Goodness-of-fit Test – Model (Distribution) Selection

Chanseok Park (박찬석)

Applied Statistics Laboratory
Department of Industrial Engineering
Pusan National University

July 13, 2018

Hosted by SEC

Overview

1. 적합도(適合度/goodness of fit)
   - Model 선택
   - 그래프를 이용한 Model 선택 방법
2. Weibull Plots
   - Weibull random variable
   - How to draw Weibull probability plot
3. Weibull Plot Examples
   - Example 1: W-LN
   - Example 2: W-N
   - Example 3: W-W
4. Test of Weibullness
   - Real Data Example
5. Test of Weibullness using R package
   - 실습: R package (weibullness) with Real Data

Overview

1. 적합도(適合度/goodness of fit)
   - Model 선택
   - 그래프를 이용한 Model 선택 방법
2. Weibull Plots
   - Weibull random variable
   - How to draw Weibull probability plot
3. Weibull Plot Examples
   - Example 1: W-LN
   - Example 2: W-N
   - Example 3: W-W
4. Test of Weibullness
   - Real Data Example
5. Test of Weibullness using R package
   - 실습: R package (weibullness) with Real Data

Overview

1. 적합도(適合度/goodness of fit)
   - Model 선택
   - 그래프를 이용한 Model 선택 방법
2. Weibull Plots
   - Weibull random variable
   - How to draw Weibull probability plot
3. Weibull Plot Examples
   - Example 1: W-LN
   - Example 2: W-N
   - Example 3: W-W
4. Test of Weibullness
   - Real Data Example
5. Test of Weibullness using R package
   - 실습: R package (weibullness) with Real Data

Overview

1. 적합도(適合度/goodness of fit)
   - Model 선택
   - 그래프를 이용한 Model 선택 방법
2. Weibull Plots
   - Weibull random variable
   - How to draw Weibull probability plot
3. Weibull Plot Examples
   - Example 1: W-LN
   - Example 2: W-N
   - Example 3: W-W
4. Test of Weibullness
   - Real Data Example
5. Test of Weibullness using R package
   - 실습: R package (weibullness) with Real Data

Overview

1. 적합도(適合度/goodness of fit)
   - Model 선택
   - 그래프를 이용한 Model 선택 방법
2. Weibull Plots
   - Weibull random variable
   - How to draw Weibull probability plot
3. Weibull Plot Examples
   - Example 1: W-LN
   - Example 2: W-N
   - Example 3: W-W
4. Test of Weibullness
   - Real Data Example
5. Test of Weibullness using R package
   - 실습: R package (weibullness) with Real Data

Popular methods

- **Pearson Goodness-of-Fit Test:**
  Use Observations and Expected values.
  For continuous case, we have to determine bins (intervals).
  For discrete case, it is advantageous.
- **Kolmogorov-Smirnov Goodness-of-Fit Test (K-S test):**
  Compare the CDFs. It needs true (population) parameter values.
- Anderson-Darling: Similar to K-S test.

Common mistakes in Model Selection

Common mistakes in using K-S or A-D Test

- It only applies to continuous distributions. (연속형만 가능).
- It tends to be more sensitive near the center of the distribution than at the tails. Kurtosis (尖度/峯度)가 normal model departure에서 중요한 역할을 함. 즉, tail이 각종 통계량에 미치는 영향이 훨씬 중요.
- Perhaps the most serious limitation is that the distribution must be fully specified. (모든 모수는 추정치가 아닌 **모집단의 값을 알고 있어야 함**). That is, if location, scale, and shape parameters are estimated from the data, the critical region of the K-S test is no longer valid. It typically must be determined by simulation. (If it is pivotal, then it would be OK).
- See https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm

그래프를 이용한 Model 선택 방법

- **Basic idea - linearization(선형화)**
- Q-Q plot (quantile versus quantile plot. **The most widely used**).
- Q-Q norm (Q-Q plot under normal distribution).
- Q-Q Weibull (Q-Q plot under Weibull. It is **impossible**).
- Other plots – Weibull plot (not QQ Weibull), etc.

장단점

- **장점:** Seeing is believing.
- **단점:** Seeing is illusive (착시) and subjective (주관적).

Can we make the graphical method objective?

- Calculate the **sample correlation** in the plot.
- Then how to find the critical value for the sample correlation?

