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&T: 7.3)

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

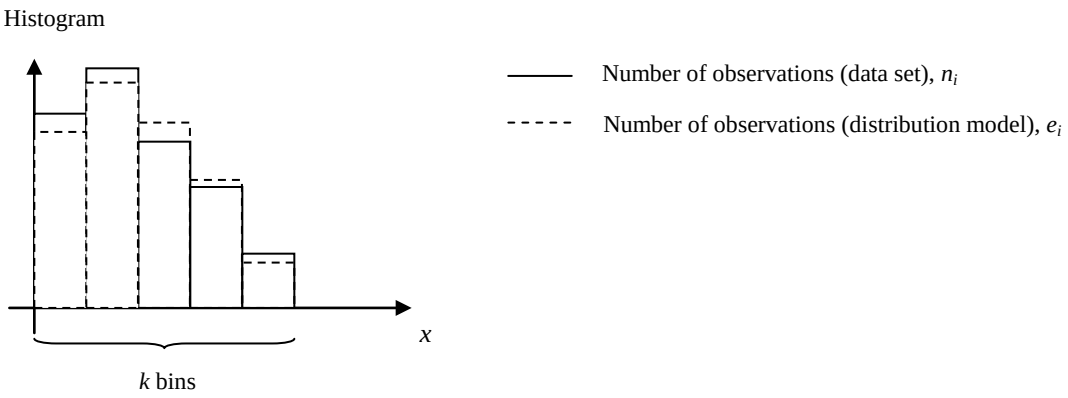
Question: Does it follow a certain type of distribution or not? (e.g. Normal, Lognormal…)

"Goodness-of-Fit" Test

(1) Visual: by probability paper (plot)

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

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 $N \to \infty$

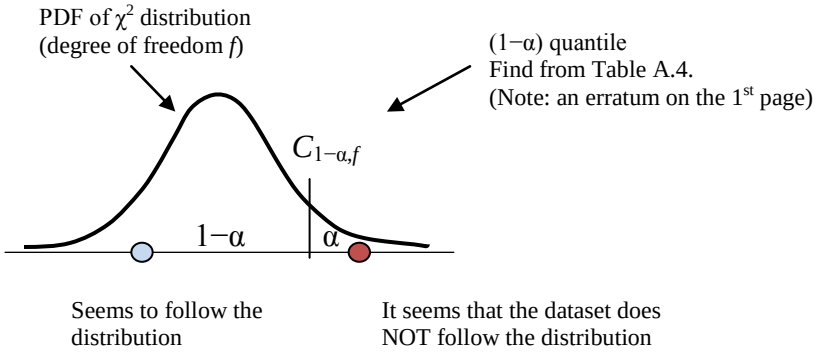
Note: If the parameters of the assumed distributions are estimated from the dataset, the degree of freedom is

$$f = k - 1 - r$$

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: $f =$

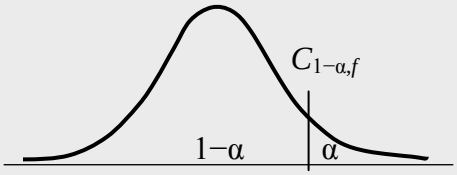
(e) (1−α) quantile of the Chi-square distribution: $C_{1-\alpha,f}$

(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?

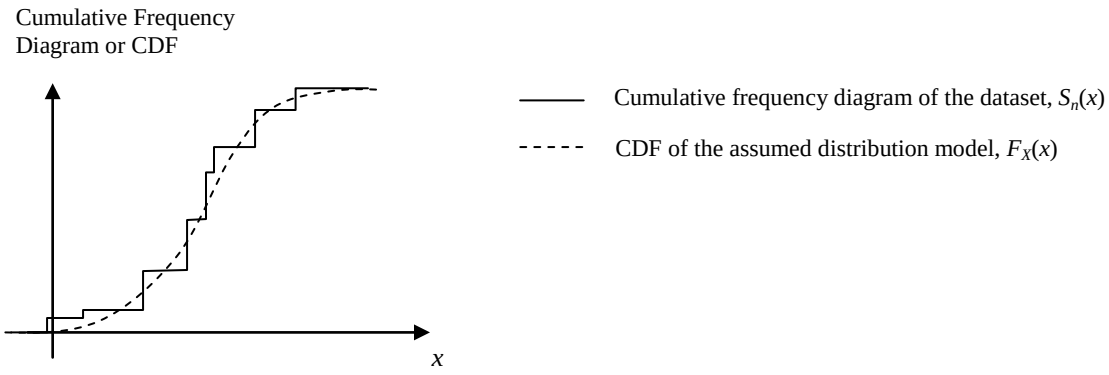


Instructor: Junho Song
junhosong@snu.ac.kr

(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)

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

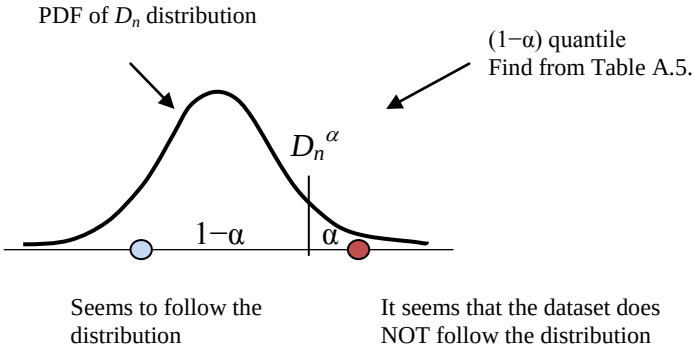


(a) The distribution of the maximum difference between the CDF of the assumed distribution and the cumulative frequency diagram, i.e.

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

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 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?