Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

457.212 Statistics for Civil & Environmental Engineers

In-Class Material: Class 24

Testing Validity of Distributions: (2) Chi-square Test, K-S Test, A-D Test (A&T: 7.3)

Given: Sample data set $\{x_1, x_2, ..., x_n\}$

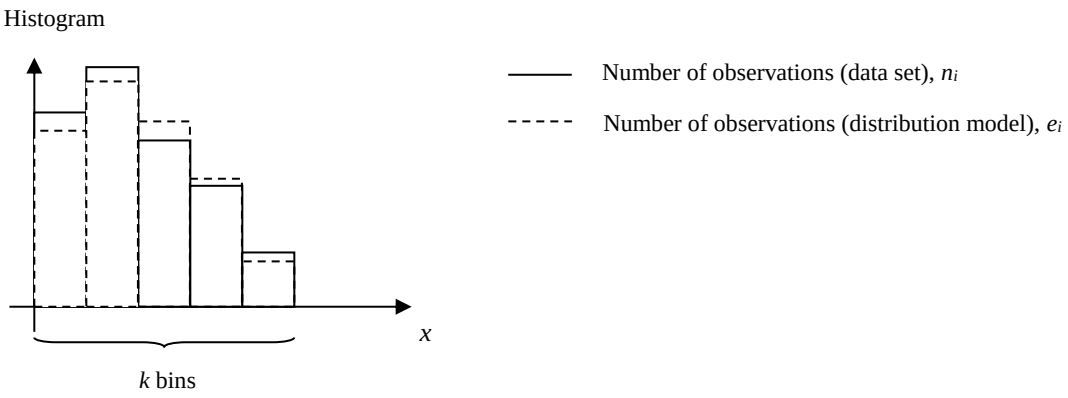
Question: Does it follow a certain type of distribution or not? (e.g. Normal, Lognormal…)

"Goodness-of-Fit" Test

(1) Visual: by probability paper (plot)

(2) Numerical: Chi-square test, K-S test, A-D test

1. **Chi-square (χ^2) Test:** Use histogram



(a) It is known that, if the dataset follows the same distribution, the error measure

$$\sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$$

follows "Chi-square (χ^2)" distribution with degree of freedom $f = k - 1$ as $N \to \infty$

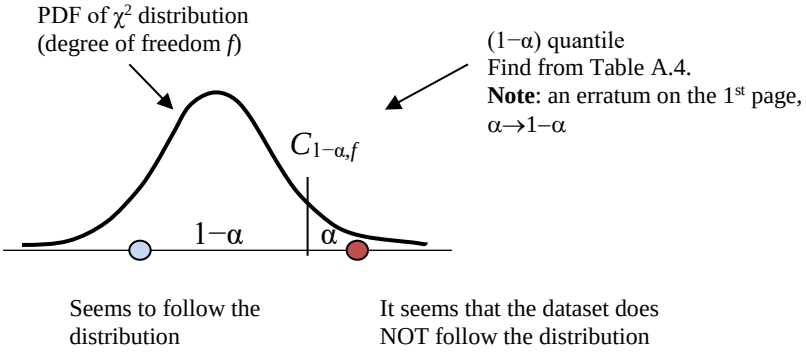
Note: If the parameters of the assumed distributions are estimated from the dataset, (i.e. instead of being given), the degree of freedom is adjusted to

$$f = k - 1 - r$$

in which r is the number of estimated parameters.

Instructor: Junho Song
junhosong@snu.ac.kr

(b) If the error measure for the given dataset is too large, it is more likely that the dataset does not follow the distribution model.



(c) If $\sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} > C_{1-\alpha,f}$

the hypothesis: "The given dataset follows the assumed distribution model" is rejected with the "significance level" α (i.e. probability of "wrong rejection")

Otherwise, i.e. $\sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \le C_{1-\alpha,f}$, the hypothesis is not rejected with $(1-\alpha)\%$ confidence.

(d) Significance level α: the level of strictness.