Q-Q plot 원리

Let $y_1, y_2, \ldots, y_n$ be order statistics of $x_1, x_2, \ldots, x_n$. Then the empirical CDF of y_i is approximated by

$$F(y_i) \approx \frac{i}{n} \text{ or } \frac{i}{n+1} \text{ or } \frac{i-0.5}{n} \text{ or } \frac{i-0.375}{n+0.25}$$

etc. Take an inverse, F^{-1} on both sides.

$$y_i \approx F^{-1}\left(\frac{i-0.5}{n}\right)$$

Plot y_i (on y-axis) versus $F^{-1}\left(\frac{i-0.5}{n}\right)$ (on x-axis):

$$y_1, y_2, \ldots, y_n \text{ vs. } F^{-1}\left(\frac{1-0.5}{n}\right), \ldots, F^{-1}\left(\frac{2-0.5}{n}\right), F^{-1}\left(\frac{n-0.5}{n}\right).$$

Q-Q norm 원리

For the normal distribution, we have a very nice property

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

where $X \sim N(\mu, \sigma)$. Note that $(X - \mu)/\sigma$ is a **pivotal quantity**. Thus, the CDF of $(X - \mu)/\sigma$ is given by $\Phi(z)$ which does **not include any unknown parameters**. We have

$$\frac{y_i - \mu}{\sigma} \approx \boldsymbol{\Phi^{-1}}\left(\frac{i - 0.5}{n}\right),$$

which is equivalent to

$$y_i = \mu + \sigma\Phi^{-1}\left(\frac{i - 0.5}{n}\right) = a + bt_i,$$

where $a = \mu$ (intercept), $b = \sigma$ (slope), and $t_i = \Phi^{-1}\left(\frac{i-0.5}{n}\right)$. Regardless of intercept and slope, it is a normal if the plot is on the straight line.

Weibull random variable

- One of the most popular distributions used to model the lifetimes and reliability data is the Weibull distribution, named after a Swedish **mechanical engineer** by the name of Walodie Weibull 1939.
- Indeed, this distribution is as central to the parametric analysis of **reliability engineering** and **survival** data as the normal distribution in statistics.

- For Weibull case, it is impossible to derive the pivotal quantity with Quantile – Quantile.
- Thus, we have to use a different way.
- But, its spirit is the same – $\boxed{\textbf{Linearization}}$.

Weibull random variable

Definition

A random variable X is called Weibull with shape α and scale β if its cumulative distribution function is given by

$$F(x) = 1 - \exp\left\{ -\left(\frac{x}{\beta}\right)^{\alpha} \right\}, \qquad x \geq 0.$$

It is easily shown that its pdf is given by

$$f(x) = \frac{\alpha x^{\alpha-1}}{\beta^{\alpha}} \exp\left\{ -\left(\frac{x}{\beta}\right)^{\alpha} \right\}.$$

The mean of the Weibull random variable is $\beta \cdot \Gamma\left[(\alpha+1)/\alpha\right]$, where $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$.

Weibull random variable

Reparametrized Form

The following reparametrized form is often used

$$F(x) = 1 - \exp\big(-\lambda x^{\alpha}\big) \quad \text{and} \quad f(x) = \lambda\alpha x^{\alpha-1}\exp\big(-\lambda x^{\alpha}\big),$$

where $\lambda = \beta^{-\alpha}$.

- The parameter λ is called the rate parameter.
- It should be noted that with the shape parameter $\alpha = 1$, the Weibull distribution becomes the exponential distribution with mean β (or rate λ).

How to draw Weibull probability plot

The Weibull CDF is given by

$$F(x) = 1 - \exp\big(-\lambda x^{\alpha}\big).$$

Then we have

$$\log(1-p) = -\lambda x_p^{\alpha},$$

where $p = F(x)$. It follows that

$$\log\big\{-\log(1-p)\big\} = \log\lambda + \alpha\log x_p.$$

This implies that the plot of

$$\boxed{\log\big\{-\log(1-p)\big\}} \text{ versus } \boxed{\log x_p}$$

draws a straight line with the slope α and the intercept $\log\lambda$. The widely-used Weibull *probability paper* is based on this idea.

How to draw Weibull probability plot

- Need to find $p = F(x)$ and x_p with real experimental data. That is, we need to estimate $p = F(x)$ and x_p in the following plot:
- The empirical CDF $\hat{F}(x)$ is used for $p = F(x)$ which is an increasing step function jumping $1/n$ at $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where $x_{(i)}$ is sorted from the smallest.
- Thus $\hat{p}_i = \hat{F}(x_{(i)})$ has values, $1/n, 2/n, \ldots, n/n$.

$$\boxed{\log\left\{-\log(1-\hat{p}_i)\right\}} \text{ versus } \boxed{\log x_{(i)}}.$$

Actually, Blom (1958) method is more popular, which uses $\hat{p}_i = (i - 0.375)/(n + 0.25)$ to have better power – plotting position problem.

