S_03 Normality Test and Sample Size

Statistical Analysis

References

- Text and figures from the DATAtab site (https://datatab.net/) and the book "Statistics made easy" published by DATAtab.

- Statistics Page from Scribbr site (https://www.scribbr.com/category/statistics/)

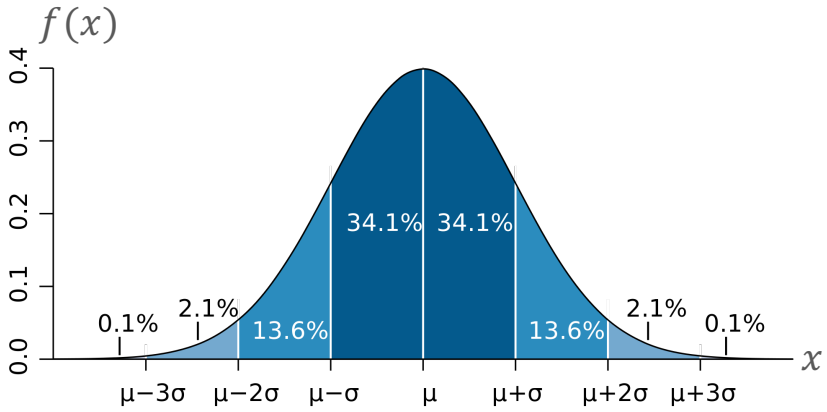
Normal Distribution (정규분포)

- Probability density function (pdf)
  - Bell 모양의 대칭형 곡선

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ: mean
σ: standard deviation

- Central Limit Theorem (CLT, 중심극한정리)
  - Sample이 커지면 Sample 내 data의 합과 평균은 normal distribution에 근접해 간다.



Standard Normal Distribution (표준정규분포)

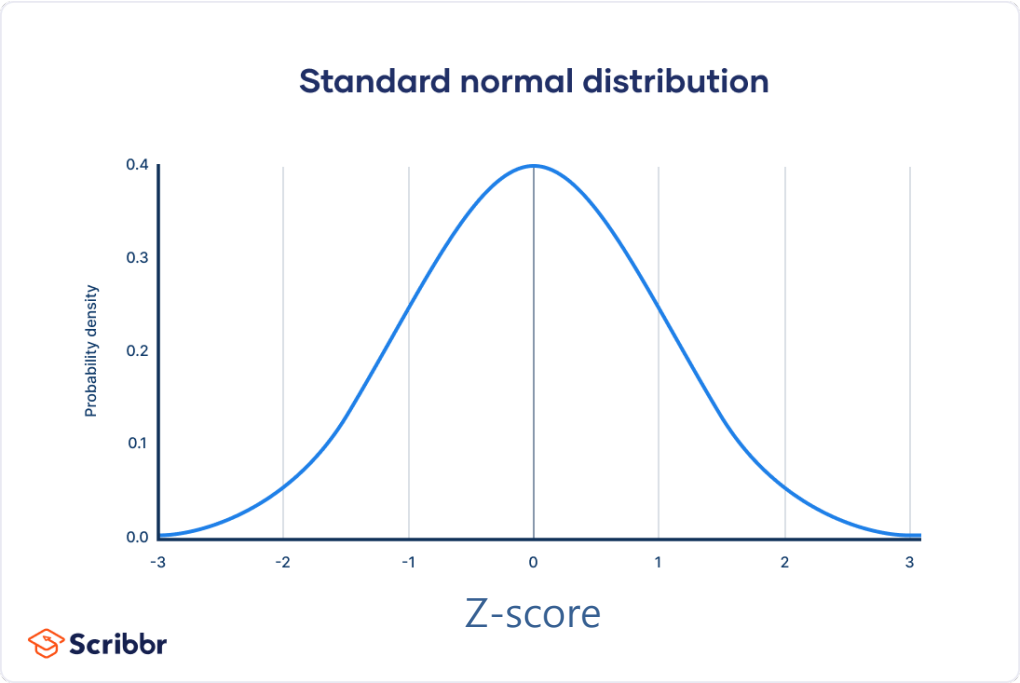
- Special case of Normal Distribution
  mean: $\mu = 0$, standard deviation: $\sigma = 1$
- pdf

