




# Weibull 분포 개념 정리

▼ 아래 링크에서도 확인 가능합니다.

 [weibull 분포 개념 정리](#)

## Weibull 분포란?

- 수명자료 분석을 위한 모수적 추정방법의 한 종류
  - 지수분포, 와이블분포 등이 있음
- 고장률이 (일정/증가/감소) 중 어느 것인지 모를 때,  $f(t)$ 는 Weibull 분포를 따름

## cf) 지수분포(Exponential Distribution)의 개념

- 고장률함수가 상수로 시간변화에 관계없이 고장률이 일정한 분포
- 평균수명 : 기대시간
  - 여기서 평균 수명은 시스템을 수리하여 사용하는 경우와 사용할 수 없는 경우로 나뉨
    - 수리 가능한 경우의 평균 수명 : MTBF
    - 수리 불가능한 경우의 평균 수명 : MTTF

## 왜 Weibull 분포를 사용하는가?

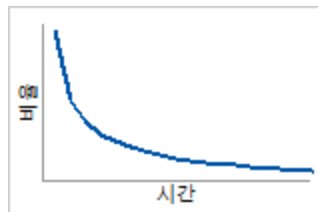
- 지수 분포의 경우 시간변화에 관계없이 고장률이 항상 일정함
  - 따라서, weibull 분포를 사용해 시간에 따른 고장률 변화를 확인하고자 함.
- Weibull 분포의 경우 보다 범용적으로 사용이 가능하고 reasonable한 기법으로 알려져 있음

## Weibull 분포의 모수

- The 2-parameter Weibull distribution has a scale and shape parameter. The 3-parameter Weibull includes a location parameter.
- $\beta$ (beta)  
is the **shape parameter**, also known as the **Weibull slope**
  - Weibull 분포의 모양은 형상모수  $\beta$ 에 의해 전적으로 결정됨
- $\eta$ (eta or alpha)  
is the **scale parameter**
- $\gamma$   
is the location parameter

## Weibull 분포의 $\beta$ 에 따른 확률밀도함수

- $0 < \beta < 1$  : 초기 고장
  - 초기의 높은 고장률이 시간이 지남에 따라 감소함



- 제품 수명의 초기를 모델링 : "burn-in"기간
- 초기 고장은 제품 수명의 초기 단계에서 발생합니다. 이러한 고장으로 인해 초기 고장의 위험을 줄이기 위해 제품 "초기 고장" 기간이 반드시 필요할 수도 있습니다.
- $\beta = 1$  : 우발 고장
  - 고장률이 일정하게 유지됨



- 고장이 랜덤하게 발생하는 경우 제품의 '실제 사용 수명'을 모델링
- 고장의 원인이 여러가지일 수 있음.

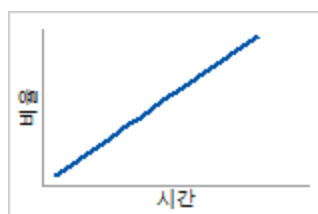
- **$\beta = 1.5$  : 조기 마모 고장**

- 고장률이 초기에 가장 크게 증가



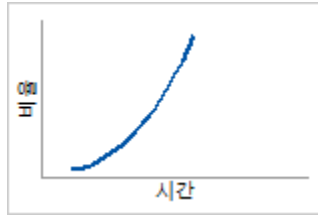
- **$\beta = 2$**

- 제품 수명기간동안 마모고장 위험이 꾸준히 증가
- 고장률이 선형으로 증가



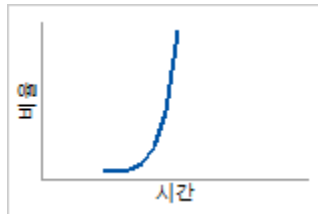
- **$3 \leq \beta \leq 4$  : 빠른 마모 고장**

- 빠른 마모 고장. 고장이 가장 많이 발생하는 제품 수명 기간의 최종 기간을 모델링



- $\beta > 10$  : 매우 빠른 마모 고장

- 매우 빠른 마모 고장. 고장이 가장 많이 발생하는 제품 수명 기간의 최종 기간을 모델링.