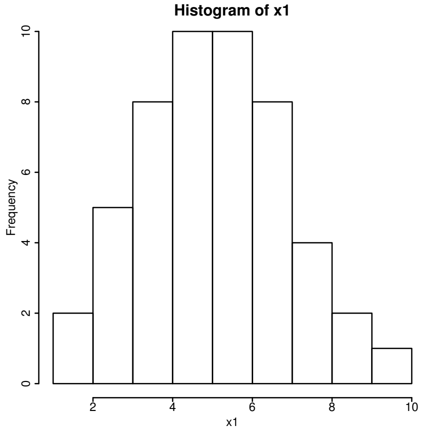
Example 1: W-LN

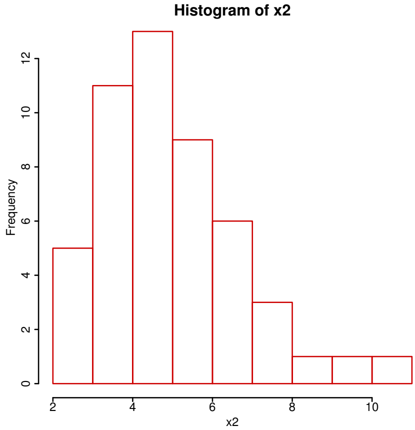
Data I

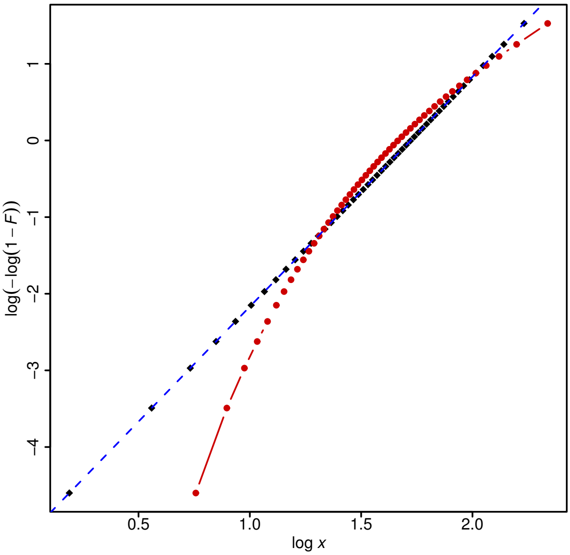
1.208339 1.748644 2.080409 2.335513 2.548681 2.734928 2.902281 3.055600 3.198078 3.331945 3.458826 3.579955 3.696291 3.808603 3.917520 4.023568 4.127191 4.228776 4.328661 4.427151 4.524519 4.621021 4.716894 4.812366 4.907658 5.002987 5.098571 5.194633 5.291403 5.389125 5.488057 5.588484 5.690718 5.795109 5.902057 6.012024 6.125552 6.243292 6.366035 6.494765 6.630734 6.775576 6.931493 7.101568 7.290332 7.504885 7.757387 8.071660 8.506590 9.315591

Data II

2.132357 2.451394 2.652611 2.809841 2.943303 3.061764 3.169903 3.270548 3.365555 3.456222 3.543500 3.628113 3.710634 3.791524 3.871172 3.949906 4.028015 4.105758 4.183369 4.261071 4.339071 4.417576 4.496786 4.576906 4.658146 4.740724 4.824872 4.910837 4.998892 5.089334 5.182497 5.278755 5.378541 5.482349 5.590762 5.704470 5.824303 5.951271 6.086630 6.231970 6.389342 6.561469 6.752075 6.966453 7.212504 7.502789 7.859157 8.324999 9.008337 10.356137