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

- Normal to Standard Normal:
  - Z-score conversion:

$$Z = \frac{x - \mu}{\sigma}$$

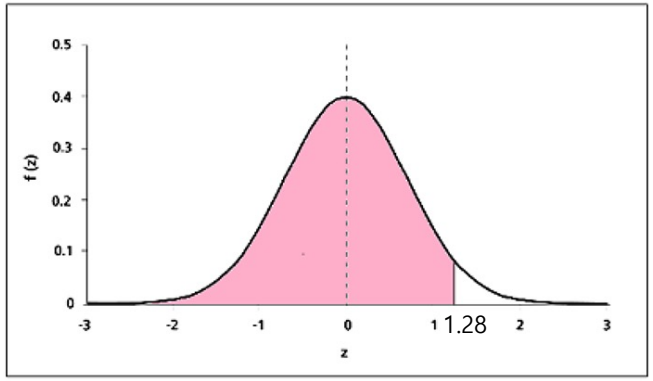
  - x와 μ의 차이를 σ 단위로 나타낸 것
  - 서로 다른 dataset 들에서의 stat 값을 비교할 때 사용 가능
- Standard Normal Distribution을 "Z-Distribution" 이라고도 부름



Standard Normal Table (Unit Normal Table, Z-Table) (1/2)

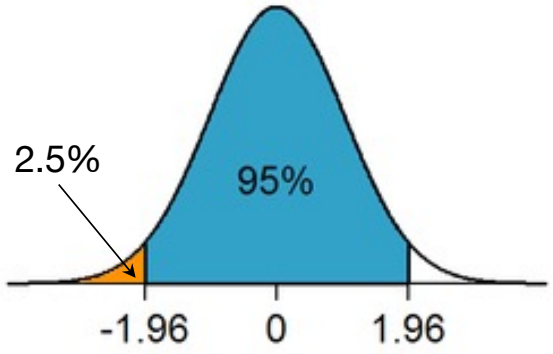
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

- Z가 특정값일 확률
  - Pr(Z < 0.00) = 0.50
  - Pr(Z < 0.53) = 0.7019
  - Pr(Z < 1.28) = 0.8997



Standard Normal Table (Unit Normal Table, Z-Table) (2/2)

- 확률이 0.975 가 되는 Z 값은?
  - Table에서 Z ≈ 1.96
  - 그림을 참고하면 이것은 Significance Level 0.05 일 때의 Z 값을 말함



- 확률이 0.995 가 되는 Z 값은?
  - 앞 슬라이드의 Table에서는 확인이 불가능하지만 Z ≈ 2.576 (Significance Level 0.01 일 때)

t-Distribution (1/2)

