69.0000 Max. 89.0000 Std.Dev. 11.9220 NA 1	0: 304 1: 224	Min. 2.0000 1st Qu. 260.0000 Median 430.0000 Mean 740.0000 3rd Qu. 814.5000 Max. 6417.0000 Std.Dev. 922.0532 NA 305	Min. 11.0000 1st Qu. 529.5000 Median 851.0000 Mean 1042.9000 3rd Qu. 1404.0000 Max. 5480.0000 Std.Dev. 801.1199 NA 225	
		tumor_tissue_site	pathologic_stage	pathology_T_stage	pathology_N_stage	pathology_M_stage
	head and neck: 528	stage i: 27 stage ii: 77 stage iii: 82 stage iv: 257 stage ivb: 12 stage ivc: 1 NA 72	t0: 1 t4a: 160 t1: 49 t4b: 4 t2: 140 tx: 39 t3: 101 NA 23 t4: 11	n0: 180 n2c: 48 n1: 68 n3: 8 n2: 12 nx: 75 n2a: 8 NA 25 n2b: 104	m0: 191 m1: 1 mx: 65 NA 271	
		gender	date_of_initial_pathologic_diagnosis	days_to_last_known_alive	radiation_therapy	histological_type
	female: 142 male: 386	Min. 1992.0000 1st Qu. 2007.0000 Median 2010.0000 Mean 2008.0000 3rd Qu. 2011.0000 Max. 2013.0000 Std.Dev. 5.1484 NA 1	Min. 98.0000 1st Qu. 483.0000 Median 848.5000 Mean 1226.9000 3rd Qu. 1647.5000 Max. 3930.0000 Std.Dev. 1074.6822 NA 506	no: 163 yes: 303 NA 62	head & neck squamous cell carcinoma head & neck squamous cell carcinoma basaloid type head & neck squamous cell carcinoma, spindle cell variant	517 10 1
		number_pack_years_smoked	year_of_tobacco_smoking_onset	number_of_lymph_nodes	race	ethnicity
	Min. 0.0169 1st Qu. 25.00000 Median 40.0000 Mean 45.7550 3rd Qu. 60.0000 Max. 300.0000 Std.Dev. 35.2134 NA 246	Min. 1936.0000 1st Qu. 1959.0000 Median 1968.0000 Mean 1967.0000 3rd Qu. 1975.0000 Max. 2001.0000 Std.Dev. 35.2134 NA 246	Min. 0.0000 1st Qu. 0.0000 Median 1.0000 Mean 2.1860 3rd Qu. 3.0000 Max. 44.0000 Std.Dev. 4.2702 NA 115	american indian or alaska native : 2 asian: 11 black or african american : 48 white: 452 NA 15	hispanic or latino : 26 not hispanic or latino : 465 NA 37	

Table 91: HNSC 허이터의 clinical part 요약

25.2.13 KIRC: Kidney renal clear cell carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)

Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 322 2014. Broad Institute of MIT and Harvard, doi:10.7908/C1DN43V9 (2014).

<https://gdac.broadinstitute.org/>

여기에서 RNA-Seq expression and clinical data를 활용하였다.

- Clinical Data: Clinical_Pick_Tier1 (MD5)
- mRNASeq: illuminahisseq_rnaseqv2-RSEM_genes_normalized (MD5)
- miRSeq: illuminahisseq_mirnaseq-miR_gene_expression (MD5)

25.2.13.1 변수 요약

Instance 수는 537명이며, 각 데이터는 다음과 같이 구성되어 있다. 분석할 때는 이들을 머지하여 한다.

- clinical data는 20개의 변수
- mRNASeq data는 607명의 환자에 대한 20533개의 변수
- miRSeq data는 979명의 환자에 대한 1048개의 변수

mRNASeq와 miRSeq의 환자 수가 clinical data와 차이가 있으며, repeated measured case도 있음을 유의할 필요가 있다. 또한 데이터에 missing value도 다수 존재하는 편이다.

- Composite
 - Value: (모든 값이 Value이므로 의미가 없다.)
- years_to_birth, a numeric vector
- vital_status
 - 0
 - 1
- days_to_death, a numeric vector
- days_to_last_followup, a numeric vector
- tumor_tissue_site
 - kidney: (모든 값이 kidney이다.)
- pathologic_stage
 - stage i
 - stage ii
 - stage iii
- pathology_T_stage
 - t1
 - t1a
 - t1b
 - t2
 - t2a
 - t2b
 - t3
 - t3a
 - t3b
 - t3c
 - t4
- pathology_N_stage
 - n0
 - n1
 - nx

- pathology_M_stage
 - m0
 - m1
 - mx
- gender
 - female
 - male
- date_of_initial_pathologic_diagnosis, a numeric vector
- days_to_last_known_alive, a numeric vector
- radiation_therapy
 - no
 - yes
- karnofsky_performance_score, a numeric vector
- histological_type
 - kidney clear cell renal carcinoma
- number_pack_years_smoked, a numeric vector
- year_of_tobacco_smoking_onset, a numeric vector
- number_of_lymph_nodes, a numeric vector
- race
 - asian
 - black or african american
 - white
- ethnicity
 - hispanic or latino
 - not hispanic or latino

	Composite Element REF	years_to_birth	vital_status	days_to_death	days_to_last_followup
All	Value: 537	Min. 26.0000 1st Qu. 52.0000 Median 61.0000 Mean 60.5700 3rd Qu. 70.0000 Max. 90.0000 Std.Dev. 12.1503 NA 1	0: 360 1: 177	Min. 2.0000 1st Qu. 333.0000 Median 819.0000 Mean 961.2000 3rd Qu. 1432.0000 Max. 3615.0000 Std.Dev. 763.4074 NA 360	Min. 0.0000 1st Qu. 710.5000 Median 1454.5000 Mean 1536.9000 3rd Qu. 2172.0000 Max. 4537.0000 Std.Dev. 1021.4460 NA 177
	tumor_tissue_site	pathologic_stage	pathology_T_stage	pathology_N_stage	pathology_M_stage
	kidney: 537	stage i: 269 stage ii: 57 stage iii: 125 stage iv: 84 NA: 2	t1: 22 t3: 5 t1a: 142 t3a: 122 t1b: 111 t3b: 53 t2: 55 t3c: 2 t2a: 10 t4: 11 t2b: 4	n0: 240 n1: 17 nx 280	m0: 426 m1: 79 mx: 30 NA 2
	gender	date_of_initial_pathologic_diagnosis	days_to_last_known_alive	radiation_therapy	karnofsky_performance_score
	female: 191 male: 346	Min. 1998.0000 1st Qu. 2004.0000 Median 2006.0000 Mean 2006.0000 3rd Qu. 2007.0000 Max. 2013.0000 Std.Dev. 910.4067 NA 2.7613	Min. 0.0000 1st Qu. 191.0000 Median 1172.0000 Mean 1117.0000 3rd Qu. 1887.0000 Max. 2799.0000 Std.Dev. 910.4067 NA 510	no: 142 yes: 2 NA 393	Min. 0.0000 1st Qu. 90.0000 Median 90.0000 Mean 85.5600 3rd Qu. 100.0000 Max. 100.0000 Std.Dev. 25.8199 NA 483
	histological_type	number_pack_years_smoked	year_of_tobacco_smoking_onset	race	ethnicity
	kidney clear cell renal carcinoma: 537	Min. 7.0000 1st Qu. 14.0000 Median 30.0000 Mean 28.30300 3rd Qu. 40.0000 Max. 65.0000 Std.Dev. 16.1224 NA 516	Min. 1946.0000 1st Qu. 1966.0000 Median 1978.0000 Mean 1979.0000 3rd Qu. 1996.0000 Max. 2001.0000 Std.Dev. 17.8230 NA 525	asian: 8 black or african american : 565 white: 466 NA 7	hispanic or latino : 26 not hispanic or latino : 359 NA 152

Table 92: KIRC 데이터의 clinical part 요약

25.2.14 LGG: Brain Lower Grade Glioma from Broad Institute TCGA Genome Data Analysis Center (2014)

Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 322 2014. Broad Institute of MIT and Harvard, doi:10.7908/C1DN43V9 (2014).

<https://gdac.broadinstitute.org/>

여기에서 RNA-Seq expression and clinical data를 활용하였다.

- Clinical Data: Clinical_Pick_Tier1 (MD5)
- mRNASeq: illuminahisq_rnaseqv2-RSEM_genes_normalized (MD5)
- miRSeq: illuminahisq_mirnaseq-miR_gene_expression (MD5)

25.2.14.1 변수 요약

Instance 수는 515명이며, 각 데이터는 다음과 같이 구성되어 있다. 분석할 때는 이들을 머지하여여 한다.

- clinical data는 14개의 변수
- mRNASeq data는 531명의 환자에 대한 20533개의 변수
- miRSeq data는 1579명의 환자에 대한 1048개의 변수

mRNASeq와 miRSeq의 환자 수가 clinical data와 차이가 있으며, repeated measured case도 있음을 유의할 필요가 있다. 또한 데이터에 missing value도 다수 존재하는 편이다.

- Composite
 - Value: (모든 값이 Value이므로 의미가 없다.)
- years_to_birth, a numeric vector
- vital_status
 - 0
 - 1
- days_to_death, a numeric vector
- days_to_last_followup, a numeric vector
- tumor_tissue_site
 - central nervous system: (모든 값이 central nervous system이다.)
- gender
 - female
 - male
- days_to_last_known_alive, a numeric vector
- radiation_therapy
 - no
 - yes
- karnofsky_performance_score, a numeric vector
- histological_type
 - astrocytoma
 - oligoastrocytoma
 - oligodendrogioma
- race
 - american indian or alaska native
 - asian
 - black or african american
 - white
- ethnicity
 - hispanic or latino
 - not hispanic or latino

	Composite Element REF	years_to_birth	vital_status	days_to_death	days_to_last_followup
All	Value: 515	Min. 14.0000 1st Qu. 32.0000 Median 41.0000 Mean 42.9300 3rd Qu. 53.0000 Max. 86.0000 Std.Dev. 13.3611 NA 1	0: 389 1: 126	Min. 7.0000 1st Qu. 438.0000 Median 629.0000 Mean 880.1000 3rd Qu. 1147.0000 Max. 6423.0000 Std.Dev. 1146.2006 NA 390	Min. -1.0000 1st Qu. 384.0000 Median 629.0000 Mean 880.1000 3rd Qu. 1147.0000 Max. 6423.0000 Std.Dev. 874.9760 NA 126
	tumor_tissue_site	gender	date_of_initial_pathologic_diagnosis	days_to_last_known_alive	radiation_therapy
central nervous system	515	female: 230 male: 285	Min. 1992.0000 1st Qu. 2008.0000 Median 2011.0000 Mean 2009.0000 3rd Qu. 2012.0000 Max. 2013.0000 Std.Dev. 4.4293 NA	Min. 1st Qu. Median Mean 3rd Qu. Max. Std.Dev. NA	no: 186 yes: 296 NA 33
karnofsky_performance_score	histological_type	race	ethnicity		
Min. 40.0000 1st Qu. 80.0000 Median 90.0000 Mean 86.6400 3rd Qu. 100.0000 Max. 100.0000 Std.Dev. 12.5823 NA 208	astrocytoma 194 oligoastrocytoma 130 oligodendrogloma 191	american indian or alaska native : 1 asian: 8 black or african american : 21 white: 475 NA 10	hispanic or latino : 32 not hispanic or latino : 449 NA 34		

Table 93: LGG 데이터의 clinical part 요약

25.2.15 LIHC: Liver hepatocellular carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)

Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 322 2014. Broad Institute of MIT and Harvard, doi:10.7908/C1DN43V9 (2014).

<https://gdac.broadinstitute.org/>

여기에서 RNA-Seq expression and clinical data를 활용하였다.

- Clinical Data: Clinical_Pick_Tier1 (MD5)
- mRNASeq: illuminahisseq_rnaseqv2-RSEM_genes_normalized (MD5)
- miRSeq: illuminahisseq_mirnaseq-miR_gene_expression (MD5)

25.2.15.1 변수 요약

Instance 수는 377명이며, 각 데이터는 다음과 같이 구성되어 있다. 분석할 때는 이들을 머지하여여 한다.

- clinical data는 18개의 변수
- mRNASeq data는 424명의 환자에 대한 20533개의 변수
- miRSeq data는 1273명의 환자에 대한 1048개의 변수

mRNASeq와 miRSeq의 환자 수가 clinical data와 차이가 있으며, repeated measured case도 있음을 유의할 필요가 있다. 또한 데이터에 missing value도 다수 존재하는 편이다.

• Composite

– Value: (모든 값이 Value이므로 의미가 없다.)

• years_to_birth, a numeric vector

• vital_status

– 0
– 1

• days_to_death, a numeric vector

• days_to_last_followup, a numeric vector

• tumor_tissue_site

– liver: (모든 값이 liver이다.)

• pathologic_stage

– stage i
– stage ii
– stage iii
– stage iiia
– stage iiib
– stage iiic
– stage iv
– stage iva
– stage ivb

• pathology_T_stage

– t1
– t2
– t2a
– t2b
– t3
– t3a
– t3b
– t4
– tx

• pathology_N_stage

– n0
– n1
– nx

• pathology_M_stage

– m0
– m1
– mx

- gender
 - female
 - male
- days_to_last_known_alive, a numeric vector
- radiation_therapy
 - no
 - yes
- histological_type
 - fibrolamellar carcinoma
 - hepatocellular carcinoma
 - hepatocholangiocarcinoma (mixed)
- residual_tumor
 - r0
 - r1
 - r2
 - rx
- race
 - american indian or alaska native
 - asian
 - black or african american
 - white
- ethnicity
 - hispanic or latino
 - not hispanic or latino

	Composite Element REF	years_to_birth	vital_status	days_to_death	days_to_last_followup	
All	Value: 377	Min. 16.0000 1st Qu. 51.0000 Median 61.0000 Mean 59.2700 3rd Qu. 69.0000 Max. 87.0000 Std.Dev. 13.4290 NA 4	0: 245 1: 132	Min. 9..0000 1st Qu. 194.8000 Median 417.5000 Mean 672.1000 3rd Qu. 837.0000 Max. 3258.0000 Std.Dev. 696.2465 NA 245	Min. 0.0000 1st Qu. 395.8000 Median 649.5000 Mean 885.8000 3rd Qu. 1222.0000 Max. 3675.0000 Std.Dev. 738.8077 NA 133	
	tumor_tissue_site	pathologic_stage	pathology_T.stage	pathology_N.stage	pathology_M.stage	
	liver: 377	stage i: 175 stage ii: 87 stage iii: 3 stage iiiia: 65 stage iiib: 9	stage iic: 9 stage iv: 2 stage ivb: 2 NA 24	t1: 185 t3a: 29 t2: 93 t3b: 7 t2a: 1 t4: 13 t2b: 1 tx: 1 t3: 45 NA 2	n0: 257 n1: 4 nx: 115 NA 1	m0: 272 m1: 4 mx: 101 NA
	gender	date_of_initial_pathologic_diagnosis	days_to_last_known_alive	radiation_therapy	histological_type	
	female: 122 male: 255	Min. 1995.0000 1st Qu. 2008.0000 Median 2011.0000 Mean 2010.0000 3rd Qu. 2012.0000 Max. 2013.0000 Std.Dev. 3.7625 NA 3	Min. 1st Qu. Median Mean 3rd Qu. Max. Std.Dev. NA 377	no: 345 yes: 9 NA 33	fibrolamellar carcinoma 3 hepatocellular carcinoma 367 hepatocholangiocarcinoma (mixed) 7	
	residual_tumor	race	ethnicity			
	r0: 330 r1: 17 r2: 1 rx 22 NA 7	american indian or alaska native 2 asian: 161 black or african american : 17 white: 187 NA 10	hispanic or latino 18 not hispanic or latino : 40 NA 19			

Table 94: LIHC의 데이터의 clinical part 요약

25.2.16 LUAD: Lung adenocarcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)

Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 322 2014. Broad Institute of MIT and Harvard, doi:10.7908/C1DN43V9 (2014).

<https://gdac.broadinstitute.org/>

여기에서 RNA-Seq expression and clinical data를 활용하였다.

- Clinical Data: Clinical.Pick.Tier1 (MD5)
- mRNASeq: illuminahiseq_rnaseqv2-RSEM_genes_normalized (MD5)
- miRSeq: illuminahiseq_mirnaseq-miR_gene_expression (MD5)

25.2.16.1 변수 요약

Instance 수는 522명이며, 각 데이터는 다음과 같이 구성되어 있다. 분석할 때는 이들을 머지하여여 한다.