Histogram of x1
Frequency
x1
0
2
4
6
8
10
2
4
6
8
10


Histogram of x2
Frequency
x2
0
2
4
6
8
10
12
2
4
6
8
10


log(−log(1 − F))
log x
0.5
1.0
1.5
2.0
1
0
−1
−2
−3
−4

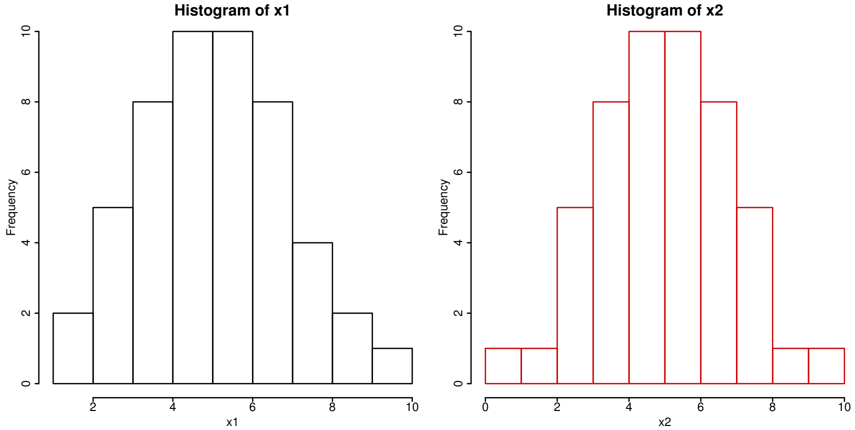
Example 2: W-N

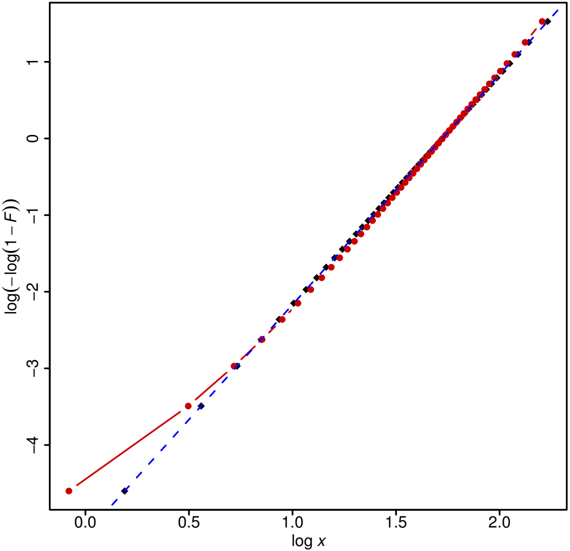
Data I

1.208339 1.748644 2.080409 2.335513 2.548681 2.734928 2.902281 3.055600 3.198078 3.331945 3.458826 3.579955 3.696291 3.808603 3.917520 4.023568 4.127191 4.228776 4.328661 4.427151 4.524519 4.621021 4.716894 4.812366 4.907658 5.002987 5.098571 5.194633 5.291403 5.389125 5.488057 5.588484 5.690718 5.795109 5.902057 6.012024 6.125552 6.243292 6.366035 6.494765 6.630734 6.775576 6.931493 7.101568 7.290332 7.504885 7.757387 8.071660 8.506590 9.315591

Data II

0.923350 1.642687 2.049681 2.346765 2.586173 2.789748 2.968822 3.130079 3.277814 3.414961 3.543624 3.665370 3.781399 3.892659 3.999913 4.103790 4.204817 4.303443 4.400057 4.495004 4.588591 4.681098 4.772786 4.863899 4.954671 5.045329 5.136101 5.227214 5.318902 5.411409 5.504996 5.599943 5.696557 5.795183 5.896210 6.000087 6.107341 6.218601 6.334630 6.456376 6.585039 6.722186 6.869921 7.031178 7.210252 7.413827 7.653235 7.950319 8.357313 9.076650