- pdf:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

  ν: 자유도 (df), 일반적으로 Sample size - 1

  Γ: Gamma 함수, factorial의 일반화, 양의 정수 n에 대해 $\Gamma(n) = (n-1)!$

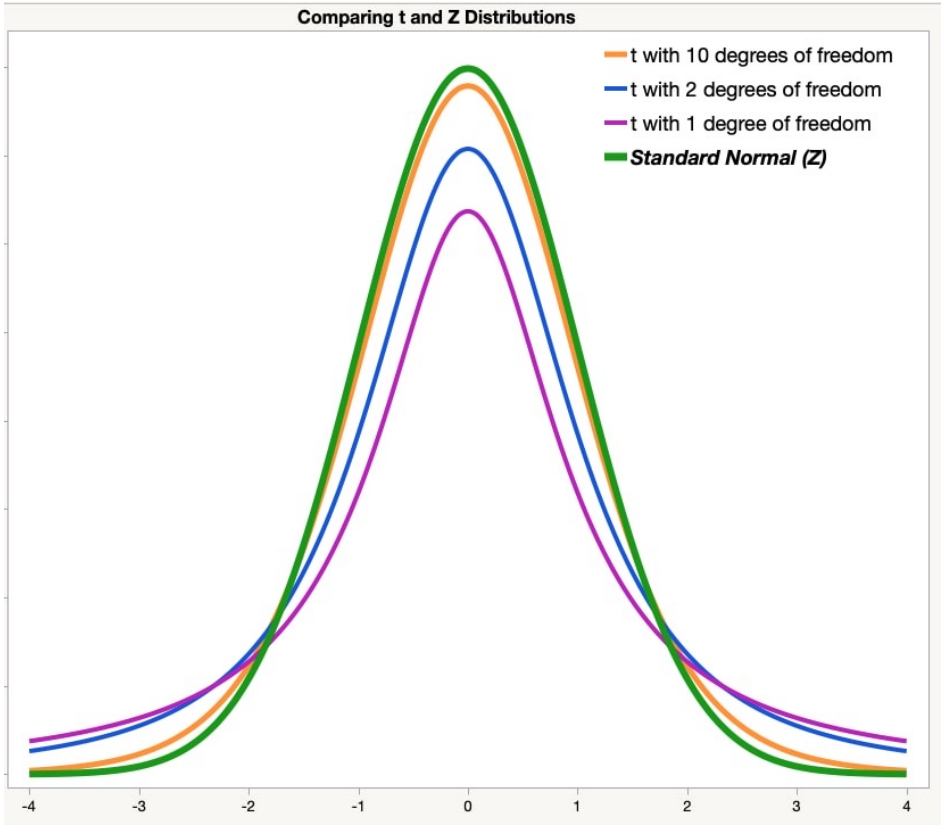
- t-score $t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$

  $\bar{X}$: Sample mean, μ: Population mean (가설로부터의 기대값)
  s: Sample 표준편차, n: Sample size

- Why t-distribution?
  - Sample size가 작을 때 (일반적으로 n < 30), Population의 표준편차를 정확하게 알 수 없음
  - t-dist는 Sample size가 작은 상황에서 Sample mean이 Population mean을 얼마나 잘 추정하는지를 더 정확하게 반영
  - t-dist는 양옆 꼬리가 더 두꺼워 극단값의 발생 가능성을 더 높게 평가

t-Distribution (2/2)

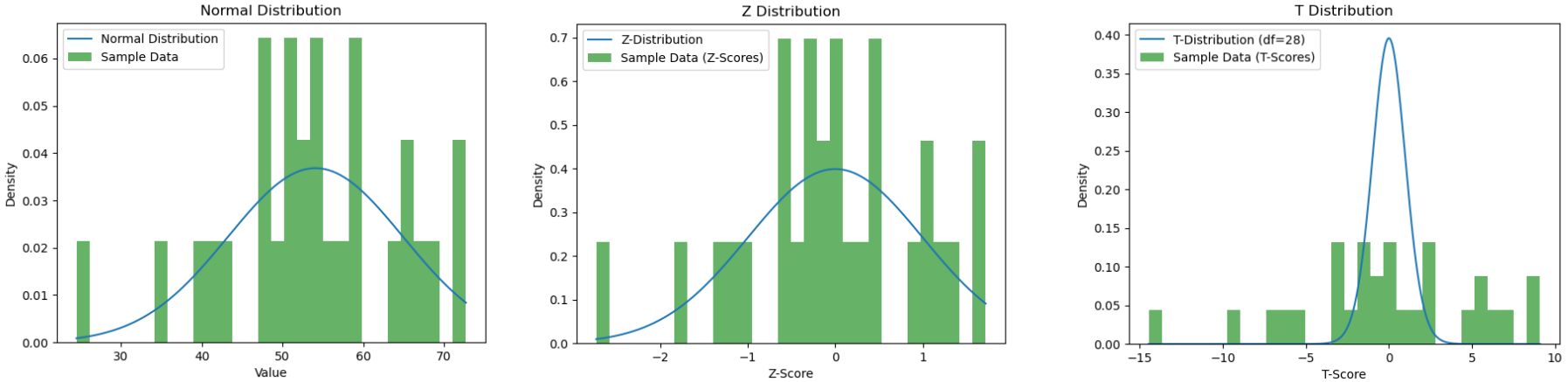
- Population의 표준편차를 모를 때
  - 대신 Sample 표준편차를 사용해야 함
  - Sample 표준편차는 Sample size가 작을 때 더 변동성이 심하므로 이를 보완하기 위해 t-분포를 사용
- Hypothesis testing 때
  - mean의 차이를 testing할 때 t-분포를 사용하여 보다 정확한 p-value를 계산
- t-dist 는 자유도 (df) 에 따라 모양이 달라짐
  - df가 커질수록 normal dist.에 가까와짐
  - df가 무한대로 간다는 것은 sample size가 충분히 커 진다는 것임. 그 때는 t-dist.가 normal dist.와 거의 동일하다는 것을 의미