- clinical data는 21개의 변수
- mRNASeq data는 577명의 환자에 대한 20533개의 변수
- miRSeq data는 1495명의 환자에 대한 1048개의 변수

mRNASeq와 miRSeq의 환자 수가 clinical data와 차이가 있으며, repeated measured case도 있음을 유의할 필요가 있다. 또한 데이터에 missing value도 다수 존재하는 편이다.

- Composite
 - Value: (모든 값이 Value이므로 의미가 없다.)
- years_to_birth, a numeric vector
- vital_status
 - 0
 - 1
- days_to_death, a numeric vector
- days_to_last_followup, a numeric vector
- tumor_tissue_site
 - lung: (모든 값이 lung이다.)
- pathologic_stage
 - stage i
 - stage ia
 - stage ib
 - stage ii
 - stage iia
 - stage iib
 - stage iiia
 - stage iiib
 - stage iv
- pathology_T_stage
 - t1
 - t1a
 - t1b
 - t2
 - t2a
 - t2b
 - t3
 - t4
 - tx
- pathology_N_stage
 - n0
 - n1
 - n2
 - n3
 - nx
- pathology_M_stage
 - m0
 - m1
 - m1a
 - m1b
 - mx

- gender
 - female
 - male
- date_of_initial_pathologic_diagnosis, a numeric vector
- days_to_last_known_alive, a numeric vector
- radiation_therapy
 - no
 - yes
- karnofsky_performance_score, a numeric vector
- histological_type
 - lung acinar adenocarcinoma
 - lung adenocarcinoma- not otherwise specified (nos)
 - lung adenocarcinoma mixed subtype
 - lung bronchioloalveolar carcinoma mucinous
 - lung bronchioloalveolar carcinoma nonmucinous
 - lung clear cell adenocarcinoma
 - lung micropapillary adenocarcinoma
 - lung mucinous adenocarcinoma
 - lung papillary adenocarcinoma
 - lung signet ring adenocarcinoma
 - lung solid pattern predominant adenocarcinoma
 - mucinous (colloid) carcinoma
- number_pack_years_smoked, a numeric vector
- year_of_tobacco_smoking_onset, a numeric vector
- residual_tumor
 - r0
 - r1
 - r2
 - rx
- race
 - american indian or alaska native
 - asian
 - black or african american
 - white
- ethnicity
 - hispanic or latino
 - not hispanic or latino

Composite Element REF		years_to_birth	vital_status	days_to_death	days_to_last_followup	
A II	Value: 522	Min. 33.0000 1st Qu. 59.0000 Median 66.0000 Mean 65.2400 3rd Qu. 72.0000 Max. 88.0000 Std.Dev. 10.0303 NA 31	0: 334 1: 188	Min. 0.0000 1st Qu. 297.8000 Median 619.0000 Mean 791.4000 3rd Qu. 1120.0000 Max. 4961.0000 Std.Dev. 700.4983 NA 338	Min. 1991.0000 1st Qu. 2007.0000 Median 2010.0000 Mean 2008.0000 3rd Qu. 2011.0000 Max. 2013.0000 Std.Dev. 978.4594 NA 19	
		tumor_tissue_site	pathologic_stage	pathology_T_stage	pathology_N_stage	
	lung: 522	stage i: 5 stage ia: 134 stage ib: 140 stage ii: 1 stage iiia: 50 stage iiib: 73	stage iii: 74 stage iiib: 11 stage iv: 26 NA 8	t1: 69 t2b: 28 t1a: 48 t3: 47 t1b: 55 t4: 19 t2: 171 tx: 3 t2a: 82	n0: 335 n1: 98 n2: 75 n3: 2 nx: 11 NA 1	m0: 353 m1: 18 m1a: 2 m1b: 5 mx: 140 NA 4
gender		date_of_initial_pathologic_diagnosis	days_to_last_known_alive	radiation_therapy	karnofsky_performance_score	
	female: 280 male: 242	Min. 1991.0000 1st Qu. 2007.0000 Median 2010.0000 Mean 2008.0000 3rd Qu. 2011.0000 Max. 2013.0000 Std.Dev. 4.2023 NA 19	Min. 9.0000 1st Qu. 96.7500 Median 169.0000 Mean 359.0000 3rd Qu. 576.0000 Max. 1178.0000 Std.Dev. NA NA	no: 414 yes: 61 NA 47	Min. 0.0000 1st Qu. 80.0000 Median 90.0000 Mean 78.5500 3rd Qu. 100.0000 Max. 100.0000 Std.Dev. 28.4272 NA 384	
histological_type		number_pack_years_smoked	year_of_tobacco_smoking_onset	residual_tumor	race	
- not otherwise specified (nos)	lung acinar adenocarcinoma	18	lung micropapillary adenocarcinoma	3	american indian or alaska native : 1 asian: 8 black or african american : 53 white: 393 NA 67	
	lung adenocarcinoma	327	lung mucinous adenocarcinoma	2		
	lung adenocarcinoma mixed subtype	107	lung papillary adenocarcinoma	23		
	lung bronchioloalveolar carcinoma mucinous	5	lung signet ring adenocarcinoma	1		
	lung bronchioloalveolar carcinoma nonmucinous	19	lung solid pattern predominant adenocarcinoma	5		
	lung clear cell adenocarcinoma	2	mucinous (colloid) carcinoma	10		
ethnicity						
	hispanic or latino	7				
	not hispanic or latino	389				
	NA	126				

Table 95: Support 테이터 요약 (Cont'd)

25.2.17 LUSC: Lung squamous cell carcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)

Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 322 2014. Broad Institute of MIT and Harvard, doi:10.7908/C1DN43V9 (2014).

<https://gdac.broadinstitute.org/>

여기에서 RNA-Seq expression and clinical data를 활용하였다.

- Clinical Data: Clinical.Pick.Tier1 (MD5)
- mRNASeq: illuminahisq_rnaseqv2-RSEM_genes_normalized (MD5)
- miRSeq: illuminahisq_mirnaseq-miR_gene_expression (MD5)

25.2.17.1 변수 요약

Instance 수는 504명이며, 각 데이터는 다음과 같이 구성되어 있다. 분석할 때는 이들을 머지하여여 한다.

- clinical data는 21개의 변수
- mRNASeq data는 553명의 환자에 대한 20533개의 변수
- miRSeq data는 1162명의 환자에 대한 1048개의 변수

mRNASeq와 miRSeq의 환자 수가 clinical data와 차이가 있으며, repeated measured case도 있음을 유의할 필요가 있다. 또한 데이터에 missing value도 다수 존재하는 편이다.

- Composite
 - Value: (모든 값이 Value이므로 의미가 없다.)

- years_to_birth, a numeric vector

- vital_status
 - 0
 - 1

- days_to_death, a numeric vector

- days_to_last_followup, a numeric vector

- tumor_tissue_site
 - lung: (모든 값이 lung이다.)

- pathologic_stage
 - stage i
 - stage ia
 - stage ib
 - stage ii
 - stage iia
 - stage iib
 - stage iii
 - stage iiib
 - stage iv

- pathology_T_stage
 - t1
 - t1a
 - t1b
 - t2
 - t2a
 - t2b
 - t3
 - t4

- pathology_N_stage
 - n0
 - n1
 - n2
 - n3
 - nx

- pathology_M_stage
 - m0
 - m1
 - m1a
 - m1b
 - mx

- gender
 - female
 - male
- date_of_initial_pathologic_diagnosis, a numeric vector
- days_to_last_known_alive, a numeric vector
- radiation_therapy
 - no
 - yes
- karnofsky_performance_score, a numeric vector
- histological_type
 - lung basaloid squamous cell carcinoma
 - lung papillary squamous cell carcinoma
 - lung small cell squamous cell carcinoma
 - lung squamous cell carcinoma- not otherwise specified (nos)
- number_pack_years_smoked
- year_of_tobacco_smoking_onset
- residual_tumor
 - r0
 - r1
 - r2
 - rx
- race
 - asian
 - black or african american
 - white
- ethnicity
 - hispanic or latino
 - not hispanic or latino

	Composite Element REF	years_to_birth	vital_status	days_to_death	days_to_last_followup
All	Value: 504	Min. 39.0000	0: 284 1: 220	Min. 1.0000	Min. 1992.0000
		1st Qu. 62.0000 Median 68.0000 Mean 67.2600 3rd Qu. 73.0000 Max. 90.0000 Std.Dev. 8.6140 NA 10		1st Qu. 280.0000 Median 550.0000 Mean 872.3000 3rd Qu. 1110.5000 Max. 5287.0000 Std.Dev. 907.4474 NA 289	1st Qu. 2005.0000 Median 2009.0000 Mean 2008.0000 3rd Qu. 2011.0000 Max. 2013.0000 Std.Dev. 996.3079 NA 221
lung	tumor_tissue_site	pathologic_stage	pathology_T_stage	pathology_N_stage	pathology_M_stage
	lung: 504	stage i: 3 stage ia: 90 stage ib: 152 stage ii: 3 stage iiia: 65 stage iib: 95	t1: 50 t2a: 87 t1a: 24 t2b: 34 t1b: 40 t3: 71 t2: 174 t4: 24	n0: 320 n1: 133 n2: 40 n3: 5 nx: 6	m0: 414 m1: 5 m1a: 1 m1b: 1 mx: 79 NA 4
		gender	date_of_initial_pathologic_diagnosis	days_to_last_known_alive	radiation_therapy karnofsky_performance_score
female: 131	male: 373	Min. 1992.0000 1st Qu. 2005.0000 Median 2009.0000 Mean 2008.0000 3rd Qu. 2011.0000 Max. 2013.0000 Std.Dev. 4.1717 NA 25	Min. 4.0000 1st Qu. 298.8000 Median 706.0000 Mean 904.8000 3rd Qu. 1051.8000 Max. 2734.0000 Std.Dev. 860.5114 NA 436	no: 387 yes: 53 NA 64	Min. 0.0000 1st Qu. 0.0000 Median 80.0000 Mean 60.3000 3rd Qu. 90.0000 Max. 100.0000 Std.Dev. 41.0532 NA 338
		histological_type	number_pack_years_smoked	year_of_tobacco_smoking_onset	residual_tumor race
lung basaloid	squamous cell carcinoma	Min. 1.0000	Min. 1933.0000	r0: 401 r1: 12 r2: 4 rx: 23 NA 64	asian: 9 black or african american : 31 white: 351 NA 113
		1st Qu. 31.1200 Median 50.0000 Mean 52.9100 3rd Qu. 64.5000 Max. 240.0000 Std.Dev. 31.1570 NA 77	1st Qu. 1952.0000 Median 1960.0000 Mean 1960.0000 3rd Qu. 1968.0000 Max. 1997.0000 Std.Dev. 11.5653 NA 183		
ethnicity	hispanic or latino	8			
	not hispanic or latino	319 NA 177			

Table 96: LUSC 데이터의 clinical part 요약

25.2.18 OV: Ovarian serous cystadenocarcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)

Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 322 2014. Broad Institute of MIT and Harvard, doi:10.7908/C1DN43V9 (2014).

<https://gdac.broadinstitute.org/>

여기에서 RNA-Seq expression and clinical data를 활용하였다.

- Clinical Data: Clinical_Pick_Tier1 (MD5)
- mRNASeq: illuminahisseq_rnaseqv2-RSEM_genes_normalized (MD5)
- miRSeq: illuminahisseq_mirnaseq-miR_gene_expression (MD5)

25.2.18.1 변수 요약

Instance 수는 591명이며, 각 데이터는 다음과 같이 구성되어 있다. 분석할 때는 이들을 머지하여 한다.

- clinical data는 21개의 변수
- mRNASeq data는 308명의 환자에 대한 20533개의 변수
- miRSeq data는 1384명의 환자에 대한 707개의 변수

mRNASeq와 miRSeq의 환자 수가 clinical data와 차이가 있으며, repeated measured case도 있음을 유의할 필요가 있다. 또한 데이터에 missing value도 다수 존재하는 편이다.

- Composite
 - Value: (모든 값이 Value이므로 의미가 없다.)
- years_to_birth, a numeric vector
- vital_status
 - 0
 - 1
- days_to_death, a numeric vector
- days_to_last_followup, a numeric vector
- tumor_tissue_site
 - omentum
 - ovary
 - peritoneum ovary
- pathologic_stage
 - (NA 값만 있음)
- pathology_T_stage
 - (NA 값만 있음)
- pathology_N_stage
 - (NA 값만 있음)
- pathology_M_stage
 - (NA 값만 있음)
- gender
 - female
 - male
- date_of_initial_pathologic_diagnosis, a numeric vector
- radiation_therapy
 - no
 - yes
- karnofsky_performance_score, a numeric vector

- histological_type
 - serous cystadenocarcinoma
- tumor_grade
 - (NA 값만 있음)
- tumor_stage
 - (NA 값만 있음)
- days_to_tumor_recurrence
 - (NA 값만 있음)
- chemo_therapy
 - (NA 값만 있음)
- residual_tumor
 - r0
 - r1
 - r2
 - rx
- ethnicity
 - hispanic or latino
 - not hispanic or latino

	Composite Element REF	years_to_birth	vital_status	days_to_death	days_to_last_followup
All	Value: 591	Min. 26.0000		Min. 8.0000	Min. 16.0000
		1st Qu. 51.0000		1st Qu. 567.5000	1st Qu. 268.0000
		Median 59.0000		Median 1073.0000	Median 837.0000
		Mean 59.7800	0: 247	Mean 1147.4000	Mean 1214.0000
		3rd Qu. 68.7500	1: 344	3rd Qu. 1557.0000	3rd Qu. 1919.0000
		Max. 89.0000		Max. 4624.0000	Max. 5481.0000
		Std.Dev. 11.5850		Std.Dev. 772.6693	Std.Dev. 1124.4820
		NA 21		NA 248	NA 358
	tumor_tissue_site	pathologic_stage	pathology_T_stage	pathology_N_stage	pathology_M_stage
	omentum: 3				
	ovary: 576	NA 591	NA 591	NA 591	NA 591
	peritoneum ovary: 2				
	NA 10				
	gender	date_of_initial_pathologic_diagnosis	radiation_therapy	karnofsky_performance_score	histological_type
	female: 581	Min. 1992.0000		40: 2	
	NA 10	1st Qu. 2001.0000		50: 2	
		Median 2004.0000		60: 20	
		Mean 2004.0000	no: 557	70: 1	serous cystadenocarcinoma 581
		3rd Qu. 2007.0000	yes: 5	80: 49	NA 10
		Max. 2013.0000	NA 29	90:	
		Std.Dev. 1124.4820		100: 10	
		NA 358		NA 507	
	tumor_grade	tumor_stage	days_to_tumor_recurrence	chemo_tumor	residual_tumor
	NA 591	NA 591	NA 591	NA 591	r0: 16 r1: 31 r2: 5 rx: 3 NA 536
	ethnicity				
	hispanic or latino 11				
	not hispanic or latino : 338				
	NA 242				

Table 97: OV 테이터의 clinical part 요약

25.2.19 STAD: Stomach adenocarcinoma from Broad Institute TCGA Genome Data Analysis Center (2014)

Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 322 2014. Broad Institute of MIT and Harvard, doi:10.7908/C1DN43V9 (2014).

<https://gdac.broadinstitute.org/>

여기에서 RNA-Seq expression and clinical data를 활용하였다.

- Clinical Data: Clinical.Pick.Tier1 (MD5)
- mRNASeq: illuminahiseq_rnaseqv2-RSEM_genes_normalized (MD5)
- miRSeq: illuminahiseq_mirnaseq-miR_gene_expression (MD5)

25.2.19.1 변수 요약

Instance 수는 443명이며, 각 데이터는 다음과 같이 구성되어 있다. 분석할 때는 이들을 머지하여여 한다.

- clinical data는 18개의 변수
- mRNASeq data는 451명의 환자에 대한 20533개의 변수
- miRSeq data는 1291명의 환자에 대한 707개의 변수

mRNASeq와 miRSeq의 환자 수가 clinical data와 차이가 있으며, repeated measured case도 있음을 유의할 필요가 있다. 또한 데이터에 missing value도 다수 존재하는 편이다.