Example 1 (A&T 7.6): Severe rainstorms per year for a 66-year period: {0, 2, 1, 0, … }

Histogram:

No. of rainstorms	No. of years (n_i)
0	20
1	23
2	15
3	6
4	2

Following Poisson distribution? (PMF: $p_X(x) = \frac{(\nu t)^x}{x!}\exp(-\nu t),\ t = 1$)

Chi-square test with $\alpha = 5\%$ significance level

Instructor: Junho Song
junhosong@snu.ac.kr

(a) $\hat{\nu}$

(b) Combine the 3rd and 4th bins: (Why? want to avoid small e_i's)

(c) The number of bins: $k =$

(d) The d.o.f. of the Chi-square distribution: $f =$

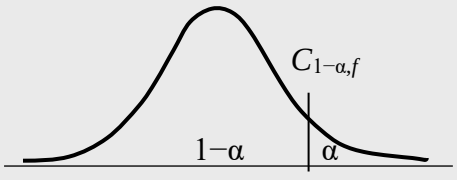
(e) (1−α) quantile of the Chi-square distribution: $C_{1-\alpha, f}$

(f) The error measure $\sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$?

No. of rainstorms	No. of years (n_i)	By Poisson (e_i)	$(n_i - e_i)^2 / e_i$
0	20	19.94	0.0002
1	23		
2	15		
3-4	8		
Sum	66		

(g) Comparison between the error measure and the quantile of Chi-square distribution:

(h) Conclusion?

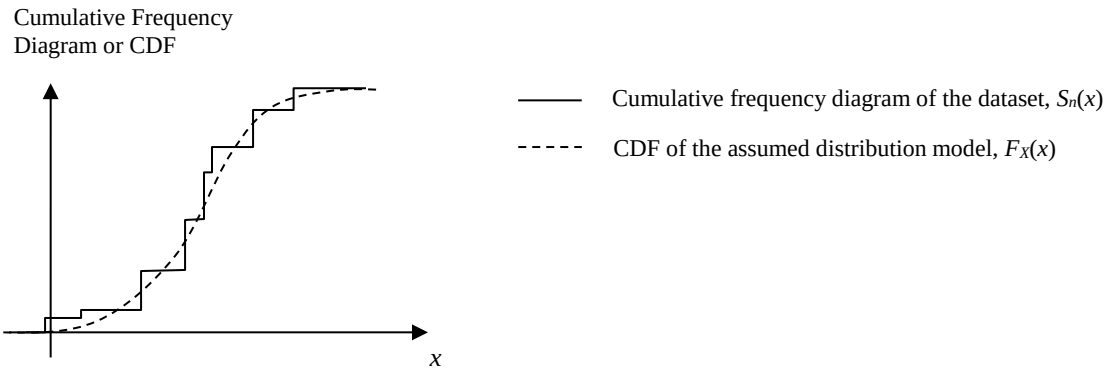


Instructor: Junho Song
junhosong@snu.ac.kr

(e) Issues of Chi-square tests:

- The error measure is sensitive to small e_i's
- Subjective test: the number of bins?
- Small number of intervals → inaccurate tests (trade-off between the first and third)
- Inaccurate test results in terms of tail behavior (bins usually merged)

2. Kolmogorov-Smirnov (K-S) Test: Use CDF (less subjective than Chi-square test)

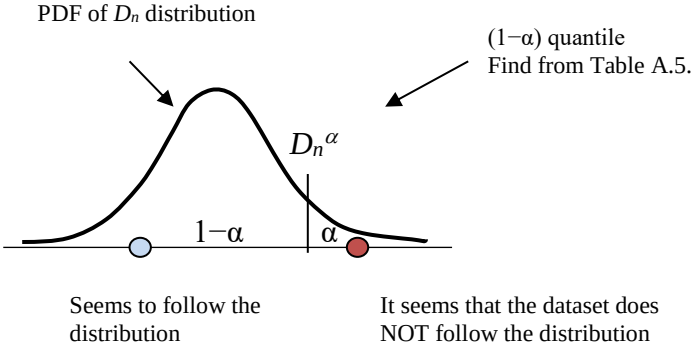


(a) The distribution of the maximum difference between the CDF of the assumed distribution and the cumulative frequency diagram, i.e.

$$D_n = \max_x \left| F_X(x) - S_n(x) \right|$$

has been derived. The only parameter n is the number of data points.