t-score

CODE: 02_normalDistribution.py

- https://github.com/iklee99/StatCode
  - Data에 최적의 Normal distribution 계산
  - Data의 z-score 계산
  - Standard normal table 계산 (주어진 Z-score에 대한 확률계산)
  - 주어진 확률을 가지는 Z-score 계산
  - Data에 대한 t-score 계산
  - Plotting normal, z-dist., and t-dist. curves

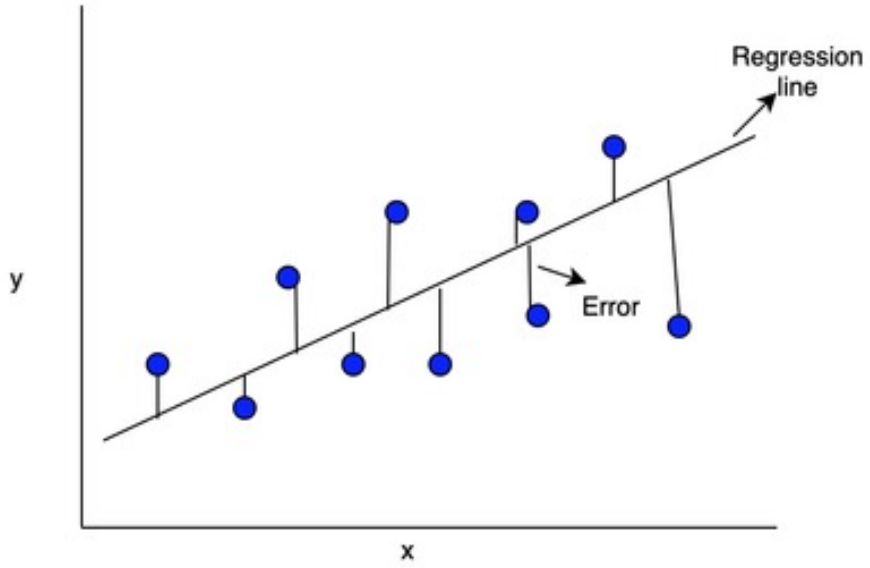


Normality Test

- Normality Test에 통과할 경우에는 Parametric test를 사용
- 통과하지 못할 경우에는 Nonparametric test를 사용