- Composite
 - Value: (모든 값이 Value이므로 의미가 없다.)
- years_to_birth, a numeric vector
- vital_status
 - 0
 - 1
- days_to_death, a numeric vector
- days_to_last_followup, a numeric vector
- tumor_tissue_site
 - stomach: (모든 값이 stomach이다.)
- pathologic_stage
 - stage i
 - stage ia
 - stage ib
 - stage ii
 - stage iia
 - stage iib
 - stage iii
 - stage iiia
 - stage iiib
 - stage iiic
 - stage iv
- pathology_T_stage
 - t1
 - t1a
 - t1b
 - t2
 - t2a
 - t2b
 - t3
 - t4
 - t4a
 - t4b
 - tx
- pathology_N_stage
 - n0
 - n1
 - n2
 - n3
 - n3a
 - n3b
 - nx

- pathology_M_stage
 - m0
 - m1
 - mx
- gender
 - female
 - male
- date_of_initial_pathologic_diagnosis, a numeric vector
- radiation_therapy
 - no
 - yes
- karnofsky_performance_score, a numeric vector
- histological_type
 - stomach adenocarcinoma, signet ring type
 - stomach, adenocarcinoma, diffuse type
 - stomach, adenocarcinoma, not otherwise specified (nos)
 - stomach, intestinal adenocarcinoma, mucinous type
 - stomach, intestinal adenocarcinoma, not otherwise specified (nos)
 - stomach, intestinal adenocarcinoma, papillary type
 - stomach, intestinal adenocarcinoma, tubular type
- residual_tumor
 - r0
 - r1
 - r2
 - rx
- number_of_lymph_node, a numeric vector
- race
- asian
- black or african american
- native hawaiian or other pacific islander
- white
- ethnicity
 - hispanic or latino
 - not hispanic or latino

Composite Element REF		years_to_birth		vital_status		days_to_death		days_to_last_followup	
All	Value: 443	Min.	30.0000			Min.	0.0000	Min.	0.0000
		1st Qu.	58.0000			1st Qu.	194.0000	1st Qu.	335.5000
		Median	67.0000			Median	346.0000	Median	547.5000
		Mean	65.7300	0:	268	Mean	423.7000	Mean	673.7000
		3rd Qu.	73.0000	1:	175	3rd Qu.	553.5000	3rd Qu.	912.0000
		Max.	90.0000			Max.	3720.0000	Max.	3720.0000
		Std.Dev.	10.7517			Std.Dev.	369.9151	Std.Dev.	609.4170
		NA	9			NA	273	NA	177
tumor_tissue_site	pathologic_stage	pathology_T_stage		pathology_N_stage		pathology_M_stage		residual_tumor	
stomach: 443	stage i: 2 stage ia: 16 stage ib: 41 stage ii: 33 stage iii: 41 stage iiib: 56	stage iii: 3 stage iiia: 81 stage iiib: 63 stage iiic: 39 stage iv: 44 NA	stage iii: 3 stage iiia: 81 stage iiib: 63 stage iiic: 39 stage iv: 44 NA	t1: 6 t1a: 2 t1b: 15 t2: 69 t2a: 9 t2b: 15	t3: 198 t4: 32 t4a: 60 t4b: 27 tx: 10	n0: 132 n1: 119 n2: 86 n3: 32 n3a: 49 n3b: 7 nx: 17 NA: 1		m0: 391 m1: 30 mx: 22	
gender	date_of_initial_pathologic_diagnosis	radiation_therapy		histological_type		residual_tumor			
female: 158 male: 285	Min. 1996.0000 1st Qu. 2010.0000 Median 2011.0000 Mean 2010.0000 3rd Qu. 2012.0000 Max. 2013.0000 Std.Dev. 2.4563 NA 6	no: 323 yes: 77 NA: 43		stomach adenocarcinoma, signet ring type: stomach adenocarcinoma, diffuse type: stomach adenocarcinoma, not otherwise specified (nos): stomach, intestinal adenocarcinoma, mucinous type:	13 72 164 22	stomach, intestinal adenocarcinoma, not otherwise specified (nos): stomach, intestinal adenocarcinoma, papillary type: stomach, intestinal adenocarcinoma, tubular type: NA	82 8 79 3	r0: 350 r1: 18 r2: 19 rx: 25 NA: 31	
number_of_lymph_nodes	race	ethnicity							
Min. 0.0000 1st Qu. 0.0000 Median 3.0000 Mean 5.6350 3rd Qu. 8.0000 Max. 57.0000 Std.Dev. 8.4079 NA 51	asian: 89 black or african american : 13 native hawaiian or other pacific islander : 1 white: 278 NA 62	hispanic or latino : 5 not hispanic : 318 or latino : 120							

Table 98: STAD 데이터의 clinical part 요약

25.2.20 Acute Myeloid Leukemia from Bullinger et al. (2004)

Bullinger et al.(2004)[19]에 의해 제공된 데이터로, 다음 URL에 접속한 후, accession number "GSE425"를 입력하여 다운로드할 수 있다.

<https://www.ncbi.nlm.nih.gov/geo/>

따로 검색하지 않고 바로 데이터를 확인하려면 다음 링크로 접속하면 된다.

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE425>

25.2.20.1 변수 요약

특이하게도, 이 데이터셋은 7개의 테이블로 구성되어 있다.

- [Supplementary Table 1](#): Clinical, morphological, cytogenetic and molecular genetic information on 116 AML patient samples.
- [Supplementary Table 2](#): Summary of the distribution of clinical and molecular genetic characteristics within the AML sample set.
- [Supplementary Table 3](#): Fluorescence ratios of the 6,283 well-measured and variably-expressed genes.
- [Supplementary Table 4](#): Clinical and laboratory characteristics of normal karyotype predominant subtypes I and II.
- [Supplementary Table 5](#): Supervised analysis of group-specific gene expression signatures.
- [Supplementary Table 6](#): Gene-expression outcome class predictor.
- [Supplementary Table 7](#): Multivariate proportional hazards analysis. Keywords: other

작성중...

25.2.21 Node-positive breast Cancer from German Breast Cancer Study Group (GBCS2)

Byar and Green(1980)[20]의 연구를 통해 소개된 전립선암 데이터이다. 506명의 환자 데이터이며, R의 'clustMD' 패키지에 내장되어 있으나, clustMD 패키지에서는 475명의 환자 데이터를 제공하여 준다.⁴⁴ 다음과 같이 불러올 수 있다.

Code 22. Byar and Green(1980)의 전립선암 데이터 불러오기(R 코드)

```
1 library("clustMD")
2 data("Byar")
```

25.2.21.1 변수 요약

Instance 수는 475명이며, 총 15개의 변수가 있다. 다만 **Observation** 변수는 환자를 구분하기 위한 용도 외에는 큰 의미가 없다.

- **Age**, a numeric vector indicating the age of the patient.
- **Weight**, a numeric vector indicating the weight of the patient.
- **Performance.rating**, an ordinal variable indicating how active the patient is
 - 0: normal activity
 - 1: in bed less than 50% of daytime
 - 2: in bed more than 50% of daytime
 - 3: confined to bed.
- **Cardiovascular.disease.history**, a binary variable indicating if the patient has a history of cardiovascular disease
 - 0: no
 - 1: yes
- **Systolic.Blood.pressure**, a numeric vector indicating the systolic blood pressure of the patient in units of ten.
- **Diastolic.blood.pressure**, a numeric vector indicating the diastolic blood pressure of the patient in units of ten.
- **Electrocardiogram.code**, a nominal variable indicating the electorcardiogram code
 - 0: normal
 - 1: benign
 - 2: rythmic disturbances and electrolyte changes
 - 3: heart blocks or conduction defects
 - 4: heart strain
 - 5: old myocardial infarct
 - 6: recent myocardial infarct
- **Serum.haemoglobin**, a numeric vector indicating the serum haemoglobin levels of the patient measured in g/100ml.
- **Size.of.primary.tumour**, a numeric vector indicating the estimated size of the patient's primary tumour in centimeters squared.
- **Index.of.tumour.stage.and.histologic.grade**, a numeric vector indicating the combined index of tumour stage and histologic grade of the patient.
- **Serum.prostatic.acid.phosphatase**, a numeric vector indicating the serum prostatic acid phosphatase levels of the patient in King-Armstrong units.
- **Bone.metastases**, a binary vector indicating the presence of bone metastasis:
 - 0: no
 - 1: yes
- **Stage**, the stage of the patient's prostate cancer.
- **Observation**, a patient ID number.
- **SurvStat**, the post trial survival status of the patient:
 - 0: alive
 - 1: dead from prostatic cancer
 - 2: dead from heart or vascular disease
 - 3: dead from cerebrovascular accident
 - 4: dead form pulmonary embolus
 - 5: dead from other cancer
 - 6: dead from respiratory disease
 - 7: dead from other specific non-cancer cause
 - 8: dead from other unspecified non-cancer cause
 - 9: dead from unknown cause

⁴⁴이 데이터셋은 R뿐만 아니라 다른 링크에서도 구할 수 있었지만, 506명의 환자 데이터를 제공하는 곳을 찾지는 못했다. clustMD 패키지와 같은 데이터를 다음 링크의 07번 항목에서도 제공하고 있다.

<http://portal.uni-freiburg.de/imbi/Royston-Sauerbrei-book>

	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
All	Min. 48.000 1st Qu. 70.000 Median 73.000 Mean 71.560 3rd Qu. 76.000 Max. 89.000 Std.Dev. 6.920	Min. 69.000 1st Qu. 90.000 Median 98.000 Mean 99.010 3rd Qu. 107.000 Max. 152.000 Std.Dev. 13.341	0: 428 1: 32 2: 13 3: 2	0: 268 1: 207	8: 1 17: 33 9: 3 18: 16 10: 13 19: 12 11: 26 20: 2 12: 60 21: 2 13: 70 22: 3 14: 90 23: 1 15: 70 24: 1 16: 71 30: 1	4: 4 10: 62 5: 5 11: 9 6: 40 12: 5 7: 99 13: 1 8: 155 14: 1 9: 93 18: 1	0: 161 1: 23 2: 50 3: 25 4: 145 5: 70 6: 1	Min. 59.000 1st Qu. 122.500 Median 137.000 Mean 134.200 3rd Qu. 147.000 Max. 182.000 Std.Dev. 19.382
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostactic.acid.phosphatase	Bone.metastases	Stage	Observation	Surv.Stat	
Performance.rating ==0	Min. 0.000 1st Qu. 5.000 Median 10.000 Mean 14.290 3rd Qu. 21.000 Max. 69.000 Std.Dev. 12.236	5: 3 11: 110 6: 8 12: 26 7: 6 13: 70 8: 66 14: 5 9: 132 15: 16 10: 33	Min. 1.000 1st Qu. 5.000 Median 7.000 Mean 125.700 3rd Qu. 29.500 Max. 9999.000 Std.Dev. 638.486	0: 398 1: 77	3: 273 4: 202	Not Meaningful	0: 137 5: 24 1: 121 6: 16 2: 93 7: 27 3: 31 8: 6 4: 14 9: 6	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Performance.rating ==1	Min. 49.000 1st Qu. 70.000 Median 73.000 Mean 71.460 3rd Qu. 76.000 Max. 89.000 Std.Dev. 6.802	Min. 69.000 1st Qu. 90.000 Median 99.000 Mean 99.480 3rd Qu. 108.000 Max. 152.000 Std.Dev. 13.094	0: 428 1: 0 2: 0 3: 0	0: 249 1: 179	8: 1 17: 30 9: 3 18: 14 10: 13 19: 12 11: 24 20: 2 12: 56 21: 1 13: 64 22: 1 14: 75 23: 1 15: 66 24: 1 16: 65 30: 1	4: 4 10: 59 5: 3 11: 8 6: 38 12: 2 7: 88 13: 1 8: 136 14: 1 9: 87 18: 1	0: 151 1: 20 2: 46 3: 23 4: 128 5: 59 6: 1	Min. 59.000 1st Qu. 123.000 Median 137.000 Mean 135.300 3rd Qu. 147.000 Max. 182.000 Std.Dev. 18.964
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostactic.acid.phosphatase	Bone.metastases	Stage	Observation	Surv.Stat	
Performance.rating ==1	Min. 0.000 1st Qu. 5.000 Median 10.000 Mean 13.830 3rd Qu. 20.000 Max. 69.000 Std.Dev. 11.838	5: 2 11: 97 6: 8 12: 23 7: 6 13: 59 8: 62 14: 3 9: 125 15: 13 10: 30	Min. 1.000 1st Qu. 5.000 Median 7.000 Mean 89.100 3rd Qu. 24.250 Max. 5960.000 Std.Dev. 418.059	0: 371 1: 57	3: 255 4: 173	Not Meaningful	0: 133 5: 24 1: 101 6: 16 2: 84 7: 23 3: 25 8: 4 4: 12 9: 6	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Performance.rating ==2	Min. 48.000 1st Qu. 69.750 Median 73.000 Mean 72.190 3rd Qu. 78.000 Max. 87.000 Std.Dev. 8.731	Min. 71.000 1st Qu. 87.750 Median 96.000 Mean 95.340 3rd Qu. 99.250 Max. 36.000 Std.Dev. 14.098	0: 0 1: 32 2: 0 3: 0	0: 10 1: 22	8: 17: 3 9: 18: 1 10: 19: 1 11: 2 20: 1 12: 2 21: 1 13: 3 22: 2 14: 12 23: 1 15: 2 24: 1 16: 3 30: 1	4: 4 10: 3 5: 1 11: 1 6: 2 12: 3 7: 7 13: 1 8: 13 14: 1 9: 2 18: 1	0: 7 1: 2 2: 3 3: 1 4: 12 5: 7 6:	Min. 91.000 1st Qu. 117.000 Median 128.500 Mean 127.400 3rd Qu. 140.000 Max. 176.000 Std.Dev. 17.816
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostactic.acid.phosphatase	Bone.metastases	Stage	Observation	Surv.Stat	
Performance.rating ==2	Min. 1.000 1st Qu. 4.000 Median 15.000 Mean 18.000 3rd Qu. 25.000 Max. 61.000 Std.Dev. 15.149	5: 1 11: 8 6: 12: 2 7: 13: 5 8: 3 14: 2 9: 6 15: 3 10: 3	Min. 2.000 1st Qu. 6.000 Median 9.000 Mean 80.090 3rd Qu. 23.500 Max. 1278.000 Std.Dev. 237.014	0: 23 1: 9	3: 15 4: 17	Not Meaningful	0: 2 5: 1: 9 6: 2: 8 7: 3 3: 6 8: 2 4: 2 9:	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Performance.rating ==2	Min. 60.000 1st Qu. 72.000 Median 74.000 Mean 73.920 3rd Qu. 78.000 Max. 80.000 Std.Dev. 5.235	Min. 73.000 1st Qu. 79.000 Median 93.000 Mean 94.777 3rd Qu. 105.000 Max. 134.000 Std.Dev. 17.331	0: 0 1: 0 2: 13 3: 0	0: 7 1: 6	8: 17: 9: 18: 2 10: 19: 11: 20: 12: 2 21: 13: 2 22: 14: 2 23: 15: 2 24: 16: 3 30:	4: 4 10: 5: 1 11: 6: 12: 7: 4 13: 8: 5 14: 9: 4 18:	0: 2 5: 1: 1 1: 2: 1 2: 3: 3 4: 4: 5 4: 5: 4 4: 6:	Min. 72.000 1st Qu. 101.000 Median 123.000 Mean 116.500 3rd Qu. 135.000 Max. 148.000 Std.Dev. 25.877
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostactic.acid.phosphatase	Bone.metastases	Stage	Observation	Surv.Stat	
Performance.rating ==3	Min. 1.000 1st Qu. 11.000 Median 17.000 Mean 20.920 3rd Qu. 26.000 Max. 54.000 Std.Dev. 15.141	5: 1 11: 4 6: 12: 1 7: 13: 5 8: 1 14: 9: 1 15: 10: 1	Min. 3.000 1st Qu. 12.000 Median 204.000 Mean 691.500 3rd Qu. 430.000 Max. 3160.000 Std.Dev. 1101.967	0: 4 1: 9	3: 3 4: 10	Not Meaningful	0: 2 5: 1: 9 6: 2: 1 7: 1 3: 3 8: 4: 4 9:	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Performance.rating ==3	Min. 59.000 1st Qu. 62.750 Median 66.500 Mean 66.500 3rd Qu. 70.250 Max. 74.000 Std.Dev. 10.607	Min. 74.000 1st Qu. 80.250 Median 86.500 Mean 86.500 3rd Qu. 92.750 Max. 99.000 Std.Dev. 17.678	0: 0 1: 0 2: 0 3: 2	0: 2 1: 0	8: 17: 9: 18: 10: 19: 11: 20: 12: 21: 13: 1 22: 14: 1 23: 15: 2 24: 16: 3 30:	4: 4 10: 5: 1 11: 6: 12: 7: 13: 8: 14: 9: 18:	0: 1 1: 1 2: 2 3: 3 4: 4	Min. 112.000 1st Qu. 115.800 Median 119.500 Mean 119.500 3rd Qu. 123.200 Max. 127.000 Std.Dev. 10.607
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostactic.acid.phosphatase	Bone.metastases	Stage	Observation	Surv.Stat	
	Min. 2.000 1st Qu. 5.750 Median 9.500 Mean 9.500 3rd Qu. 13.250 Max. 17.000 Std.Dev. 10.607	5: 1 11: 1 6: 12: 1 7: 13: 1 8: 14: 9: 15: 10: 1	Min. 37.000 1st Qu. 2528.000 Median 5018.000 Mean 5018.000 3rd Qu. 7508.000 Max. 9999.000 Std.Dev. 7044.198	0: 1: 2	3: 4: 2	Not Meaningful	0: 2 5: 1: 2 6: 2: 7: 3: 8: 4: 9:	