Histogram of x1
Frequency
x1
Histogram of x2
Frequency
x2


log(−log(1 − F))
1
0
−1
−2
−3
−4
0.0
0.5
1.0
1.5
2.0
log x

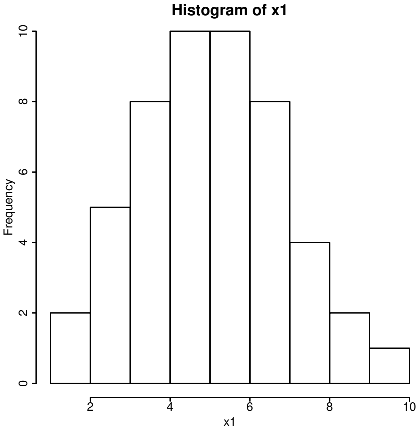
Example 3: W-W

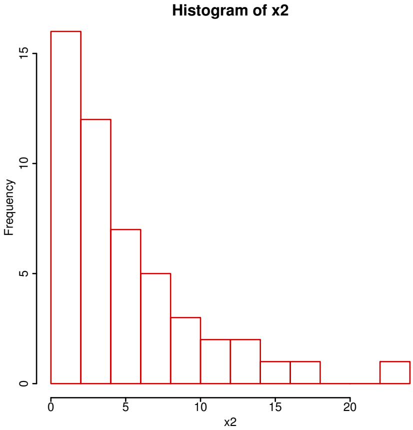
Data I

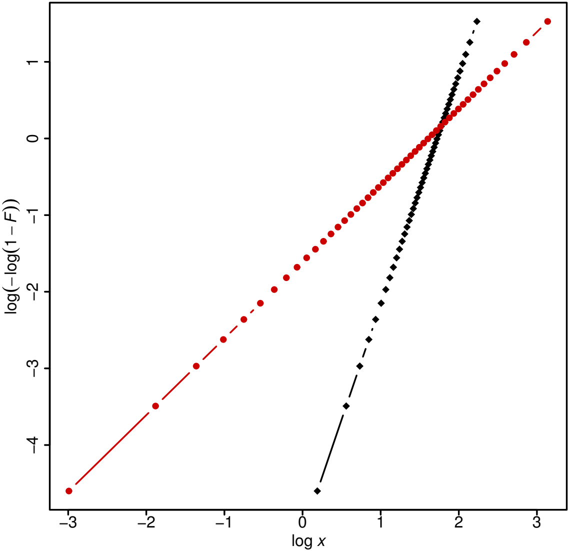
1.208339 1.748644 2.080409 2.335513 2.548681 2.734928 2.902281 3.055600 3.198078 3.331945 3.458826 3.579955 3.696291 3.808603 3.917520 4.023568 4.127191 4.228776 4.328661 4.427151 4.524519 4.621021 4.716894 4.812366 4.907658 5.002987 5.098571 5.194633 5.291403 5.389125 5.488057 5.588484 5.690718 5.795109 5.902057 6.012024 6.125552 6.243292 6.366035 6.494765 6.630734 6.775576 6.931493 7.101568 7.290332 7.504885 7.757387 8.071660 8.506590 9.315591

Data II

0.0502516 0.1522960 0.2564664 0.3628534 0.4715534 0.5826690 0.6963103 0.8125946 0.9316478 1.0536051 1.1786116 1.3068238 1.4384103 1.5735537 1.7124515 1.8553184 2.0023878 2.1539145 2.3101773 2.4714816 2.6381637 2.8105945 2.9891850 3.1743913 3.3667227 3.5667494 3.7751129 3.9925384 4.2198503 4.4579906 4.7080427 4.9712613 5.2491106 5.5433131 5.8559149 6.1893717 6.5466666 6.9314718 7.3483798 7.8032387 8.3036560 8.8597842 9.4855999 10.2011041 11.0363745 12.0397280 13.2963001 14.9786613 17.5327894 23.0258509


Histogram of x1
Frequency
x1
0
2
4
6
8
10
2
4
6
8
10


Histogram of x2
Frequency
x2
0
5
10
15
0
5
10
15
20


log(−log(1 − F))
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
log x

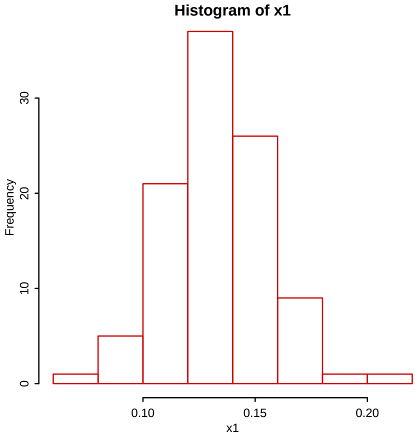
Real Data: Birnbaum-Saunders and Leemis Data

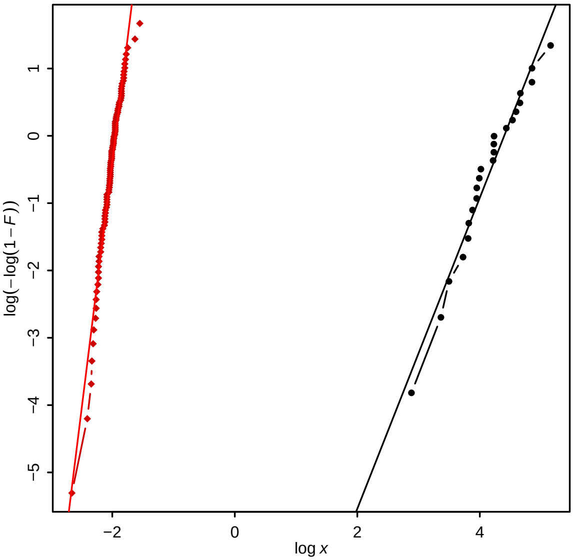
Data I - from Birnbaum and Saunders (1969)