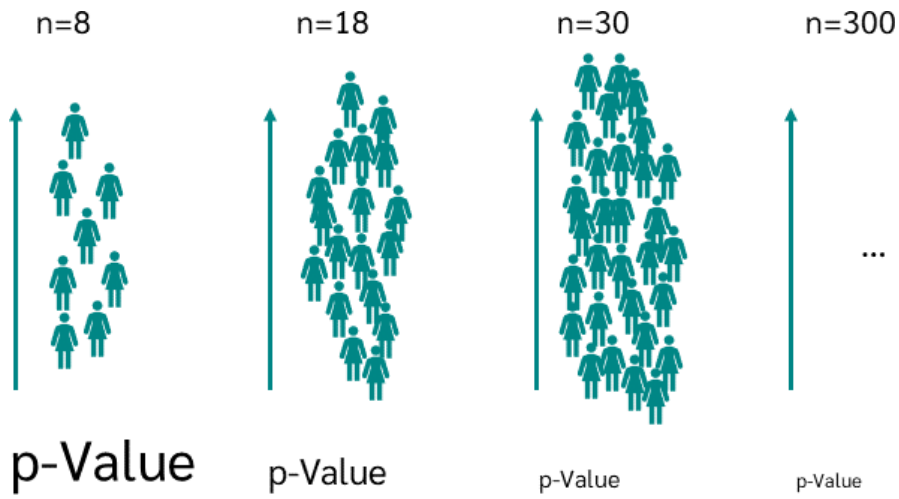
Normality Test in Linear Regression

- Dataset의 normality보다 model이 만들어내는 error가 normal distribution임을 확인해야 함



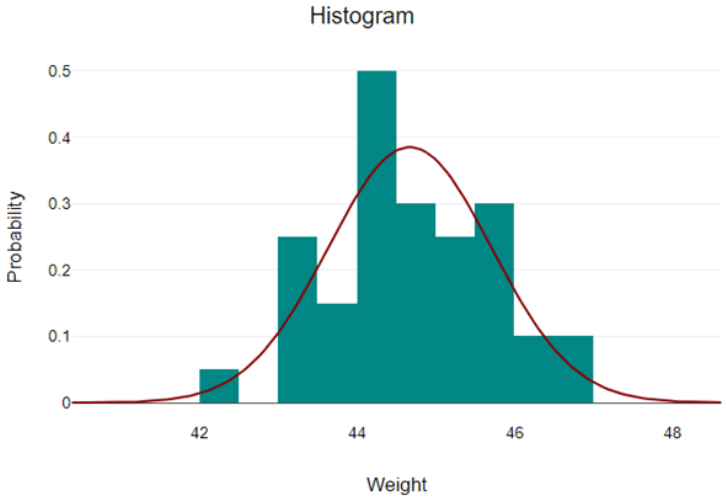
Normality Test Methods

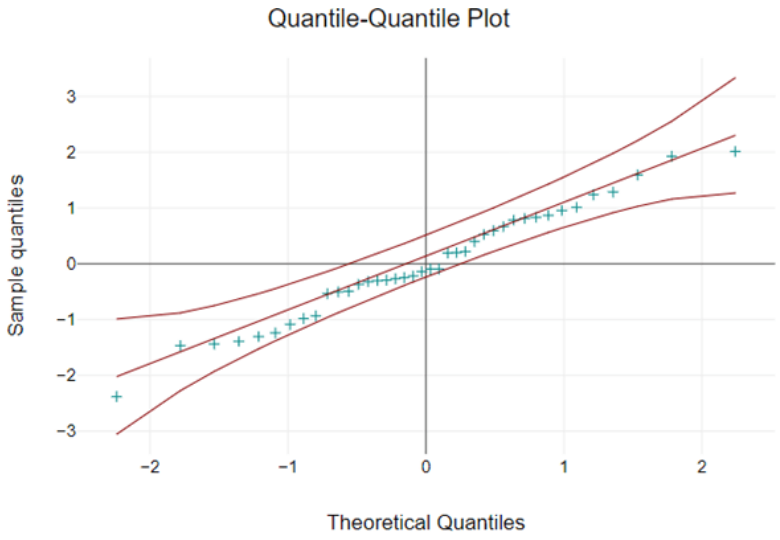
- Analytical Test Methods
  - Kolmogorov-Smirnov Test
  - Shapiro-Wilk Test
  - Anderson-Darling Test
- Procedure
  - H0: "Data가 normally distributed 되어 있다" 로 설정
  - 선택한 method로 p-value를 계산
  - p-value ≤ 0.05 이면 H0를 reject, 즉, data는 normal distribution이 아님
- Analytical Test의 단점
  - Sample 크기가 작을 수록 p-value가 더 커짐