Table 99: Byar and Green(1980)의 전립선암 데이터 요약

	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Cardiovascular.disease.history == 0	Min. 49.000 1st Qu. 68.000 Median 72.000 Mean 70.520 3rd Qu. 75.000 Max. 87.000 Std.Dev. 7.201	Min. 69.000 1st Qu. 90.000 Median 98.000 Mean 98.320 3rd Qu. 106.250 Max. 145.000 Std.Dev. 13.178	0: 249 1: 10 2: 7 3: 2	0: 268 1: 1 2: 21 3: 23 4: 24 5: 30	8: 1 17: 17 9: 2 18: 6 10: 4 19: 10 11: 16 20: 1 12: 43 21: 1 13: 39 22: 1 14: 53 23: 14 15: 44 24: 1 16: 34 30:	4: 2 10: 35 5: 2 11: 5 6: 24 12: 2 7: 58 13: 8: 83 14: 9: 57 18: 10: 113 15: 11: 79 6: 12 12: 29 7: 14 13: 9 8: 4 14: 7 9: 1	0: 113 15: 1: 15 2: 28 3: 15 4: 72 5: 25 6: 6	Min. 59.000 1st Qu. 122.800 Median 137.000 Mean 133.900 3rd Qu. 147.000 Max. 175.000 Std.Dev. 19.799
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	SurvStat	
Cardiovascular.disease.history == 1	Min. 5.000 1st Qu. 6.000 Median 7.000 Mean 15.260 3rd Qu. 23.000 Max. 69.000 Std.Dev. 12.692	5: 1 11: 66 6: 4 12: 16 7: 3 13: 48 8: 29 14: 2 9: 65 15: 11 10: 23 23:	Min. 1.000 1st Qu. 5.000 Median 8.000 Mean 144.950 3rd Qu. 38.250 Max. 9999.000 Std.Dev. 784.587	0: 218 1: 50	3: 143 4: 125	Not Meaningful	0: 99 5: 14 1: 79 6: 12 2: 29 7: 14 3: 9 8: 4 4: 7 9: 1	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Cardiovascular.disease.history == 1	Min. 48.000 1st Qu. 71.000 Median 74.000 Mean 72.900 3rd Qu. 76.000 Max. 89.000 Std.Dev. 6.306	Min. 71.000 1st Qu. 90.500 Median 98.000 Mean 99.920 3rd Qu. 107.500 Max. 152.000 Std.Dev. 13.528	0: 179 1: 22 2: 6 3:	0: 207 1: 1 2: 1 3: 1	8: 17 16 9: 1 18: 10 10: 9 19: 5 11: 10 20: 1 12: 17 21: 1 13: 31 22: 3 14: 37 23: 1 15: 26 24: 1 16: 37 30: 1	4: 2 10: 27 5: 3 11: 4 6: 16 12: 3 7: 41 13: 1 8: 72 14: 1 9: 36 18: 1	0: 48 1: 8 2: 22 3: 10 4: 73 5: 45 6: 1	Min. 82.000 1st Qu. 122.500 Median 136.000 Mean 134.600 3rd Qu. 147.000 Max. 182.000 Std.Dev. 18.869
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	SurvStat	
Bone.metastases == 0	Min. 0.000 1st Qu. 5.000 Median 9.000 Mean 13.020 3rd Qu. 18.000 Max. 62.000 Std.Dev. 11.525	5: 2 11: 44 6: 4 12: 10 7: 3 13: 22 8: 37 14: 3 9: 67 15: 5 10: 10 10:	Min. 1.000 1st Qu. 4.000 Median 7.000 Mean 100.800 3rd Qu. 18.500 Max. 353.5000 Std.Dev. 372.909	0: 180 1: 27	3: 130 4: 77	Not Meaningful	0: 38 5: 10 1: 42 6: 4 2: 64 7: 13 3: 22 8: 2 4: 7 9: 5	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Bone.metastases == 0	Min. 48.000 1st Qu. 70.000 Median 73.000 Mean 71.777 3rd Qu. 76.000 Max. 89.000 Std.Dev. 6.656	Min. 72.000 1st Qu. 91.000 Median 99.000 Mean 100.100 3rd Qu. 108.000 Max. 152.000 Std.Dev. 13.203	0: 371 1: 23 2: 4 3:	0: 218 1: 180	8: 17 27 9: 3 18: 13 10: 10 19: 11 11: 23 20: 1 12: 47 21: 2 13: 62 22: 2 14: 68 23: 1 15: 63 24: 1 16: 63 30: 1	4: 3 10: 51 5: 3 11: 9 6: 27 12: 4 7: 86 13: 1 8: 133 14: 1 9: 79 18: 1	0: 142 1: 17 2: 37 3: 22 4: 118 5: 61 6: 1	Min. 59.000 1st Qu. 125.200 Median 138.000 Mean 136.800 3rd Qu. 148.000 Max. 182.000 Std.Dev. 17.858
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	SurvStat	
Bone.metastases == 1	Min. 0.000 1st Qu. 4.000 Median 9.000 Mean 12.920 3rd Qu. 18.750 Max. 69.000 Std.Dev. 11.455	5: 2 11: 81 6: 8 12: 20 7: 6 13: 44 8: 65 14: 3 9: 132 15: 8 10: 29 10:	Min. 1.000 1st Qu. 5.000 Median 7.000 Mean 39.850 3rd Qu. 14.500 Max. 353.5000 Std.Dev. 199.774	0: 398 1:	3: 272 4: 126	Not Meaningful	0: 127 5: 23 1: 77 6: 16 2: 81 7: 24 3: 26 8: 5 4: 13 9: 6	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Bone.metastases == 1	Min. 49.000 1st Qu. 68.000 Median 72.000 Mean 70.430 3rd Qu. 76.000 Max. 87.000 Std.Dev. 8.105	Min. 69.000 1st Qu. 84.000 Median 93.000 Mean 93.480 3rd Qu. 102.000 Max. 123.000 Std.Dev. 12.746	0: 57 1: 9 2: 9 3:	0: 50 1: 1 2: 27	8: 1 17: 6 9: 2 18: 3 10: 3 19: 1 11: 3 20: 1 12: 13 21: 2 13: 8 22: 1 14: 22 23: 1 15: 7 24: 1 16: 8 30:	4: 1 10: 11 5: 2 11: 1 6: 13 12: 1 7: 13 13: 1 8: 22 14: 1 9: 14 18:	0: 19 1: 6 2: 13 3: 3 4: 1	Min. 70.000 1st Qu. 105.000 Median 123.000 Mean 121.000 3rd Qu. 137.000 Max. 160.000 Std.Dev. 21.579
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	SurvStat	
Stage == 3	Min. 2.000 1st Qu. 10.000 Median 19.000 Mean 21.340 3rd Qu. 30.000 Max. 62.000 Std.Dev. 13.716	5: 1 11: 29 6: 8 12: 6 7: 7 13: 26 8: 1 14: 2 9: 15: 8 10: 4	Min. 4.000 1st Qu. 20.000 Median 93.000 Mean 569.600 3rd Qu. 385.000 Max. 9999.000 Std.Dev. 1447.698	0: 398 1: 77	3: 1 4: 76	Not Meaningful	0: 10 5: 1 1: 44 6: 1 2: 12 7: 3 3: 5 8: 1 4: 1 9:	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Stage == 3	Min. 49.000 1st Qu. 70.000 Median 73.000 Mean 71.900 3rd Qu. 76.000 Max. 89.000 Std.Dev. 6.490	Min. 72.000 1st Qu. 91.000 Median 99.000 Mean 100.200 3rd Qu. 109.000 Max. 152.000 Std.Dev. 13.094	0: 255 1: 15 2: 3 3:	0: 143 1: 130	8: 1 17: 19 9: 2 18: 9 10: 6 19: 9 11: 3 20: 1 12: 31 21: 1 13: 51 22: 2 14: 45 23: 1 15: 38 24: 1 16: 40 30: 1	4: 3 10: 38 5: 2 11: 6 6: 17 12: 3 7: 63 13: 1 8: 90 14: 1 9: 48 18: 1	0: 95 1: 14 2: 21 3: 16 4: 81 5: 45 6: 1	Min. 59.000 1st Qu. 125.000 Median 138.000 Mean 137.200 3rd Qu. 149.000 Max. 182.000 Std.Dev. 18.519
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	SurvStat	
Stage == 4	Min. 0.000 1st Qu. 4.000 Median 8.000 Mean 11.420 3rd Qu. 16.000 Max. 69.000 Std.Dev. 10.294	5: 2 11: 34 6: 8 12: 6 7: 6 13: 6 8: 65 14: 5 9: 130 15: 16 10: 16	Min. 1.000 1st Qu. 4.000 Median 5.000 Mean 6.689 3rd Qu. 7.000 Max. 297.000 Std.Dev. 17.974	0: 272 1:	3: 273 4:	Not Meaningful	0: 91 5: 19 1: 31 6: 12 2: 63 7: 18 3: 21 8: 2 4: 10 9: 6	
	Age	Weight	Performance.rating	Cardiovascular.disease.history	Systolic.Blood.pressure	Diastolic.blood.pressure	Electrocardiogram.code	Serum.haemoglobin
Stage == 4	Min. 48.000 1st Qu. 70.000 Median 73.000 Mean 71.900 3rd Qu. 76.000 Max. 87.000 Std.Dev. 7.454	Min. 69.000 1st Qu. 89.000 Median 97.000 Mean 97.480 3rd Qu. 105.000 Max. 150.000 Std.Dev. 13.548	0: 173 1: 17 2: 10 3:	0: 125 1: 77	8: 1 17: 14 9: 1 18: 7 10: 7 19: 3 11: 10 20: 1 12: 29 21: 1 13: 19 22: 1 14: 45 23: 1 15: 32 24: 1 16: 31 30:	4: 1 10: 24 5: 3 11: 3 6: 23 22: 2 7: 36 13: 8: 65 14: 9: 45 18: 10: 1	0: 66 1: 9 2: 29 3: 9 4: 64 5: 25 6: 6	Min. 70.000 1st Qu. 118.000 Median 134.000 Mean 130.100 3rd Qu. 144.000 Max. 168.000 Std.Dev. 19.820
	Size.of.primary.tumour	Index.of.tumour.stage.and.histologic.grade	Serum.prostatic.acid.phosphatase	Bone.metastases	Stage	Observation	SurvStat	
	Min. 0.000 1st Qu. 6.250 Median 16.000 Mean 18.160 3rd Qu. 26.000 Max. 62.000 Std.Dev. 13.544	5: 1 11: 76 6: 12: 20 7: 13: 64 8: 1 14: 5 9: 2 15: 16 10: 17	Min. 2.000 1st Qu. 16.000 Median 39.500 Mean 286.600 3rd Qu. 200.000 Max. 9999.000 Std.Dev. 956.902	0: 126 1: 76	3: 202 4:	Not Meaningful	0: 46 5: 5 1: 90 6: 4 2: 30 7: 9 3: 10 8: 4 4: 4 9:	

Table 100: Byar and Green(1980)의 전립선암 데이터 요약(cont'd)

25.2.22 Primary biliary cirrhosis data from Fleming and Harrington(1991) (pbc)

Fleming T. and Harrington D.(1991)[25]에서 소개된 데이터이다.

R의 randomSurvivalForest⁴⁵ 패키지의 pbc 험수로 내장되어 있다.

또한, 아래의 웹사이트를 통해 다운로드할 수도 있다. 다만 변수의 레이블이 R의 PBC와 다소 차이는 있다. 이것만 감안하면 동일한 데이터이다.

<http://www4.stat.ncsu.edu/~boos/var.select/pbc.html>

Data from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data.

Code 23. Byar and Green(1980)의 전립선암 데이터 불러오기(R 코드)

```
1 library("randomForestSRC")
2 data(pbc)
```

25.2.22.1 변수 요약

Instance 수는 418개이며, 총 19개의 변수가 있다. 데이터에 다수의 missing value가 있으므로 주의할 필요가 있다.

- **days**, number of days between registration and the earlier of death, transplantation, or study analysis time in July, 1986
- **status**,
 - 0: alive
 - 1: liver transplant
 - 2: dead
- **treatment**,
 - 1: penicillamine
 - 2: placebo
- **age**, age in days
- **sex**,
 - 0: male
 - 1: female
- **ascities**, presence of ascites: 0=no 1=yes
 - 0: no
 - 1: yes
- **hepatom**, presence of hepatomegaly 0=no 1=yes
 - 0: no
 - 1: yes
- **spiders**, presence of spiders 0=no 1=yes
 - 0: no
 - 1: yes
- **adema**, presence of edema
 - 0: no edema and no diuretic therapy for edema
 - 0.5: edema present without diuretics, or edema resolved by diuretics;
 - 1: 1 = edema despite diuretic therapy
- **bili**, serum bilirubin in mg/dl
- **chol**, serum cholesterol in mg/dl
- **albumin**, albumin in gm/dl
- **copper**, urine copper in ug/day
- **alk**, alkaline phosphatase in U/liter
- **sgot**, SGOT in U/ml
- **trig**, triglycerides in mg/dl
- **platelet**, platelets per cubic ml/1000
- **prothrombin**, prothrombin time in seconds
- **stage**, histologic stage of disease
 - 0: normal activity

⁴⁵2017/12/29일 기준으로. 이 패키지의 원래 이름은 randomSurvivalForest였으나, 바뀌었다.