0.07, 0.09, 0.096, 0.097, 0.099, 0.1, 0.103, 0.104, 0.104, 0.105, 0.107, 0.108, 0.108, 0.108, 0.109, 0.109, 0.112, 0.112, 0.113, 0.114, 0.114, 0.114, 0.116, 0.119, 0.12, 0.12, 0.12, 0.121, 0.121, 0.123, 0.124, 0.124, 0.124, 0.124, 0.124, 0.128, 0.128, 0.129, 0.129, 0.13, 0.13, 0.13, 0.131, 0.131, 0.131, 0.131, 0.131, 0.132, 0.132, 0.132, 0.133, 0.134, 0.134, 0.134, 0.134, 0.134, 0.136, 0.136, 0.137, 0.138, 0.138, 0.138, 0.139, 0.139, 0.141, 0.141, 0.142, 0.142, 0.142, 0.142, 0.142, 0.142, 0.144, 0.144, 0.145, 0.146, 0.148, 0.148, 0.149, 0.151, 0.151, 0.152, 0.155, 0.156, 0.157, 0.157, 0.157, 0.157, 0.158, 0.159, 0.162, 0.163, 0.163, 0.164, 0.166, 0.166, 0.168, 0.170, 0.174, 0.196, 0.212

Data II - Example 8.16 in Leemis (1995)

17.88, 28.92, 33.00, 41.52, 45.12, 45.60, 48.48, 51.84, 51.96, 54.12, 55.56, 67.80, 68.64, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40


Histogram of x1
Frequency
0
10
20
30
0.10
0.15
0.20
x1
Histogram of x2
Frequency
0
2
4
6
8
0
50
100
150
x2



log(−log(1 − F))
1
0
−1
−2
−3
−4
−5
−2
0
2
4
log x

How to determine whether the data are from Weibull or not

$$H_0 : \text{Weibull} \quad \text{versus} \quad H_1 : \text{non-Weibull.}$$

- Recall: $\log\Big\{-\log(1-p)\Big\} = \log\lambda + \alpha\log x_p$.
- Thus $\boxed{\log\Big\{-\log(1-\hat{p}_i)\Big\}}$ versus $\boxed{\log x_{(i)}}$ draws a straight line, where $\hat{p}_i = (i-0.375)/(n+0.25)$. Thus, the linearity measure (sample correlation) can be used for Weibullness test.
- Idea: Large r implies Weibullness.

$$\boxed{H_0 : r \geq r_0 \quad \text{versus} \quad H_1 : r < r_0}$$

Then, how large is large enough for r?

- To this end, we should know the distribution of r and then find the critical value (from pivot) with the significance level (or Type-I error).

Distribution of r sample correlations under the normal distribution

If (X, Y) is from bi-variate normal, it is known that

$$r\sqrt{\frac{n-2}{1-r^2}} \sim t\text{-distribution}$$

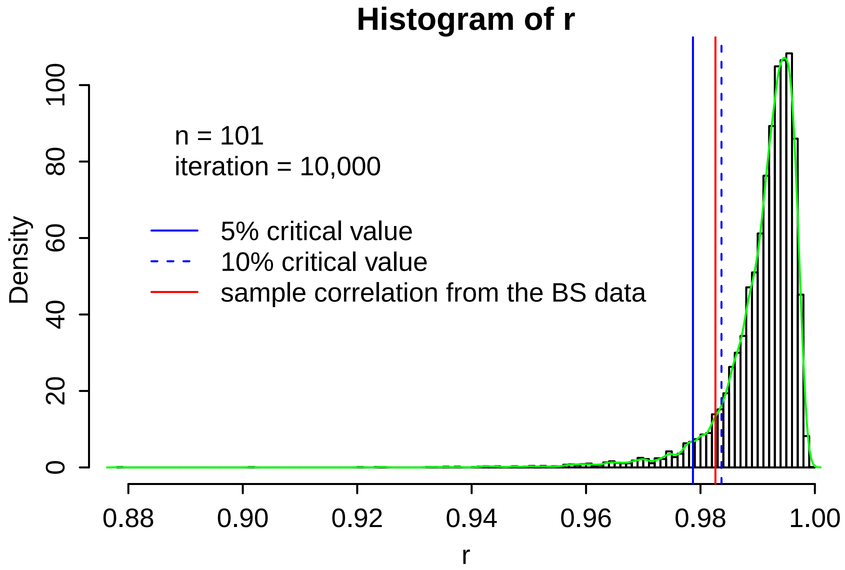
with $n-2$ degrees of freedom.