- 참고자료 ) <https://support.minitab.com/ko-kr/minitab/20/help-and-how-to/statistical-modeling/reliability/supporting-topics/distribution-models/weibull-distribution/>

## Weibull 분포 계산을 위한 수식

- $\beta$ (beta) /  $\eta$ (eta or alpha)
  - 프로그램을 사용해 결과 도출
  - Reliability Library의 alpha, beta 도출을 위한 코드(2-parameter)

▼ code

```
## https://github.com/MatthewReid854/reliability

x, y = plotting_positions(failures=list_fail, right_censored=list_censored)

x = np.array(x)
y = np.array(y)

def linear_regression(...) # 생략 -> github 코드 참조
```

```

xlin = np.log(x)
ylin = np.log(-np.log(1 - y))
slope, intercept = linear_regression(
    xlin, ylin
)
LS_beta = slope
LS_alpha = np.exp(-intercept / LS_beta)
guess = [LS_alpha, LS_beta]

print(guess)

```

- MTTF

- $MTTF = \eta \cdot \Gamma(1/\beta + 1)$

- Reliability / Probability of Failure

- Probability of Failure

- F(t) is the cumulative probability of failure from time zero till time t.

$$F(t) = 1 - \eta^{-(t/n)^\beta}$$

- Reliability

- R(t) is the chance of survival from from time zero till time t.

$$R(t) = \eta^{-(t/n)^\beta}$$

- Bearing Life

- The “BX” or “Bearing Life” nomenclature, which refers to the time at which X% of items in a population will fail.

$$B(1 - P) \text{ Life} = \eta[-\ln(1 - P)]^{1/\beta}$$

## 수명분포 파라미터 추정 방법

- 최소제곱추정법(Least Square Estimator: LSE)
  - 데이터의 형태가 완전자료인 경우 이론상 Bias 없으나, 누적고장확률 계산 선택에 따라 Bias가 있을 수 있음.
  - 분산 추정 바로 할 수 없기에 분산 추정 크게 구해짐
  - 분포의 모형 확정시 MLE보다 더 정밀함
  - 컴퓨터 통계 패키지 프로그램 사용하지 않아도 계산 가능
- 최대우도추정법(Maximum Likelihood Estimator: MLE)
  - 시료의 크기가 증가할수록 Bias 감소
  - 분산 추정 작게 구해짐
  - 파라미터 추정의 경우 LSE 보다 더 좋은 추정량 가짐
  - 컴퓨터 통계 패키지 프로그램을 사용하지 않으면 계산이 힘들

출처 : 와이블 분포와 정시중단 하에서의 MLE와 LSE의 정확도 비교, 품질경영학회지

- etc..

## 신뢰도 분석을 위한 분포 적합 평가

- Anderson-Darling 적합도 통계량 및 Pearson 상관 계수를 통한 평가
  - Anderson-Darling 값이 낮을수록 일반적으로 분포가 더 적합하다는 것을 나타냅니다. 최대우도 추정 방법(MLE)과 최소 제곱 추정 방법(LSE)에서 모두 Anderson-Darling 통계량이 계산됩니다.
  - Pearson 상관 계수 값이 클수록 분포가 더 적합하다는 것을 나타냅니다. 상관 계수는 최소 제곱 추정 방법(LSE) 방법에 사용할 수 있습니다.
- Log-likelihood
  - The **log-likelihood value** of a regression model is a way to measure the goodness of fit for a model.

- The higher the value of the log-likelihood, the better a model fits a dataset.

```
#calculate log-likelihood value of each model
logLik(model1)

'log Lik.' -91.04219 (df=3)

logLik(model2)

'log Lik.' -111.7511 (df=3)
```

→ The first model has a higher log-likelihood value (**-91.04**) than the second model (**-111.75**), which means the first model offers a better fit to the data.

- AICc

- The Akaike information criterion (AIC)

$$AIC = -2 \text{ LogLikelihood} + 2p$$

- AIC is calculated from the maximum likelihood estimate of the model (how well the model reproduces the data).
- **lower is better.**
- 불필요한 파라미터, 독립변수의 수가 증가할수록 2k를 증가시켜 패널티를 부여하여 모델의 품질을 평가
- 2k가 증가할수록 AIC 값이 증가하게 되므로 좋지 않은 모형

- BIC

- **Bayesian information criterion (BIC)**

$$BIC = -2 \text{ LogLikelihood} + \log(n)p$$

- **lower is better.**
- AIC와 유사하지만 마지막 패널티를 수정함으로써 AIC를 보완

- BIC의 경우 변수가 많을 수록 AIC보다 더 페널티를 가함
- 결국 AIC, BIC를 최소화 한다는 뜻은 **우도(likelihood)**를 가장 크게 하는 동시에 변수 갯수는 가장 적은 **최적의 모델(parsimonious & explainable)**을 의미
  - 참고자료 : <https://rk1993.tistory.com/entry/AIC-BIC-Mallows-Cp-쉽게-이해하기>

## Estimation of Parameters using R

```
library("fitdistrplus")

data <- read.csv("C:\\\\~~~")

parameters <- fitdist(data$uptime, 'weibull', 'mle')$estimate

parameters
```

## Weibull 분포 모수 추정을 위한 Python 라이브러리 설명

👉 [링크 접속](#)