		days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
All	sho	Min. 41.0000 1st Qu. 109.0000 Median 173.0000 Mean 191.8000 3rd Qu. 261.0000 Max. 479.5000 Std.Dev. 1104.6730	41.0000 1st Qu. 109.0000 Median 173.0000 Mean 191.8000 3rd Qu. 261.0000 Max. 479.5000 Std.Dev. 1104.6730	0: 257 1: 161	1: 158 2: 154 Mean 97.6500 3rd Qu. 123.0000 Max. 588.0000 Std.Dev. 85.6139	9598 1st Qu. 15644 Median 18628 Mean 18533 3rd Qu. 21273 Max. 28650 Std.Dev. 3815.8451	0: 44 1: 374	0: 288 1: 24 NA: 106	0: 152 1: 160 NA: 106	0: 222 1: 90 NA: 106	0: 354 1: 44 2: 20	Min. 0.3000 1st Qu. 0.8000 Median 1.4000 Mean 3.2210 3rd Qu. 3.4000 Max. 28.0000 Std.Dev. 4.4075
		Min. 120.0000 1st Qu. 249.5000 Median 309.5000 Mean 369.5000 3rd Qu. 400.0000 Max. 1775.0000 Std.Dev. 231.9445	1: 9.9000 2: 4.340 Median 3.5300 Mean 3.4070 3rd Qu. 3.7000 Max. 4.6400 Std.Dev. 0.4250	Min. 41.0000 1st Qu. 41.2500 Median 42.5000 Mean 44.0000 3rd Qu. 44.0000 Max. 64.0000 Std.Dev. 0.3711	Min. 289.0000 1st Qu. 871.5000 Median 1259.0000 Mean 1982.7000 3rd Qu. 18000.0000 Max. 13862.4000 Std.Dev. 2140.3888	Min. 26.3500 1st Qu. 80.6000 Median 114.7000 Mean 122.5600 3rd Qu. 151.9000 Max. 457.2500 Std.Dev. 56.6995	Min. 33.0000 1st Qu. 84.2500 Median 108.0000 Mean 124.7000 3rd Qu. 151.0000 Max. 598.0000 Std.Dev. 65.1486	Min. 62.0000 1st Qu. 10.0000 Median 16.0000 Mean 17.7300 3rd Qu. 11.1000 Max. 18.0000 Std.Dev. 1.0220	Min. 9.0000 1st Qu. 10.0000 Median 10.6000 Mean 10.7300 3rd Qu. 11.1000 Max. 18.0000 Std.Dev. NA 6	Min. 0.3000 1st Qu. 0.7000 Median 1.4000 Mean 3.2210 3rd Qu. 3.4000 Max. 28.0000 Std.Dev. 4.4075		
		Min. 553.0000 1st Qu. 1139.0000 Median 2157.0000 Mean 2257.0000 3rd Qu. 2862.0000 Max. 4795.0000 Std.Dev. 1000.2467	0: 257 1: 1	1: 93 2: 94 NA: 70	Min. 3598.0000 1st Qu. 14899.0000 Median 17841.0000 Mean 17806.0000 3rd Qu. 20507.0000 Max. 28650.0000 Std.Dev. 3782.5513	Min. 289.0000 1st Qu. 871.5000 Median 1259.0000 Mean 1982.7000 3rd Qu. 18000.0000 Max. 13862.4000 Std.Dev. NA 108	Min. 26.3500 1st Qu. 80.6000 Median 114.7000 Mean 122.5600 3rd Qu. 151.0000 Max. 598.0000 Std.Dev. NA 106	Min. 33.0000 1st Qu. 84.2500 Median 108.0000 Mean 124.7000 3rd Qu. 151.0000 Max. 598.0000 Std.Dev. NA 106	Min. 62.0000 1st Qu. 10.0000 Median 16.0000 Mean 17.7300 3rd Qu. 11.1000 Max. 18.0000 Std.Dev. NA 11	Min. 9.0000 1st Qu. 10.0000 Median 10.6000 Mean 10.7300 3rd Qu. 11.1000 Max. 18.0000 Std.Dev. 1.0220		
		Min. 4.0000 1st Qu. 35.2500 Median 57.0000 Mean 72.4700 3rd Qu. 83.5000 Max. 444.0000 Std.Dev. 192.6167	Min. 2.3100 1st Qu. 2.3500 Median 3.6000 Mean 3.5830 3rd Qu. 3.8200 Max. 444.0000 Std.Dev. 64.8002	Min. 4.0000 1st Qu. 35.2500 Median 57.0000 Mean 72.4700 3rd Qu. 83.5000 Max. 1104.6000 Std.Dev. 1569.2611	Min. 289.0000 1st Qu. 80.5000 Median 113.2000 Mean 157.8000 3rd Qu. 164.8000 Max. 11046.6000 Std.Dev. 51.7914	Min. 26.3500 1st Qu. 73.1900 Median 98.0000 Mean 114.1000 3rd Qu. 133.3000 Max. 382.0000 Std.Dev. 51.2994	Min. 33.0000 1st Qu. 80.0000 Median 104.0000 Mean 126.0000 3rd Qu. 139.0000 Max. 539.0000 Std.Dev. 91.0355	Min. 76.0000 1st Qu. 9.9000 Median 10.3000 Mean 12.6000 3rd Qu. 10.8000 Max. 18.0000 Std.Dev. NA 2	Min. 9.0000 1st Qu. 9.9000 Median 10.3000 Mean 10.4400 3rd Qu. 10.8000 Max. 18.0000 Std.Dev. 1.0220			
		Min. 41.0000 1st Qu. 99.0000 Median 183.0000 Mean 187.7000 3rd Qu. 207.0000 Max. 419.1000 Std.Dev. 1049.2280	0: 257 1: 1	1: 65 2: 60 NA: 36	Min. 11273.0000 1st Qu. 17000.0000 Median 19540.0000 Mean 19695.0000 3rd Qu. 22388.0000 Max. 28018.0000 Std.Dev. 3584.6469	Min. 15.0000 1st Qu. 17.0000 Median 18.0000 Mean 18.7800 3rd Qu. 21.0000 Max. 21545.0000 Std.Dev. 4020.3673	Min. 21.0000 1st Qu. 22.0000 Median 23.0000 Mean 24.0000 3rd Qu. 25.0000 Max. 29500.0000 Std.Dev. 36.3710	Min. 102 1: 23 NA: 36	0: 37 1: 88 NA: 36	0: 73 1: 52 NA: 36	0: 116 1: 26 2: 60	Min. 0.3000 1st Qu. 1.2000 Median 3.5390 Mean 7.1000 3rd Qu. 7.1000 Max. 28.0000 Std.Dev. 5.8371
		Min. 127.0000 1st Qu. 257.5000 Median 339.0000 Mean 415.8000 3rd Qu. 454.0000 Max. 1775.0000 Std.Dev. 257.0308	Min. 1.9600 1st Qu. 3.0800 Median 3.4000 Mean 3.3610 3rd Qu. 3.6500 Max. 4.5200 Std.Dev. 0.4687	Min. 13.0000 1st Qu. 35.5000 Median 111.0000 Mean 135.4000 3rd Qu. 199.2000 Max. 588.0000 Std.Dev. 2677.1086	Min. 516.0000 1st Qu. 102.0000 Median 164.0000 Mean 259.0000 3rd Qu. 246.0000 Max. 338.0000 Std.Dev. 58.3795	Min. 28.3800 1st Qu. 99.3300 Median 124.8500 Mean 141.9300 3rd Qu. 176.7000 Max. 598.0000 Std.Dev. 79.2582	Min. 49.0000 1st Qu. 91.0000 Median 122.0000 Mean 140.5000 3rd Qu. 171.0000 Max. 598.0000 Std.Dev. 107.8762	Min. 62.0000 1st Qu. 10.4700 Median 12.2000 Mean 14.2000 3rd Qu. 11.8000 Max. 15.2000 Std.Dev. 1.0490	Min. 9.0000 1st Qu. 9.9000 Median 10.3000 Mean 10.4400 3rd Qu. 10.8000 Max. 18.0000 Std.Dev. NA 1	Min. 0.3000 1st Qu. 1.2000 Median 3.5390 Mean 7.1000 3rd Qu. 7.1000 Max. 28.0000 Std.Dev. 5.8371		
		Min. 41.0000 1st Qu. 123.0000 Median 201.6000 Mean 216.0000 3rd Qu. 2632.0000 Max. 4556.0000 Std.Dev. 1094.1233	0: 93 1: 65	1: 58 2: 60 NA: 106	Min. 9598.0000 1st Qu. 15688.0000 Median 18938.0000 Mean 18781.0000 3rd Qu. 21545.0000 Max. 28650.0000 Std.Dev. 4020.3673	Min. 13.0000 1st Qu. 35.5000 Median 111.0000 Mean 135.4000 3rd Qu. 199.2000 Max. 588.0000 Std.Dev. 2677.1086	Min. 28.3800 1st Qu. 99.3300 Median 124.8500 Mean 141.9300 3rd Qu. 171.0000 Max. 598.0000 Std.Dev. 79.2582	Min. 49.0000 1st Qu. 91.0000 Median 122.0000 Mean 140.5000 3rd Qu. 171.0000 Max. 598.0000 Std.Dev. 107.8762	Min. 62.0000 1st Qu. 10.4700 Median 12.2000 Mean 14.2000 3rd Qu. 11.8000 Max. 15.2000 Std.Dev. NA 4	Min. 0.3000 1st Qu. 0.8000 Median 1.4000 Mean 2.8730 3rd Qu. 3.0000 Max. 2.0000 Std.Dev. 3.6289		
		Min. 127.0000 1st Qu. 247.8000 Median 315.0000 Mean 365.0000 3rd Qu. 417.0000 Max. 1712.0000 Std.Dev. 209.5439	Min. 2.1000 1st Qu. 3.2120 Median 3.5650 Mean 3.5160 3rd Qu. 3.8300 Max. 4.6400 Std.Dev. 90.5901	Min. 9.0000 1st Qu. 40.0000 Median 73.0000 Mean 97.6400 3rd Qu. 121.0000 Max. 588.0000 Std.Dev. 2183.4358	Min. 369.0000 1st Qu. 840.8000 Median 124.5000 Mean 2021.3000 3rd Qu. 2028.0000 Max. 11552.0000 Std.Dev. 54.5412	Min. 26.3500 1st Qu. 76.7200 Median 111.6000 Mean 120.2100 3rd Qu. 151.5100 Max. 383.0000 Std.Dev. 71.5391	Min. 33.0000 1st Qu. 83.7800 Median 125.4000 Mean 142.9700 3rd Qu. 151.9000 Max. 457.2500 Std.Dev. 58.5211	Min. 62.0000 1st Qu. 189.5000 Median 224.0000 Mean 242.5000 3rd Qu. 322.0000 Max. 563.0000 Std.Dev. 100.3247	Min. 9.0000 1st Qu. 10.0300 Median 12.6000 Mean 16.5000 3rd Qu. 11.0000 Max. 14.1000 Std.Dev. 0.8514	Min. 0.3000 1st Qu. 1.2000 Median 3.5390 Mean 7.1000 3rd Qu. 7.1000 Max. 28.0000 Std.Dev. 5.8371		
		Min. 51.0000 1st Qu. 1153.0000 Median 181.0000 Mean 1997.0000 3rd Qu. 2771.0000 Max. 4523.0000 Std.Dev. 106.0000	0: 94 1: 60	1: 58 2: 154 NA: 106	Min. 11167.0000 1st Qu. 15134.0000 Median 17733.0000 Mean 17453.0000 3rd Qu. 20383.0000 Max. 2720.0000 Std.Dev. 3637.1012	Min. 15.0000 1st Qu. 18988.0000 Median 20838.0000 Mean 21545.0000 3rd Qu. 21880.0000 Max. 28018.0000 Std.Dev. 3573.4334	Min. 21.0000 1st Qu. 22.0000 Median 23.0000 Mean 24.0000 3rd Qu. 25.0000 Max. 29500.0000 Std.Dev. 36.3710	0: 21 1: 137	0: 144 1: 14	0: 85 1: 73	0: 113 1: 45 NA: 106	0: 132 1: 16 2: 35
		Min. 120.0000 1st Qu. 254.2400 Median 303.5000 Mean 373.9000 3rd Qu. 377.0000 Max. 1775.0000 Std.Dev. 252.4846	Min. 1.9600 1st Qu. 3.3420 Median 3.5450 Mean 3.5240 3rd Qu. 3.7770 Max. 4.8000 Std.Dev. 0.3958	Min. 4.0000 1st Qu. 43.0000 Median 73.0000 Mean 97.6500 3rd Qu. 139.0000 Max. 558.0000 Std.Dev. 2161.0873	Min. 289.0000 1st Qu. 84.8000 Median 124.5000 Mean 194.0000 3rd Qu. 149.8000 Max. 13862.4000 Std.Dev. 58.9313	Min. 28.3800 1st Qu. 83.7800 Median 117.4000 Mean 124.9700 3rd Qu. 151.9000 Max. 383.0000 Std.Dev. 58.5211	Min. 44.0000 1st Qu. 84.5000 Median 113.0000 Mean 125.3000 3rd Qu. 155.0000 Max. 598.0000 Std.Dev. 71.0000	Min. 62.0000 1st Qu. 10.0300 Median 12.6000 Mean 16.5000 3rd Qu. 11.0000 Max. 14.1000 Std.Dev. 0.8514	Min. 9.0000 1st Qu. 10.0300 Median 12.6000 Mean 16.5000 3rd Qu. 11.0000 Max. 14.1000 Std.Dev. NA 106	Min. 0.3000 1st Qu. 0.7250 Median 1.3000 Mean 3.4000 3rd Qu. 4.6000 Max. 28.0000 Std.Dev. 5.2819		
		Min. 41.0000 1st Qu. 99.0000 Median 157.0000 Mean 167.0000 3rd Qu. 242.0000 Max. 479.0000 Std.Dev. 1008.5589	0: 70 1: 36	1: 58 2: 154 NA: 106	Min. 12053.0000 1st Qu. 16802.0000 Median 19580.0000 Mean 19310.0000 3rd Qu. 22880.0000 Max. 27394.0000 Std.Dev. 3573.4334	Min. 15.0000 1st Qu. 35.5000 Median 111.0000 Mean 135.4000 3rd Qu. 199.2000 Max. 588.0000 Std.Dev. 2677.1086	Min. 21.0000 1st Qu. 22.0000 Median 23.0000 Mean 24.0000 3rd Qu. 25.0000 Max. 29500.0000 Std.Dev. 36.3710	0: 8 1: 98	0: 67 1: 87	0: 109 1: 45	0: 131 1: 13 2: 35	Min. 0.3000 1st Qu. 0.7250 Median 1.3000 Mean 3.4000 3rd Qu. 4.6000 Max. 28.0000 Std.Dev. NA 106
		Min. 120.0000 1st Qu. 254.2400 Median 303.5000 Mean 373.9000 3rd Qu. 377.0000 Max. 1775.0000 Std.Dev. 252.4846	Min. 2.3100 1st Qu. 3.1250 Median 3.4700 Mean 3.4400 3rd Qu. 3.7200 Max. 4.5200 Std.Dev. 0.4948	Min. 4.0000 1st Qu. 43.0000 Median 73.0000 Mean 97.6500 3rd Qu. 139.0000 Max. 558.0000 Std.Dev. 106.0654	Min. 289.0000 1st Qu. 84.8000 Median 124.5000 Mean 194.0000 3rd Qu. 149.8000 Max. 10397.0000 Std.Dev. 2418.4462	Min. 28.3800 1st Qu. 83.7800 Median 117.4000 Mean 124.9700 3rd Qu. 151.9000 Max. 457.2500 Std.Dev. 58.5211	Min. 44.0000 1st Qu. 91.0000 Median 121.6700 Mean 133.4000 3rd Qu. 163.2500 Max. 424.0000 Std.Dev. 58.5211	Min. 62.0000 1st Qu. 10.0300 Median 12.6000 Mean 16.5000 3rd Qu. 11.0000 Max. 14.1000 Std.Dev. 0.8514	Min. 9.0000 1st Qu. 10.0300 Median 12.6000 Mean 16.5000 3rd Qu. 11.0000 Max. 14.1000 Std.Dev. NA 106	Min. 0.3000 1st Qu. 0.7250 Median 1.3000 Mean 3.4000 3rd Qu. 4.6000 Max. 28.0000 Std.Dev. 5.2819		
		Min. 140.0000 1st Qu. 1061.0000 Median 1507.0000 Mean 1610.0000 3rd Qu. 2176.0000 Max. 4459.0000 Std.Dev. 1213.0344	0: 20 1: 24	1: 21 2: 15 NA: 8	Min. 12227.0000 1st Qu. 17886.0000 Median 19724.0000 Mean 20318.0000 3rd Qu. 23589.0000 Max. 28650.0000 Std.Dev. 4009.6417	Min. 15.0000 1st Qu. 43.0000 Median 123.0000 Mean 183.0000 3rd Qu. 238.0000 Max. 588.0000 Std.Dev. 106.0654	Min. 21.0000 1st Qu. 22.0000 Median 23.0000 Mean 24.0000 3rd Qu. 25.0000 Max. 29500.0000 Std.Dev. 36.3710	0: 33 1: 3	0: 15 1: 21	0: 32 1: 4	0: 36 1: 5	Min. 0.6000 1st Qu. 1.3000 Median 2.0500 Mean 3.0566 3rd Qu. 3.5000 Max. 5.9000 Std.Dev. 2.3195
		Min. 151.0000 1st Qu. 245.0000 Median 317.0000 Mean 362.5000 3rd Qu. 426.5000 Max. 1000.0000 Std.Dev. 178.9933	Min. 2.2700 1st Qu. 3.3280 Median 3.6450 Mean 3.5350 3rd Qu. 3.7680 Max. 4.3000 Std.Dev. 0.4566	Min. 13.0000 1st Qu. 43.0000 Median 111.0000 Mean 151.2800 3rd Qu. 211.7500 Max. 444.0000 Std.Dev. 106.0654	Min. 289.0000 1st Qu. 84.8000 Median 124.5000 Mean 194.0000 3rd Qu. 149.8000 Max. 10397.0000 Std.Dev. 2418.4462	Min. 26.3500 1st Qu. 83.7800 Median 117.4000 Mean 124.9700 3rd Qu. 151.9000 Max. 457.2500 Std.Dev. 58.5211	Min. 49.0000 1st Qu. 91.0000 Median 121.6700 Mean 133.4000 3rd Qu. 163.2500 Max. 424.0000 Std.Dev. 58.5211	Min. 70.0000 1st Qu. 165.5000 Median 217.0000 Mean 231.0000 3rd Qu. 298.5000 Max. 394.0000 Std.Dev. 58.5211	Min. 9.0000 1st Qu. 10.0300 Median 12.6000 Mean 16.5000 3rd Qu. 11.0000 Max. 14.1000 Std.Dev. 0.8514	Min. 0.3000 1st Qu. 0.7250 Median 1.3000 Mean 3.4000 3rd Qu. 4.6000 Max. 28.0000 Std.Dev. 5.2819		
		Min. 41.0000 1st Qu. 1096.0000 Median 1921.0000 Mean 2066.0000 3rd Qu. 2716.0000 Max. 4795.0000 Std.Dev. 1092.9532	0: 237 1: 137	1: 137 2: 139 NA: 98	Min. 9598.0000 1st Qu. 15479.0000 Median 18333.0000 Mean 1820.0000 3rd Qu. 20819.0000 Max. 28650.0000 Std.Dev. 3740.3974	Min. 15.0000 1st Qu. 40.0000 Median 87.5000 Mean 195.8000 3rd Qu. 196.3000 Max. 457.2500 Std.Dev. 87.5000	Min. 21.0000 1st Qu. 22.0000 Median 23.0000 Mean 24.0000 3rd Qu. 25.0000 Max. 29500.0000 Std.Dev. 36.3710	0: 255 1: 21	0: 137 1: 139	0: 190 1: 86		