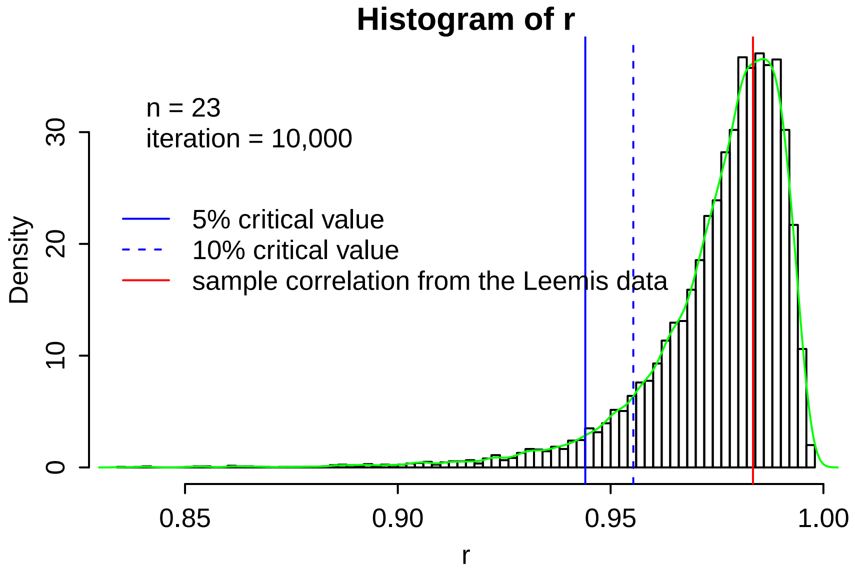
- In general, $-1 \leq r \leq 1$, and $-\infty < r\sqrt{\frac{n-2}{1-r^2}} < \infty$.
- As $n \to \infty$, CLT works in general. I.e., $r\sqrt{\frac{n-2}{1-r^2}} \xrightarrow{d} N(0,1)$
- In the Weibull plot, $0 < r \leq 1$. Thus, $0 < r\sqrt{\frac{n-2}{1-r^2}} < \infty$.
- CLT can not work for the Weibull case.

- We can not use normal approximation. We need to find the distribution of r.
- We can find the distribution of r under Weibull distribution using Monte Carlo simulation.
- Again, recall $\log\big\{-\log(1-p)\big\} = \log\lambda + \alpha\log x_p$.
- Since $\mathrm{cor}(aX+b, cY+d) = \mathrm{cor}(X, Y)$, it is enough to generate random sample of size n from any Weibull distribution. This implies that the correlation from the Weibull plot is independent of the parameters α and λ.
- Thus, the sample correlation is a **pivotal quantity**.

Algorithm

- Generate Weibull random observations of size n from any Weibull, say, Weibull (1,1).
- Sort the data. Denote $x_{(i)}$.
- Calculate $\log\{-\log(1-\hat{p}_i)\}$ where $\hat{p}_i = (i-0.375)/(n+0.25)$.
- Calculate the sample correlation between $\log\{-\log(1-\hat{p}_i)\}$ and $\log x_{(i)}$.
- Repeat the above (say, up to N iteration numbers).
- Find the empirical quantiles for critical values (say, 5%, 10%, etc.)


Histogram of r
n = 101
iteration = 10,000
5% critical value
10% critical value
sample correlation from the BS data
Density
0
20
40
60
80
100
0.88
0.90
0.92
0.94
0.96
0.98
1.00
r


Histogram of r
n = 23
iteration = 10,000
5% critical value
10% critical value
sample correlation from the Leemis data
Density
0
10
20
30
0.85
0.90
0.95
1.00
r

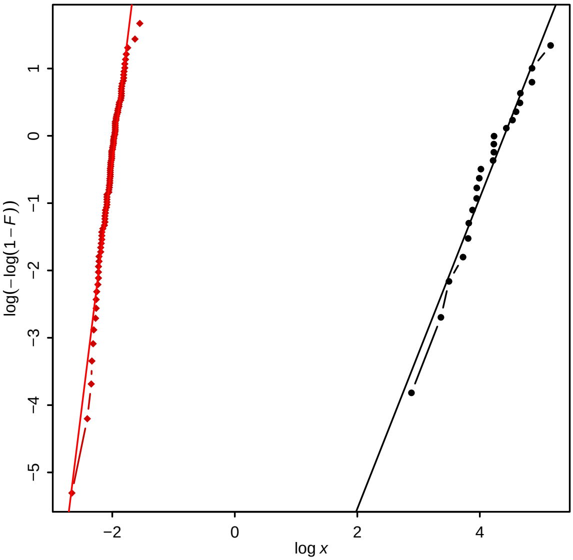
Critical Values

n	1.0%	2.0%	2.5%	5.0%	10%	20%
101	0.9593	0.9686	0.9710	0.9777	0.9833	0.9878
23	0.9085	0.9239	0.9284	0.9429	0.9553	0.9665

Sample Correlations for the BS and Leemis Data Sets

- Data I (BS Data)
  $r = 0.982614$ with $n = 101$
  Using 5% Type-I error, the Data are from Weibull.
  However, with 10% or more Type-I error, the Data are not from Weibull.
- Data II (Leemis Data)
  $r = 0.983456$ with $n = 23$
  The Data are from Weibull for any above Type-I errors.