(b) If this error measure is too large, i.e. $D_n > D_n^\alpha$ in which D_n^α is the (1−α) quantile (See Table A.5), the hypothesis of following the assumed distribution is rejected with the significance level α (with probability of "wrong rejection" α)



Instructor: Junho Song
junhosong@snu.ac.kr

Example 2 (A&T 7.9): The dataset in Example 7.1

Follow a Normal distribution? (i.e. $F_X(x) = \Phi\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)$)

Perform a K-S test with $\alpha = 5\%$ significance level

(a) Estimates on μ and σ (by MLE)

$\hat{\mu} =$

$\hat{\sigma} =$

(b) $n =$

(c) $D_n^{\alpha} = D_{(\ )}^{(\ )} =$ (Table A.5)

(d) $D_n =$ (See Table E7.9)

(e) Conclusion?

```
# Example 2 (A&T 7.9)
e = 10^-5 # small value added to avoid duplicate entries
FT =
c(69.5,71.9,72.6,73.1,73.3,73.5,74.1,74.2,75.3,75.5,75.7,75.8,76.1,76.2,76
.2+e,76.9,77.0,77.9,78.1,79.6,79.7,79.9,80.1,82.2,83.7,93.7)
# Fracture Toughness from Example 7.1
ks.test(FT,"pnorm",mean(FT),sd(FT), alternative=c("two.sided"))
ks.test(FT,"pnorm",mean(FT),sd(FT)*sqrt(25/26),
alternative=c("two.sided"))

# Kolmogorov-Smirnov Test by R
KS1 = rnorm(100,0,1)
KS2 = rnorm(100,1,1)
KS3 = rexp(100,1)

ks.test(KS1, KS2, alternative=c("two.sided"))
# two-sided with same dist
ks.test(KS1, KS2, alternative=c("greater"))
ks.test(KS1, KS2, alternative=c("less"))
# one-sided with same dist

ks.test(KS1,KS3, alternative=c("two.sided"))
# two-sided with different dist
ks.test(KS1, "pnorm",0,1)
ks.test(KS3, "pnorm",0,1)
# one-sample test(with specific dist)
```

Instructor: Junho Song
junhosong@snu.ac.kr

Example 3 (A&T 7.6): K-S test

x	$S_n(x)$	$p_X(x)$	$F_X(x)$	$\lvert S_n(x)-F_X(x)\rvert$
0	20/66	0.302	0.302	0.00103
1	43/66	0.362	0.664	0.0125
2	58/66	0.216	0.880	0.00121
3	64/66	0.0864	0.966	0.00370
4	1	0.0258	0.992	0.008

$D_n =$

$D_n^{\alpha} = D_{66}^{0.5} = \text{—} =$

Conclusion?

3. Anderson-Darling (A-D) Test: Use CDF but more weight at _____

Neither Chi-square nor K-S tests can reveal discrepancy between the empirical and theoretical frequencies at the _____ of the proposed distribution.

To this end, Anderson and Darling (1954) proposed the following alternative A-D statistic (valid for $n > 7$), which uses logarithmic functions to place more weights on the tails:

$$A^2 = -\frac{1}{n}\sum_{i=1}^{n}(2i-1)\{\ln F_X(x_i) + \ln[1 - F_X(x_{n+1-i})]\} - n$$

where $x_i, i = 1, \ldots, n$ denotes the *i*-th value in the data set sorted in the increasing order, and $F_X(\cdot)$ denotes the cumulative distribution function of the proposed distribution.

Adjust the statistics based on the sample size n, i.e. $A^2 \rightarrow A^*$, and compare with the critical value c_α to reject the null hypothesis (if $A^* > c_\alpha$) or not (if $A^* < c_\alpha$).

Note: Null hypothesis is “the sample follows the proposed distribution”

Instructor: Junho Song
junhosong@snu.ac.kr

Proposed Distribution	Critical value c_α	Adjusted Statistic A^*
Normal	$c_\alpha = a_\alpha\left(1+\frac{b_0}{n}+\frac{b_1}{n^2}\right)$ See Table A.6a for a_α, b_0 and b_1	$A^* = A^2\left(1.0+\frac{0.75}{n}+\frac{2.25}{n^2}\right)$
Exponential	See Table A.6b for c_α	$A^* = A^2\left(1.0+\frac{0.6}{\sqrt{n}}\right)$
Gamma	See Table A.6c for c_α	For $k=1$, $A^* = A^2\left(1.0+\frac{0.6}{n}\right)$ For $k \geq 2$, $A^* = A^2 + (0.2+0.3/k)/n$
Extremal distributions	See Table A.6d for c_α	$A^* = A^2\left(1.0+\frac{0.2}{\sqrt{n}}\right)$

Example 4 (A&T 7.11): The dataset in Example 7.1

Follow a Normal distribution? (i.e. $F_X(x) = \Phi\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)$)

Perform an A-D test with 5% significance level

$A^2 =$

$A^* =$

$c_{0.05} =$

Therefore,

```
# Anderson-Darling Test
install.packages("goftest")
library(goftest)
ad.test(FT, null="pnorm", mean(FT),sd(FT))# Caution: no adjustment
```