	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili
ascites==0	Min. 71.0000 1st Quartile 1299.0000 Median 1959.0000 Mean 2106.0000 3rd Quartile 2778.0000 Max. 4556.0000 Std.Dev. 1081.5422	0: 186 1: 102 NA 106	1: 144 2: 144 NA: 106	Min. 9598.0000 1st Quartile 15118.0000 Median 17911.0000 Mean 18011.0000 3rd Quartile 20508.0000 Max. 27398.0000 Std.Dev. 3761.2515 NA 106	0: 33 1: 255 NA: 106	0: 288 1: 141 NA: 106	0: 147 1: 141 NA: 106	0: 211 1: 77 NA: 106	0: 257 0.5: 25 1: 6 NA: 106	Min. 0.3000 1st Quartile 0.7000 Median 1.3000 Mean 2.7620 3rd Quartile 3.2000 Max. 28.0000 Std.Dev. 3.7856 NA 106
	chol	albumin	copper	alk	sgot	ldl	platelet	prothrombin	stage	
	Min. 120.0000 1st Quartile 253.0000 Median 315.0000 Mean 373.3000 3rd Quartile 406.0000 Max. 1775.0000 Std.Dev. 233.3072 NA 131	Min. 1.9600 1st Quartile 3.3500 Median 3.5700 Mean 3.5670 3rd Quartile 3.8220 Max. 4.6400 Std.Dev. 0.3812 NA 106	0: 4.0000 1st Quartile 4.0000 Median 4.2500 Mean 4.2500 3rd Quartile 4.5000 Max. 5.8800 Std.Dev. 0.3812 NA 107	Min. 289.0000 1st Quartile 857.8000 Median 1250.0000 Mean 1976.1000 3rd Quartile 1968.8000 Max. 13862.4000 Std.Dev. 2169.4860 NA 106	Min. 26.3500 1st Quartile 79.7500 Median 113.1500 Mean 121.0500 3rd Quartile 150.3500 Max. 457.2500 Std.Dev. 55.6339 NA 106	Min. 33.0000 1st Quartile 84.0000 Median 106.5000 Mean 121.0000 3rd Quartile 146.0000 Max. 382.0000 Std.Dev. 36.3025 NA 132	Min. 70.0000 1st Quartile 206.0000 Median 265.0000 Mean 267.9000 3rd Quartile 326.2000 Max. 563.0000 Std.Dev. 93.5341 NA 110	Min. 70.0000 1st Quartile 206.0000 Median 265.0000 Mean 267.9000 3rd Quartile 326.2000 Max. 563.0000 Std.Dev. 93.5341 NA 110	Min. 70.0000 1st Quartile 206.0000 Median 265.0000 Mean 267.9000 3rd Quartile 326.2000 Max. 563.0000 Std.Dev. 93.5341 NA 110	Min. 0.3000 1st Quartile 0.7000 Median 1.3000 Mean 2.7620 3rd Quartile 3.2000 Max. 28.0000 Std.Dev. 3.7856 NA 106
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili
	ascites==1	Min. 41.0000 1st Quartile 188.0000 Median 368.0000 Mean 813.2000 3rd Quartile 1197.5000 Max. 3222.0000 Std.Dev. 924.9491 NA 106	0: 1 1: 23 NA: 106	1: 14 2: 10 NA: 106	Min. 14317.0000 1st Quartile 18007.0000 Median 21000.0000 Mean 21368.0000 3rd Quartile 24219.0000 Max. 28650.0000 Std.Dev. 3818.9427 NA 106	0: 3 1: 21 NA: 106	0: 5 1: 19 NA: 106	0: 11 1: 13 NA: 106	0: 6 0.5: 4 1: 14	Min. 0.8000 1st Quartile 1.9250 Median 6.8500 Mean 9.1830 3rd Quartile 17.1250 Max. 24.5000 Std.Dev. 7.6795 NA 106
	chol	albumin	copper	alk	sgot	ldl	platelet	prothrombin	stage	
	Min. 151.0000 1st Quartile 200.0000 Median 261.0000 Mean 322.0000 3rd Quartile 344.0000 Max. 1092.0000 Std.Dev. 213.5978 NA 109	Min. 2.1000 1st Quartile 2.6300 Median 3.0200 Mean 2.9550 3rd Quartile 3.2500 Max. 3.9000 Std.Dev. 0.4572 NA 106	Min. 16.0000 1st Quartile 65.5000 Median 150.0000 Mean 206.1000 3rd Quartile 214.5000 Max. 558.0000 Std.Dev. 1791.2643 NA 106	Min. 55.9000 1st Quartile 96.8000 Median 147.8500 Mean 201.0000 3rd Quartile 215.0000 Max. 598.0000 Std.Dev. 66.9293 NA 106	Min. 43.4000 1st Quartile 98.1300 Median 135.9700 Mean 140.6500 3rd Quartile 176.5100 Max. 338.0000 Std.Dev. 122.7073 NA 106	Min. 49.0000 1st Quartile 109.5000 Median 148.5000 Mean 173.4000 3rd Quartile 182.0000 Max. 598.0000 Std.Dev. 128.5840 NA 110	Min. 62.0000 1st Quartile 124.2000 Median 161.0000 Mean 191.2000 3rd Quartile 224.0000 Max. 401.0000 Std.Dev. 93.2694 NA 106	Min. 10.3000 1st Quartile 11.0700 Median 11.6000 Mean 11.8400 3rd Quartile 12.2000 Max. 15.2000 Std.Dev. 1.0741 NA 106	Min. 0.8000 1st Quartile 1.9250 Median 6.8500 Mean 9.1830 3rd Quartile 17.1250 Max. 24.5000 Std.Dev. 7.6795 NA 106	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili
	ascites==NA	Min. 41.0000 1st Quartile 998.0000 Median 1397.0000 Mean 1657.0000 3rd Quartile 2262.0000 Max. 4795.0000 Std.Dev. 1008.5589	0: 70 1: 36	1: NA: 106	Min. 1.2053.0000 1st Quartile 16802.0000 Median 1937.0000 Mean 19310.0000 3rd Quartile 22280.0000 Max. 27394.0000 Std.Dev. 3573.4334	0: 8 1: 98 NA: 106	0: 1: NA: 106	0: 1: NA: 106	0: 91 0.5: 15 1: 1	Min. 0.4000 1st Quartile 0.7250 Median 1.4000 Mean 3.1170 3rd Quartile 3.0750 Max. 18.0000 Std.Dev. 4.0429
	chol	albumin	copper	alk	sgot	ldl	platelet	prothrombin	stage	
	Min. 1.2100 1st Quartile 3.1250 Median 3.4700 Mean 3.4310 3rd Quartile 3.7200 Max. 4.5200 Std.Dev. 0.4348 NA 106	Min. 2.3100 1st Quartile 3.1250 Median 3.4700 Mean 3.4310 3rd Quartile 3.7200 Max. 4.5200 Std.Dev. 0.4348 NA 106	Min. 16.0000 1st Quartile 65.5000 Median 150.0000 Mean 206.1000 3rd Quartile 214.5000 Max. 558.0000 Std.Dev. 1791.2643 NA 106	Min. 55.9000 1st Quartile 96.8000 Median 147.8500 Mean 201.0000 3rd Quartile 215.0000 Max. 598.0000 Std.Dev. 122.7073 NA 106	Min. 43.4000 1st Quartile 98.1300 Median 135.9700 Mean 140.6500 3rd Quartile 176.5100 Max. 338.0000 Std.Dev. 128.5840 NA 110	Min. 62.0000 1st Quartile 124.2000 Median 161.0000 Mean 191.2000 3rd Quartile 224.0000 Max. 401.0000 Std.Dev. 93.2694 NA 106	Min. 9.0000 1st Quartile 10.1000 Median 10.6000 Mean 10.7500 3rd Quartile 11.0000 Max. 18.0000 Std.Dev. 1.0781 NA 2	Min. 0.4000 1st Quartile 0.7250 Median 1.4000 Mean 3.1170 3rd Quartile 3.0750 Max. 18.0000 Std.Dev. 4.0429		
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili
	hepatom==0	Min. 41.0000 1st Quartile 1434.0000 Median 2338.0000 Mean 2338.0000 3rd Quartile 3092.0000 Max. 4556.0000 Std.Dev. 1091.7526 NA 106	0: 115 1: 37	1: 85 2: 67 NA: 106	Min. 10550.0000 1st Quartile 15008.0000 Median 16500.0000 Mean 17971.0000 3rd Quartile 20601.0000 Max. 28018.0000 Std.Dev. 3932.4003 NA 106	0: 15 1: 137 NA: 106	0: 147 1: 5 NA: 106	0: 152 1: 1 NA: 106	0: 138 0.5: 15 1: 1	Min. 0.3000 1st Quartile 0.6000 Median 1.1000 Mean 1.8569 3rd Quartile 2.9000 Max. 22.5000 Std.Dev. 2.7738 NA 106
	chol	albumin	copper	alk	sgot	ldl	platelet	prothrombin	stage	
	Min. 120.0000 1st Quartile 239.0000 Median 298.0000 Mean 336.5000 3rd Quartile 390.0000 Max. 1276.0000 Std.Dev. 157.6984 NA 121	Min. 2.1000 1st Quartile 3.4280 Median 3.6600 Mean 3.6490 3rd Quartile 3.8950 Max. 4.6400 Std.Dev. 0.3828 NA 106	Min. 4.0000 1st Quartile 33.0000 Median 58.0000 Mean 171.4000 3rd Quartile 94.5000 Max. 464.0000 Std.Dev. 70.4553 NA 106	Min. 369.0000 1st Quartile 789.0000 Median 117.5000 Mean 174.1000 3rd Quartile 1689.8000 Max. 12258.8000 Std.Dev. 1958.0401 NA 106	Min. 28.38 1st Quartile 74.30 Median 101.5300 Mean 114.6000 3rd Quartile 137.9500 Max. 338.0000 Std.Dev. 55.0385 NA 106	Min. 44.0000 1st Quartile 109.5000 Median 123.0000 Mean 140.8000 3rd Quartile 140.8000 Max. 514.0000 Std.Dev. 53.2374 NA 107	Min. 62.0000 1st Quartile 214.5000 Median 273.0000 Mean 280.5000 3rd Quartile 336.0000 Max. 514.0000 Std.Dev. 90.4001 NA 107	Min. 9.0000 1st Quartile 10.0000 Median 10.6000 Mean 10.7500 3rd Quartile 11.0000 Max. 18.0000 Std.Dev. 0.9813 NA 106	Min. 0.3000 1st Quartile 0.6000 Median 1.1000 Mean 1.8569 3rd Quartile 2.9000 Max. 22.5000 Std.Dev. 2.7738 NA 106	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili
	hepatom==1	Min. 71.0000 1st Quartile 90.0000 Median 155.0000 Mean 1691.4000 3rd Quartile 2294.8000 Max. 4523.0000 Std.Dev. 1063.5383 NA 106	0: 72 1: 88	1: 73 2: 67 NA: 106	Min. 9598.0000 1st Quartile 14600.0000 Median 1882.0000 Mean 1853.0000 3rd Quartile 20848.0000 Max. 28650.0000 Std.Dev. 3790.0097 NA 106	0: 21 1: 139 NA: 106	0: 141 1: 19 NA: 106	0: 93 1: 67 NA: 106	0: 125 0.5: 20 1: 15 NA: 106	Min. 0.4000 1st Quartile 1.1000 Median 2.5500 Mean 4.5869 3rd Quartile 5.8000 Max. 28.0000 Std.Dev. 5.4021 NA 106
	chol	albumin	copper	alk	sgot	ldl	platelet	prothrombin	stage	
	Min. 149.0000 1st Quartile 255.0000 Median 322.0000 Mean 400.4000 3rd Quartile 420.0000 Max. 1775.0000 Std.Dev. 281.3131 NA 119	Min. 1.9600 1st Quartile 3.1350 Median 3.4350 Mean 3.3980 3rd Quartile 3.6720 Max. 4.2400 Std.Dev. 0.4181 NA 106	Min. 11.0000 1st Quartile 52.0000 Median 88.0000 Mean 117.2000 3rd Quartile 159.0000 Max. 588.0000 Std.Dev. 94.0018 NA 106	Min. 289.0000 1st Quartile 96.5000 Median 1417.0000 Mean 2212.3000 3rd Quartile 163.7100 Max. 457.2500 Std.Dev. 2282.7452 NA 106	Min. 26.5300 1st Quartile 89.9000 Median 121.0800 Mean 130.0600 3rd Quartile 157.4017 Max. 598.0000 Std.Dev. 73.3123 NA 120	Min. 33.0000 1st Quartile 91.0000 Median 117.5000 Mean 134.8000 3rd Quartile 129.0000 Max. 598.0000 Std.Dev. 97.3666 NA 109	Min. 70.0000 1st Quartile 166.0000 Median 234.0000 Mean 244.1000 3rd Quartile 298.0000 Max. 563.0000 Std.Dev. 0.9945 NA 106	Min. 9.2000 1st Quartile 10.0000 Median 10.9000 Mean 10.9100 3rd Quartile 11.5000 Max. 15.2000 Std.Dev. 0.9945 NA 106	Min. 0.4000 1st Quartile 0.7250 Median 1.1000 Mean 3.1170 3rd Quartile 3.0750 Max. 18.0000 Std.Dev. 4.0429 NA 106	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili
	hepatom==NA	Min. 41.0000 1st Quartile 99.9800 Median 1397.0000 Mean 1657.0000 3rd Quartile 2262.0000 Max. 4795.0000 Std.Dev. 1008.5589	0: 70 1: 36	1: NA: 106	Min. 1.2053.0000 1st Quartile 16802.0000 Median 19358.0000 Mean 19310.0000 3rd Quartile 22280.0000 Max. 27394.0000 Std.Dev. 3573.4334	0: 8 1: 98 NA: 106	0: 1: NA: 106	0: 1: NA: 106	0: 91 0.5: 15 1: 2	Min. 0.4000 1st Quartile 0.7250 Median 1.1000 Mean 3.1170 3rd Quartile 3.0750 Max. 18.0000 Std.Dev. 4.0429 NA 106
	chol	albumin	copper	alk	sgot	ldl	platelet	prothrombin	stage	
	Min. 1.06 1st Quartile 1.06 Median 1.06 Mean 1.06 3rd Quartile 1.06 Max. 1.06 Std.Dev. 1.06	Min. 1.06 1st Quartile 1.06 Median 1.06 Mean 1.06 3rd Quartile 1.06 Max. 1.06 Std.Dev. 1.06	Min. 1.06 1st Quartile 1.06 Median 1.06 Mean 1.06 3rd Quartile 1.06 Max. 1.06 Std.Dev. 1.06	Min. 1.06 1st Quartile 1.06 Median 1.06 Mean 1.06 3rd Quartile 1.06 Max. 1.06 Std.Dev. 1.06	Min. 1.06 1st Quartile 1.06 Median 1.06 Mean 1.06 3rd Quartile 1.06 Max. 1.06 Std.Dev. 1.06	Min. 1.06 1st Quartile 1.06 Median 1.06 Mean 1.06 3rd Quartile 1.06 Max. 1.06 Std.Dev. 1.06	Min. 1.06 1st Quartile 1.06 Median 1.06 Mean 1.06 3rd Quartile 1.06 Max. 1.06 Std.Dev. 1.06	Min. 1.06 1st Quartile 1.06 Median 1.06 Mean 1.06 3rd Quartile 1.06 Max. 1.06 Std.Dev. 1.06	Min. 1.06 1st Quartile 1.06 Median 1.06 Mean 1.06 3rd Quartile 1.06 Max. 1.06 Std.Dev. 1.06	Min. 0.4000 1st Quartile 0.7250 Median 1.1000 Mean 3.1170 3rd Quartile 3.0750 Max. 18.0000 Std.Dev. 4.0429 NA 106
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili

Table 102: PBC 테이터 요약(Cont'd)

spiders==0	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
	Min.	41.0000		Min.	10550.0000				Min.	0.3000	
	1st Quart.	1349.0000		1st Quart.	15561.0000			1st Quart.	0.7000		
	Median	2162.0000	0: 149	Median	18393.0000	0: 32	0: 211	0: 222	0: 200	1st Quart.	
	Mean	2197.0000	1: 73	Mean	18393.0000	1: 190	1: 11	1: 93	0.5: 16	Median	
	3rd Quart.	2978.0000	NA: 106	3rd Quart.	20807.0000	NA: 106	NA: 106	NA: 106	1: 6	Mean	
	Max.	4556.0000		Max.	28018.0000				6: 3rd Quart.	2.3000	
	Std.Dev.	3089.0558		Std.Dev.	3874.0140				Max.	28.0000	
	NA	106		NA	106			NA	NA: 106	Std.Dev.	
spiders==1	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
	Min.	120.0000	Min. 2.1000	Min.	289.0000	Min.	62.0000	Min.	9.0000	Min. 0.5000	
	1st Quart.	256.0000	1st Quart. 3.3500	1st Quart.	832.2000	1st Quart.	213.0000	1st Quart.	10.0000	1st Quart. 1.3250	
	Median	316.0000	Median 3.6000	Median	1209.0000	Median	110.9500	Median	10.5000	Median 3.2000	
	Mean	360.3000	Mean 3.5820	Mean	1931.1000	Mean	117.8400	Mean	10.5700	Mean 5.0340	
	3rd Quart.	396.0000	3rd Quart. 3.8500	3rd Quart.	1902.0000	3rd Quart.	151.0000	3rd Quart.	11.0000	3rd Quart. 6.4750	
	Max.	1712.0000	Max. 4.6400	Max.	2258.8000	Max.	280.0000	Max.	17.1000	Max. 24.5000	
	Std.Dev.	208.3380	Std.Dev. 0.4035	Std.Dev.	68.8175	Std.Dev.	56.9537	Std.Dev.	92.6310	Std.Dev.	
	NA	126	NA: 107	NA	NA: 106	NA	127	NA	110	NA: 106	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
spiders==1	Min.	51.0000								Min. 0.5000	
	1st Quart.	75.0000								1st Quart. 1.3250	
	Median	1439.0000	0: 38	1: 45						Median 3.2000	
	Mean	1537.0000	1: 52	2: 45						Mean 5.0340	
	3rd Quart.	2206.0000	NA: 106	NA: 106						3rd Quart. 6.4750	
	Max.	4500.0000								Max. 24.5000	
	Std.Dev.	1049.8395								Std.Dev. 5.8380	
	NA	106								NA: 106	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
spiders==NA	Min.	168.0000	Min. 1.9600	Min.	310.0000	Min.	44.0000	Min.	71.0000	Min. 0.4000	
	1st Quart.	243.0000	1st Quart. 3.1200	1st Quart.	98.0000	1st Quart.	87.0000	1st Quart.	10.2000	1st Quart. 0.7250	
	Median	301.0000	Median 3.4150	Median	102.0000	Median	124.2100	Median	230.5000	Median 1.4000	
	Mean	392.2000	Mean 3.3660	Mean	123.9000	Mean	134.1800	Mean	11.1100	Mean 3.1170	
	3rd Quart.	426.0000	3rd Quart. 3.6300	3rd Quart.	172.0000	3rd Quart.	162.1600	3rd Quart.	310.5000	3rd Quart. 3.0750	
	Max.	1775.0000	Max. 4.1920	Max.	13862.0000	Max.	598.0000	Max.	493.0000	Max. 18.0000	
	Std.Dev.	281.9969	Std.Dev. 0.4220	Std.Dev.	110.1741	Std.Dev.	54.6557	Std.Dev.	98.9387	Std.Dev.	
	NA	114	NA: 106	NA	106	NA	115	NA	106	NA: 106	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
spiders==NA	Min.	41.0000								Min. 0.4000	
	1st Quart.	99.0000								1st Quart. 0.7250	
	Median	1397.0000	0: 70	1: 36						Median 1.4000	
	Mean	1657.0000								Mean 3.1170	
	3rd Quart.	2262.0000								3rd Quart. 3.0750	
	Max.	4795.0000								Max. 18.0000	
	Std.Dev.	1008.5589								Std.Dev. 4.0429	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
edema==0	Min.	2.3100	Min. 3.1250	Min.	168.0000	Min.	76.0000	Min.	9.0000	Min. 0.4000	
	1st Quart.	3.1250	1st Quart. 3.4700	1st Quart.	168.0000	1st Quart.	87.0000	1st Quart.	10.1000	1st Quart. 0.7250	
	Median	3.4700	Median 3.4700	Median	168.0000	Median	106.0000	Median	10.6000	Median 1.4000	
	Mean	3.4310	Mean 3.4310	Mean	168.0000	Mean	124.7100	Mean	241.7000	Mean 3.1170	
	3rd Quart.	3.7200	3rd Quart. 4.5200	3rd Quart.	172.0000	3rd Quart.	151.0000	3rd Quart.	228.0000	3rd Quart. 3.0750	
	Max.	4.5200	Max. 4.5200	Max.	2106.0000	Max.	598.0000	Max.	493.0000	Max. 18.0000	
	Std.Dev.	0.4348	Std.Dev. NA: 106	Std.Dev.	NA: 106	Std.Dev.	NA: 106	Std.Dev.	NA: 106	Std.Dev.	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
edema==0	Min.	41.0000								Min. 0.4000	
	1st Quart.	1225.0000								1st Quart. 0.7250	
	Median	1882.0000	0: 238	1: 116	NA: 91	0: 8	0: 257	0: 200	0: 354	Median 1.4000	
	Mean	2045.0000				1: 98	1: 6	1: 125	1: 63	Mean 2.6920	
	3rd Quart.	2709.0000					NA: 91	NA: 91	NA: 91	3rd Quart. 3.2000	
	Max.	4795.0000								Max. 25.5000	
	Std.Dev.	1060.7220								Std.Dev. 3.5760	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
edema==0.5	Min.	120.0000	Min. 1.9600	Min.	289.0000	Min.	76.0000	Min.	9.0000	Min. 0.4000	
	1st Quart.	256.0000	1st Quart. 3.3400	1st Quart.	94.0000	1st Quart.	87.0000	1st Quart.	10.1000	1st Quart. 0.7250	
	Median	316.0000	Median 3.5700	Median	68.5000	Median	113.5000	Median	10.6000	Median 1.4000	
	Mean	379.4000	Mean 3.5457	Mean	89.6800	Mean	107.4000	Mean	24.7500	Mean 3.1170	
	3rd Quart.	404.0000	3rd Quart. 3.8000	3rd Quart.	108.0000	3rd Quart.	139.5000	3rd Quart.	325.0000	3rd Quart. 3.1380	
	Max.	1775.0000	Max. 4.6400	Max.	464.0000	Max.	13862.0000	Max.	432.0000	Max. 10.8000	
	Std.Dev.	241.8929	Std.Dev. 0.4020	Std.Dev.	75.9971	Std.Dev.	2152.9561	Std.Dev.	56.9197	Std.Dev.	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
edema==0.5	Min.	71.0000								Min. 0.4000	
	1st Quart.	607.5000								1st Quart. 1.0750	
	Median	1180.5000	0: 18	1: 26	NA: 15	0: 5	0: 25	0: 9	0: 44	Median 2.0000	
	Mean	1483.0000				1: 39	1: 4	1: 20	1: 13	Mean 4.7300	
	3rd Quart.	2096.0000					NA: 15	NA: 15	NA: 15	3rd Quart. 3.5300	
	Max.	4232.0000								Max. 28.0000	
	Std.Dev.	1097.3448								Std.Dev. 6.3369	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
edema==1	Min.	175.0000	Min. 2.5300	Min.	14.0000	Min.	51.0000	Min.	76.0000	Min. 0.4000	
	1st Quart.	244.0000 <td>1st Quart. 3.1470</td> <th>1st Quart.</th> <td>47.2500</td> <th>1st Quart.</th> <td>84.0000</td> <th>1st Quart.</th> <td>109.5000</td> <th>1st Quart. 0.7070</th>	1st Quart. 3.1470	1st Quart.	47.2500	1st Quart.	84.0000	1st Quart.	109.5000	1st Quart. 0.7070	
	Median	298.0000	Median 3.3900	Median	60.5000	Median	113.5000	Median	155.5000	Median 1.1000	
	Mean	332.3000	Mean 3.3730	Mean	113.4300	Mean	214.0000	Mean	256.0000	Mean 2.8000	
	3rd Quart.	410.2000	3rd Quart. 3.5620	3rd Quart.	182.5000	3rd Quart.	177.0000	3rd Quart.	282.0000	3rd Quart. 3.14	
	Max.	674.0000	Max. 4.0600	Max.	1132.0000	Max.	246.4500	Max.	442.0000	Max. 13.6000	
	Std.Dev.	135.0830	Std.Dev. 0.3680	Std.Dev.	78.7328	Std.Dev.	52.4079	Std.Dev.	97.3187	Std.Dev.	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
edema==1	Min.	41.0000								Min. 0.4000	
	1st Quart.	137.8000	0: 1	1: 19	NA: 15	0: 3	0: 6	0: 5	0: 6	0: 55	1st Quart. 2.4750
	Median	299.0000				1: 17	1: 14	1: 15	1: 14	1: 20	Median 2.8000
	Mean	368.0000								Mean 2.9250	
	3rd Quart.	387.0000								3rd Quart. 3.1750	
	Max.	3428.0000								Max. 22.5000	
	Std.Dev.	796.5307								Std.Dev. 6.9966	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
edema==1	Min.	151.0000	Min. 2.1000	Min.	18.0000	Min.	55.0000	Min.	49.0000	Min. 0.8000	
	1st Quart.	188.0000	1st Quart. 2.5000	1st Quart.	39.5000	1st Quart.	80.8000	1st Quart.	87.0000 <th>1st Quart. 1.1000</th>	1st Quart. 1.1000	
	Median	222.0000	Median 2.9600	Median	45.0000	Median	108.4000	Median	125.0000	Median 7.8000	
	Mean	285.9000	Mean 2.8900	Mean	18.0000	Mean	207.9000	Mean	154.2000	Mean 9.2650	
	3rd Quart.	299.0000	3rd Quart. 3.1550	3rd Quart.	211.8000	3rd Quart.	186.5000	3rd Quart.	231.0000	3rd Quart. 10.1500	
	Max.	932.0000	Max. 3.6700	Max.	58.0000	Max.	693.2000	Max.	401.0000	Max. 22.5000	
	Std.Dev.	186.2010	Std.Dev. 0.4308	Std.Dev.	150.5848	Std.Dev.	2054.5019	Std.Dev.	55.8514	Std.Dev.	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	

Table 103: PBC 데이터 요약(Cont'd)

	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
stage==1	Min.	389.0000		Min.	10550.0000	0:	16	0:	15	0:	20
	1st Quartile	1702.0000		1st Quartile	14060.0000	1:	12	1:	1	0.5:	1
	Median	2644.0000	0: 19	Median	10920.0000	2:	4	1:	1	1:	1
	Mean	2655.0000	1: 2	Mean	17100.0000	1:	3	1:	1	Mean	1.3620
	3rd Quartile	3388.0000	NA: 6	3rd Quartile	19585.0000	NA: 11	NA: 11	NA: 11	NA: 11	3rd Quartile	1.1000
	Max.	4459.0000		Max.	22836.0000					Max.	7.3000
	Std.Dev.	1092.8129		Std.Dev.	3486.5620					Std.Dev.	1.7909
	NA	6		NA	6					NA	6
	clsd	albumin	copper	alk	sgot	trig	platelet	prothrombin	stage		
	Min.	132.0000	Min.	2.8900	Min.	10.0000	Min.	49.6000	Min.	0.5000	
stage==2	1st Quartile	216.0000	1st Quartile	3.5200	1st Quartile	26.5000	1st Quartile	647.5000	1st Quartile	0.6000	
	Median	239.0000	Median	3.7700	Median	64.0000	Median	64.3300	Median	0.8000	
	Mean	267.8000	Mean	3.7050	Mean	62.8100	Mean	1694.2000	Mean	1.3220	
	3rd Quartile	253.0000	3rd Quartile	3.9700	3rd Quartile	74.7500	3rd Quartile	1621.2000	3rd Quartile	1.1000	
	Max.	614.0000	Max.	4.1900	Max.	172.0000	Max.	5890.0000	Max.	7.3000	
	Std.Dev.	120.7765	Std.Dev.	0.4391	Std.Dev.	47.0917	Std.Dev.	1895.1604	Std.Dev.	1.7909	
	NA	14	NA	6	NA	11	NA	11	NA	NA	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
	Min.	130.0000		Min.	11058.0000					Min.	0.3000
	1st Quartile	1440.0000		1st Quartile	15510.0000					1st Quartile	0.6000
stage==3	Median	2410.0000	0: 69	Median	17897.0000					Median	0.9500
	Mean	2390.0000	1: 23	Mean	18067.0000	0: 8				Mean	2.4530
	3rd Quartile	3111.0000	NA: 6	3rd Quartile	20598.0000	1: 84				3rd Quartile	2.1000
	Max.	4795.0000		Max.	2398.0000	NA: 6				Max.	25.5000
	Std.Dev.	1069.0888		Std.Dev.	3515.2810					Std.Dev.	4.1351
	NA	6		NA	6					NA	6
	clsd	albumin	copper	alk	sgot	trig	platelet	prothrombin	stage		
	Min.	127.0000	Min.	2.6400	Min.	4.0000	Min.	37.0000	Min.	0.3000	
	1st Quartile	260.0000	1st Quartile	3.3650	1st Quartile	32.0000	1st Quartile	61.5000	1st Quartile	0.6000	
	Median	280.0000	Median	3.6500	Median	49.5000	Median	141.5000	Median	0.9500	
stage==4	Mean	353.2000	Mean	3.6070	Mean	68.0300	Mean	1816.9000	Mean	2.4530	
	3rd Quartile	408.0000	3rd Quartile	3.8930	3rd Quartile	87.0000	3rd Quartile	1834.0000	3rd Quartile	2.1000	
	Max.	1480.0000	Max.	4.6400	Max.	53.7056	Max.	3382.4000	Max.	18.0000	
	Std.Dev.	185.5826	Std.Dev.	0.3842	Std.Dev.	227.0000	Std.Dev.	2142.8100	Std.Dev.	1.0848	
	NA	37	NA	6	NA	32	NA	31	NA	6	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
	Min.	14.0000		Min.	9598.0000					Min.	0.3000
	1st Quartile	1314.0000		1st Quartile	15034.0000					1st Quartile	0.8000
	Median	1810.0000	0: 107	Median	17947.0000	0: 16				Median	1.3000
	Mean	1997.0000	1: 48	Mean	17884.0000	1: 139				Mean	2.8260
stage==NA	3rd Quartile	2532.0000	NA: 6	3rd Quartile	19953.0000	NA: 6				3rd Quartile	2.9000
	Max.	4500.0000		Max.	26259.0000					Max.	28.0000
	Std.Dev.	973.8381		Std.Dev.	3695.8127					Std.Dev.	4.1943
	NA	6		NA	6					NA	6
	clsd	albumin	copper	alk	sgot	trig	platelet	prothrombin	stage		
	Min.	120.0000	Min.	2.3800	Min.	9.0000	Min.	289.0000	Min.	0.3000	
	1st Quartile	265.0000	1st Quartile	3.3750	1st Quartile	42.0000	1st Quartile	935.2000	1st Quartile	0.6000	
	Median	324.0000	Median	3.6100	Median	67.5000	Median	125.7500	Median	0.9500	
	Mean	415.9000	Mean	3.5920	Mean	92.0800	Mean	2082.1000	Mean	2.4530	
	3rd Quartile	426.5000	3rd Quartile	3.8300	3rd Quartile	114.2500	3rd Quartile	1909.5000	3rd Quartile	2.1000	
stage==NA	Max.	1775.0000	Max.	4.5200	Max.	444.0000	Max.	12258.8000	Max.	14.1000	
	Std.Dev.	292.0084	Std.Dev.	0.3761	Std.Dev.	75.0035	Std.Dev.	2353.8051	Std.Dev.	1.0848	
	NA	47	NA	6	NA	41	NA	41	NA	6	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
	Min.	41.0000		Min.	10795.0000					Min.	0.5000
	1st Quartile	706.8000		1st Quartile	16786.0000					1st Quartile	1.2000
	Median	1207.0000	0: 60	Median	19724.0000	1: 55				Median	2.5500
	Mean	1420.2000	1: 84	Mean	19638.0000	2: 54				Mean	4.4270
	3rd Quartile	1943.8000	NA: 6	3rd Quartile	22592.0000	NA: 41				3rd Quartile	5.7750
	Max.	4523.0000		Max.	28050.0000					Max.	24.5000
	Std.Dev.	1043.6933		Std.Dev.	3953.2176					Std.Dev.	4.8574
	NA	6		NA	6					NA	6
	clsd	albumin	copper	alk	sgot	trig	platelet	prothrombin	stage		
stage==NA	Min.	151.0000	Min.	1.9600	Min.	13.0000	Min.	310.0000	Min.	0.5000	
	1st Quartile	243.5000	1st Quartile	3.0670	1st Quartile	58.0000	1st Quartile	96.0000	1st Quartile	0.8750	
	Median	299.0000	Median	3.3400	Median	58.0000	Median	142.0000	Median	1.4500	
	Mean	337.8000	Mean	3.3020	Mean	127.1000	Mean	2017.0000	Mean	2.7500	
	3rd Quartile	375.0000	3rd Quartile	3.5720	3rd Quartile	150.0000	3rd Quartile	2139.0000	3rd Quartile	4.2000	
	Max.	1092.0000	Max.	4.5200	Max.	588.0000	Max.	11047.0000	Max.	5.7000	
	Std.Dev.	171.6532	Std.Dev.	0.4372	Std.Dev.	105.8719	Std.Dev.	1934.2260	Std.Dev.	6.6689	
	NA	55	NA	6	NA	42	NA	41	NA	6	
	days	status	treatment	age	sex	ascites	hepatom	spiders	edema	bili	
	Min.	41.0000		Min.	16802.0000					Min.	0.7000
stage==NA	1st Quartile	852.2000		1st Quartile	20454.0000					1st Quartile	0.8750
	Median	2148.5000	0: 2	Median	21185.0000	1: 6				Median	1.4500
	Mean	1982.7000	1: 4	Mean	20941.0000	NA: 6				Mean	2.7500
	3rd Quartile	2844.0000	NA: 6	3rd Quartile	22463.0000					3rd Quartile	4.2000
	Max.	4062.0000		Max.	23376.0000					Max.	7.1000
	Std.Dev.	1522.0838		Std.Dev.	2340.5378					Std.Dev.	2.6689
	NA	6		NA	6					NA	6
	clsd	albumin	copper	alk	sgot	trig	platelet	prothrombin	stage		
	Min.	2.4800	Min.	1.03750	Min.	1.5750	Min.	1.9600	Min.	0.5000	
	1st Quartile	3.0750	1st Quartile	3.3170	1st Quartile	3.6470	1st Quartile	3.6900	1st Quartile	0.8750	
	Median	3.5750	Median	3.6470	Median	3.6900	Median	3.6900	Median	1.4500	
	Mean	3.3170	Mean	3.6470	Mean	3.6900	Mean	3.6900	Mean	2.7500	
	3rd Quartile	3.6470	3rd Quartile	3.6470	3rd Quartile	3.6900	3rd Quartile	3.6900	3rd Quartile	4.2000	
	Max.	3.6900	Max.	3.6900	Max.	3.6900	Max.	3.6900	Max.	7.1000	
	Std.Dev.	0.4984	Std.Dev.	NA	Std.Dev.	NA	Std.Dev.	NA	Std.Dev.	2.6689	
	NA	6		NA	6					NA	6