We can also find the p-value from the empirical pdf. The p-value for the BS data is 8.5% while that for the Leemis is 63%.


log(−log(1 − F))
log x
−2
0
2
4
−5
−4
−3
−2
−1
0
1

Recall Real Data: Birnbaum-Saunders and Leemis Data

Data I - from Birnbaum and Saunders (1969)

0.07, 0.09, 0.096, 0.097, 0.099, 0.1, 0.103, 0.104, 0.104, 0.105, 0.107, 0.108, 0.108, 0.108, 0.109, 0.109, 0.112, 0.112, 0.113, 0.114, 0.114, 0.114, 0.116, 0.119, 0.12, 0.12, 0.12, 0.121, 0.121, 0.123, 0.124, 0.124, 0.124, 0.124, 0.124, 0.128, 0.128, 0.129, 0.129, 0.13, 0.13, 0.13, 0.131, 0.131, 0.131, 0.131, 0.131, 0.132, 0.132, 0.132, 0.133, 0.134, 0.134, 0.134, 0.134, 0.134, 0.136, 0.136, 0.137, 0.138, 0.138, 0.138, 0.139, 0.139, 0.141, 0.141, 0.142, 0.142, 0.142, 0.142, 0.142, 0.142, 0.144, 0.144, 0.145, 0.146, 0.148, 0.148, 0.149, 0.151, 0.151, 0.152, 0.155, 0.156, 0.157, 0.157, 0.157, 0.157, 0.158, 0.159, 0.162, 0.163, 0.163, 0.164, 0.166, 0.166, 0.168, 0.170, 0.174, 0.196, 0.212

Data II - Example 8.16 in Leemis (1995)

17.88, 28.92, 33.00, 41.52, 45.12, 45.60, 48.48, 51.84, 51.96, 54.12, 55.56, 67.80, 68.64, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40

First, download weibullness R package at:
https://cran.r-project.org/web/packages/weibullness/

R package – weibullness

```
# Data coding:
# Birnbaum-Saunders data
data1 = c(
0.070, 0.090, 0.096, 0.097, 0.099, 0.100, 0.103, 0.104, 0.104, 0.105,
0.107, 0.108, 0.108, 0.108, 0.109, 0.109, 0.112, 0.112, 0.113, 0.114,
0.114, 0.114, 0.116, 0.119, 0.120, 0.120, 0.120, 0.121, 0.121, 0.123,
0.124, 0.124, 0.124, 0.124, 0.124, 0.128, 0.128, 0.129, 0.129, 0.130,
0.13,  0.130, 0.131, 0.131, 0.131, 0.131, 0.131, 0.132, 0.132, 0.132,
0.133, 0.134, 0.134, 0.134, 0.134, 0.134, 0.136, 0.136, 0.137, 0.138,
0.138, 0.138, 0.139, 0.139, 0.141, 0.141, 0.142, 0.142, 0.142, 0.142,
0.142, 0.142, 0.144, 0.144, 0.145, 0.146, 0.148, 0.148, 0.149, 0.151,
0.151, 0.152, 0.155, 0.156, 0.157, 0.157, 0.157, 0.157, 0.158, 0.159,
0.162, 0.163, 0.163, 0.164, 0.166, 0.166, 0.168, 0.170, 0.174, 0.196,
0.212)
# Leemis data
data2 = c(
 17.88, 28.92, 33.00, 41.52, 45.12, 45.60, 48.48, 51.84,  51.96,  54.12,
 55.56, 67.80, 68.64, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84,
127.92,128.04,173.40)
```

R package – weibullness (Continued)

```
> install.packages("weibullness")

> library(weibullness)
 weibullness Package is installed.

> wp.test(data1)
        Weibullness test from the Weibull plot

data:  data1
correlation = 0.98261, p-value = 0.08924

> wp.test(data2)
        Weibullness test from the Weibull plot

data:  data2
correlation = 0.98346, p-value = 0.6294
```

References I

Birnbaum, Z. W. and S. C. Saunders (1969). Estimation for a family of life distributions with applications to fatigue. *Journal of Applied Probability 6*, 328–347.

Blom, G. (1958). *Statistical Estimates and Transformed Beta Variates*. New York: Wiley.

Leemis, L. M. (1995). *Reliability*. Englewood Cliffs, N.J.: Prentice-Hall.

Park, C. (2018). `weibullness`: Goodness-of-fit test for Weibull (Weibullness test). `https://CRAN.R-project.org/package=weibullness`. R package version 1.18.6.

Weibull, W. (1939). A statistical theory of the strength of material. *Proceedings, Royal Swedish Institute for Engineering Research 151*, 1–45.