Table 104: PBC 테이터 요약(Cont'd)

25.2.23 데이터셋 활용 시 참고사항

25.2.23.1 H5 확장자를 R에서 불러오기

확장자가 H5인 파일은 Hierarchical Data Format(HDF)으로 저장된 데이터 파일이다. 여기에는 과학 데이터의 다차원 배열이 포함된다. H5 파일은 항공 우주, 물리학, 공학, 금융, 학술 연구, 유전체학, 철학, 전자 기기 및 의료 분야에서 일반적으로 사용된다.

간혹 파이썬의 tensorflow, theano, keras 라이브러리를 사용한 예제 코드의 경우, 이 형태의 데이터 파일을 활용한 경우가 있다. 이를 R에서 불러들이려면 rhdf5라는 패키지를 설치해야하며, 이는 CRAN이 아닌 Bioconductor를 통해 설치할 수 있다. whas_train_test.h5라는 파일을 data라는 객체에 담기 위해서는 다음과 같이 입력하면 된다.

```
1 # source("http://bioconductor.org/biocLite.R")
2 # biocLite("rhdf5")
3
4 library(rhdf5)
5
6 data = h5read("whas_train_test.h5", "test")
```

Bibliography

1. 김재균 and 서대출(2009), 생존 분석의 원리와 방법, Neurointervention, Vol. 4, 6-7.
2. 송경일 and 최종수(2007), SPSS 15를 이용한 생존자료의 분석, 한나래출판사.
3. 송인식(2017), 퀄리티 바이블(Quality Bible): 신뢰성 분석 편, 이담.
4. 석진미(2017), 두 범주 결과변수에 대한 예측모형 구축시 연속형 독립변수의 비선형성을 반영하는 방법 비교, 고려대학교 대학원 의학통계학 협동과정, 석사학위논문.
5. 임성빈(????), Survival Analysis via Deep Learning, 내부 자료.
6. 추순규(2015), 생존 자료에서 새로운 인자의 예측력을 평가하는 방법의 비교, 연세대학교 대학원 의학전산통계학 협동과정, 석사학위논문.
7. Antolini L, Boracchi P and Biganzoli E.(2005), A time-dependent discrimination index for survival data, Stat Med, Vol. 24, No. 24, 3927-44.
8. Anny Xiang, Pablo Lapuerta, Alex Ryutov, Jonathan Buckley and Stanley Azen(2000), **Comparison of the performance of neural network methods and Cox regression for censored survival data**, Computational Statistics & Data Analysis, Vol. 34, 243–257.
9. Bain L. J.(1978), Statistical Analysis of Reliability and Life Testing Models, Dekker.
10. Balbir S. Dhillon(1979), A hazard rate model, IEEE Transactions on Reliability, Vol. 28, 180.
11. Balbir S. Dhillon(1981), Life Distributions, IEEE Transactions on Reliability, Vol. 30, No. 5. 457-459.
12. Bain L. J. and Engelhardt M.(1991), **Statistical Analysis of Reliability and Life-Testing Models: Theory**, Marcel Dekker.
13. B. A. Olhausen and D. J. Field(1996), **Emergence of simple-cell receptive field properties by learning a sparse code for natural images**, Nature, Vol. 381, 607-609.
14. Bender, R., T. Augustin and M. Blettner(2005), **Generating survival times to simulate Cox proportional hazards models**, Stat Med, Vol. 24, No. 11, 1713-1723.
15. Birnbaum Z. W. and Saunders S. C.(1968), **A probabilistic interpretation of miner's rule**, SIAM-Jour. Appl. Math., Vol. 16, 637–652.
16. Birnbaum Z. W. and Saunders S. C.(1969), **A new formula of life distribution**, SIAM-Jour. Appl. Prob., Vol. 6, 319–317.
17. K. O. Bowman and L. R. Shenton(1983), **Maximum Likelihood Estimators for the Gamma Distribution Revisited**, Communications in Statistics - Simulation and Computation, Vol. 12.
18. Buckley J. and James I.(1979), **Linear regression with censored data**, Biometrika, Vol. 66, 429-436.
19. Bullinger L., Dohner K., Bair E., Frohling S., Schlenk R. F., Tibshirani R., Dohner H. and Pollack J. R., **Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia**. New England Journal of Medicine, Vol. 350, 1605-1616.
20. Byar D. P. and Green D. K.(1980), **The choice of treatment for cancer patients based on covariates information: application to prostate cancer**, Bulletin of Cancer, Paris, Vol. 67, 477-488.
21. Chambless L. E. and Diao G.(2006), **Estimation of time-dependent areas under the ROC curve for long-term risk prediction**, Stat Med, Vol. 23, No. 4, 361-387.
22. David Faraggi and Richard Simon(1995), **A Neural Network Model for Survival Data**, Statistics in Medicine, Vol. 14, 73-82.
23. Drzewiecki K. T. and Andersen P. K.(1982), **Survival with malignant melanoma. A regression analysis of prognostic factors**, Cancer, Vol. 49, 2414-2419.
24. Efron B.(1988), **Logistic regression, survival analysis, and the Kaplan—Meier curve**, Journal of the American Statistical Association, Vol. 83, 414—425.
25. Fleming T. and Harrington D.(1991), **Counting Processes and Survival Analysis**, Wiley.
26. Elisa T. Lee and John Wenyu Wang(2013), **Statistical Methods for Survival Data Analysis - 4th Edition**, Wiley.

27. Greenwood J. A. and Durand D.(1960),
Aids for fitting the gamma distribution by maximum likelihood, *Technometrics*, Vol. 2, 55-65.
28. M. Gonen and G. Heller(2005),
Concordance probability and discriminatory power in proportional hazards regression, *Biometrika*, Vol. 92.
29. Giuliana Cortese, Thomas H. Scheike and Torben Martinussen(2010),
Flexible survival regression modelling, *Statistical Methods in Medical Research*, Vol. 19, 5–28.
30. Hanley, J. A. and B. J. McNeil(1982),
The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, Vol. 143, No. 1, 29-36.
31. Hanley J. A., B. J. McNeil(1983),
A method of comparing the areas under receiver operating characteristic curves derived from the same cases, *Radiology*, Vol. 148, No. 3, 839-843.
32. Heagerty, P. J. and Y. Zheng(2005),
Survival model predictive accuracy and ROC curves, *Biometrics*, Vol. 61, No. 1, 92-105.
33. Hemant Ishwaran, Udaya B. Kogalua, Eugene H. Blackstone and Michael S. Lauer(2008),
Random Survival Forests, *The Annals of Applied Statistics*, Vol. 2, No. 3, 841-860.
34. H. Lee, C. Ekanadham and A. Y. Ng(2008),
Sparse deep belief net model for visual area V2, *Advances in Neural Information Processing Systems*, Vol. 20.
35. H. Lee, R. Grosse, R. Ranganath and A. Y. Ng(2009),
Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, *Proceedings of the 26th Annual International Conference on Machine Learning*, 609-616.
36. Hothorn T., Buhlmann P., Dudoit S., Molinaro A. and van der Laan M. J.(2006).
Survival ensembles, *Biostatistics*, Vol. 7, No. 3, 355-373.
37. Ishwaran H., Blackstone E. H., Pothier C. and Lauer M. S.(2004),
Relative risk forests for exercise heart rate recovery as a predictor of mortality, *J. Amer. Statist. Assoc.*, Vol. 99, 591–600.
38. Ishwaran H. and Kogalur U. B.(2007), **Random survival forests for R**, *Rnews*, Vol. 7, No. 2, 25-31.
39. Jared L. Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang and Yuval Kluger(2016),
Deep Survival: A Deep Cox Proportional Hazards Networks, *arXiv:1606.00931v2 [stat.ML]* 25 Oct 2016.
40. Kattan M.(2003),
Comparison of Cox regression with other methods for determining prediction models and nomograms, *J. Urol.*, Vol. 170, 6–10.
41. Kalbfleisch J. D. and Prentice R. L.(1980), **The Statistical Analysis of Failure Time Data**, *Wiley*.
42. Knut Liestol, Per Kragh Andersen and Ulrich Andersen(1994),
Survival analysis and neural nets, *Statistics in medicine*, Vol. 13, No. 12, 1189–1200.
43. Koziol J. A., Jia Z.(2009), **The concordance index C and the Mann-Whitney parameter $Pr(X > Y)$ with randomly censored data.**, *Biometrical Journal*, Vol. 51, No. 3, 467-474.
44. M. Ito and H. Komatsu(2004),
Representation of angles embedded within contour stimuli in area V2 of macaque monkeys, *The Journal of Neuroscience*, Vol. 24, No. 13, 3313-3324.
45. Military Specification (MIL)-HDBK-338A, **Electronic Reliability Design Handbook** (DOD, 12 October 1988).
46. Margaux Luck, Tristan Sylvain, Meloise Cardinal, Andrea Lodi and Yoshua Bengio(2017),
Deep Learning for Patient-Specific Kidney Graft Survival Analysis, *arXiv:1705.10245v1 [cs.LG]* 29 May 2017.
47. Odd O. Aalen, Per Kragh Andersen, Ørnulf Borgan, Richard D. Gill, Niels Keiding (2010),
History of applications of martingales in survival analysis, *Electronic Journal for History of Probability and Statistics*, Vol. 5, No. 1. 28.
48. Pencina, M. J., R. B. D'Agostino, Sr., R. B. D'Agostino, Jr. and R. S. Vasan(2008).
Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond, *Stat Med*, Vol. 27, No. 2, 157–72.
49. Pencina, M. J., R. B. D'Agostino, Sr. and E. W. Steyerberg(2011).
Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers, *Stat Med*, Vol. 30, No. 1, 11–21.
50. Peter M. Ravdin and Gary M. Clark(1992).
A practical application of neural network analysis for predicting outcome of individual breast cancer patients, *Breast Cancer Research and Treatment*, Vol. 22, No. 3, 285–293.

51. Ping Wang, Yan Li and Chandan K. Reddy(2017),
Machine Learning for Survival Analysis: A Survey, ACM Computing Surveys, Vol. 1, No. 1.
52. Prentice R. L. and Gloeckler L. A.(1978),
Regression analysis of grouped survival data with application to breast cancer data, Biometrics, Vol. 34, 57-67.
53. Q. V. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean and A. Y. Ng(2012),
Building high-level features using large scale unsupervised learning, Proceedings of the 29th International Conference on International Conference on Machine Learning, 507-514.
54. Roger Newson(2007), **Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences**, The Stata Journal, Vol. 2, No. 1, 45-64.
55. Rossi P. H., Berk R. A. and Lenihan K. J.(1980),
Money, Work and Crime: Some Experimental Results, Academic Press.
56. Rupert Miller and Jerry Halpern(1982), **Regression with Censored Data**, Biometrika, Vol. 69, No. 3, 521-31.
57. Saleh J. H. and Marais Ken(2006), **Highlights from the Early (and pre-) History of Reliability Engineering**, Reliability Engineering and System Safety, Vol. 91, No. 2, 249-256.
58. Salvia A. A.(1985), **Reliability application of the alpha distribution**, IEEE Transactions on Reliability, Vol. 34, 251-252.
59. Stacy E. W.(1962), **A generalization of the gamma distribution**,
Ann. Math. Statist., Vol. 33, 1187-1192.
60. Steyerberg E. W.(2009), **Clinical Prediction Models A Practical Approach to Development, Validation, and Updating**, Springer.
61. Terry Therneau, Cynthia Crowson and Elizabeth Atkinson(2014), **Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model**
62. Weibull W. A.(1939), **A Statistical Theory of the Strength of Materials**, Ingeniors Ventenskaps Akademien Ilandlinger., Vol. 151, 5-45.
63. Van Der Laan M. J. and Dudoit, S.(2003),
Unified cross-validation methodology for selection among estimators: finite sample results, asymptotic optimality, and applications,
Technical Report 130, Division of Biostatistics, University of California, Berkeley, California,
<http://www.bepress.com/ucbbiostat/paper130>.
64. Van Der Laan M. J. and Robins J. M.(2003),
Unified Methods for Censored Longitudinal Data and Causality, Springer.
65. Venables W. N. and Ripley B. D.(2004), **Modern Applied Statistics with S-Plus, 4th edition**, Springer-Verlag.