Graphical Test

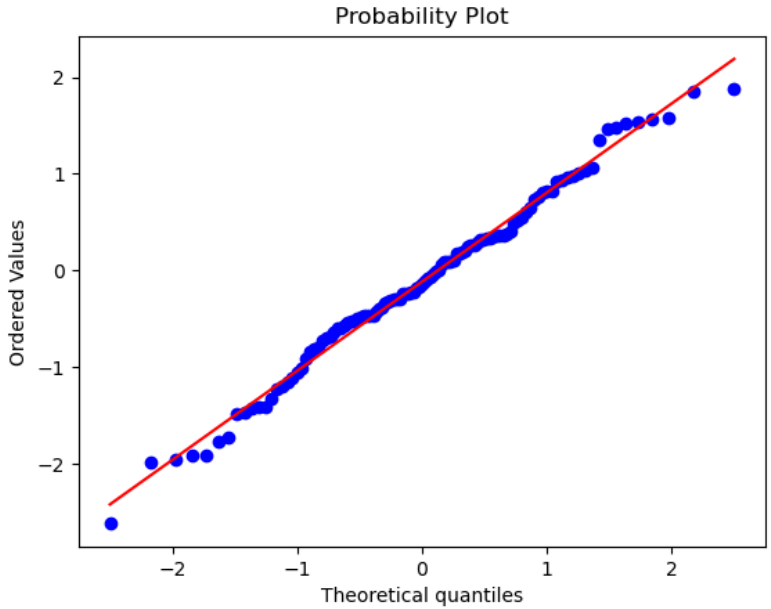
- Using Histogram
  - Histogram과 normal distribution curve를 겹쳐 그려서 차이를 관찰
- Q-Q Plot
  - Normality를 만족하는 이상적인 분포의 data curve들이 그려져 있음
  - 우리의 data를 plotting하여 normality curve들과 비교





CODE: 03_NormalityTest.py

- https://github.com/iklee99/StatCode
- 02_normalityTest.py
  - Analytical Test Methods
    - Kolmogorov-Smirnov Test
    - Shapiro-Wilk Test
    - Anderson-Darling Test
  - Graphical Test Q-Q Plot



Sample Size (Power Analysis)

- 최소 표본 크기 계산은 각 Statistical Test 마다 다름
- Sample Size 계산의 Parameter들:
  - Significant Level (SL, 유의수준): α
    - Type 1 error를 범할 최대 확률 (H0가 true인데 기각할 최대 확률)
  - Power: $1-\beta$
    - β: Type 2 error의 확률, 즉 H0가 false인데 기각하지 않을 확률
    - Type 2 error를 피할 확률 ($1-\beta$) 일반적으로 0.8 또는 0.9로 설정되며, 0.8보다 작게 설정하지 않음
  - Effect Size
    - 연구자가 검출하고자 하는 최소한의 effect size
    - Cohen's d, Odds Ratio, Pearson's r 등으로 표시
    - 보통 Cohen's d 를 사용할 때 0.2 ~ 0.5 로 사용

Effect Size – Cohen's d

- 각 group의 크기: n_1, n_2
- 따라서 $(n_1 - 1), (n_2 - 1)$ 는 df (자유도)
- 결합 표준 편차(pooled standard deviation):

$$\sigma_p = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}$$

- Cohen's $d =$ $\mu_1 - \mu_0$ / $\underline{\mu_1 - \mu_0}$

Sample Size for t-Test

- One-Sample and Paired t-Test

$$n = \left(\frac{\left(Z_{\alpha/2} + Z_{\beta} \right) \cdot \sigma}{\mu_1 - \mu_0} \right)^2$$

  - $Z_{\alpha/2}$: α에 대응하는 Z-score (1.96 for α = 0.05, 2.5758 for α = 0.01)
  - Z_{β} : Power $1 - \beta$ 에 대응하는 Z-값 (보통 0.84 for 80% power)
  - σ: 표준편차
  - $\mu_1 - \mu_0$: 기대되는 효과 크기, effect size

- Independent Samples t-Test

$$n = \frac{2 \cdot \left(Z_{\alpha/2} + Z_{\beta} \right)^2 \cdot \sigma^2}{(\mu_1 - \mu_0)^2}$$

Sample Size for Other Tests

- ANOVA (One-way)

$$n = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2 \cdot 2 \cdot \sigma^2}{(\mu_{\max} - \mu_{\min})^2}$$

  - $\mu_{\max}$: 여러 sample mean 중 최대값
  - $\mu_{\min}$: 여러 sample mean 중 최소값

- Chi-Square Test

$$n = \frac{\left(\sum_i \left(E_i \cdot Z_{\alpha/2} + \sqrt{E_i} \cdot Z_{\beta}\right)\right)^2}{\left(\sum_i (O_i - E_i)\right)^2}$$

  - E_i: Expected Frequency (기대빈도)
  - O_i: Observed Frequency (관찰빈도)

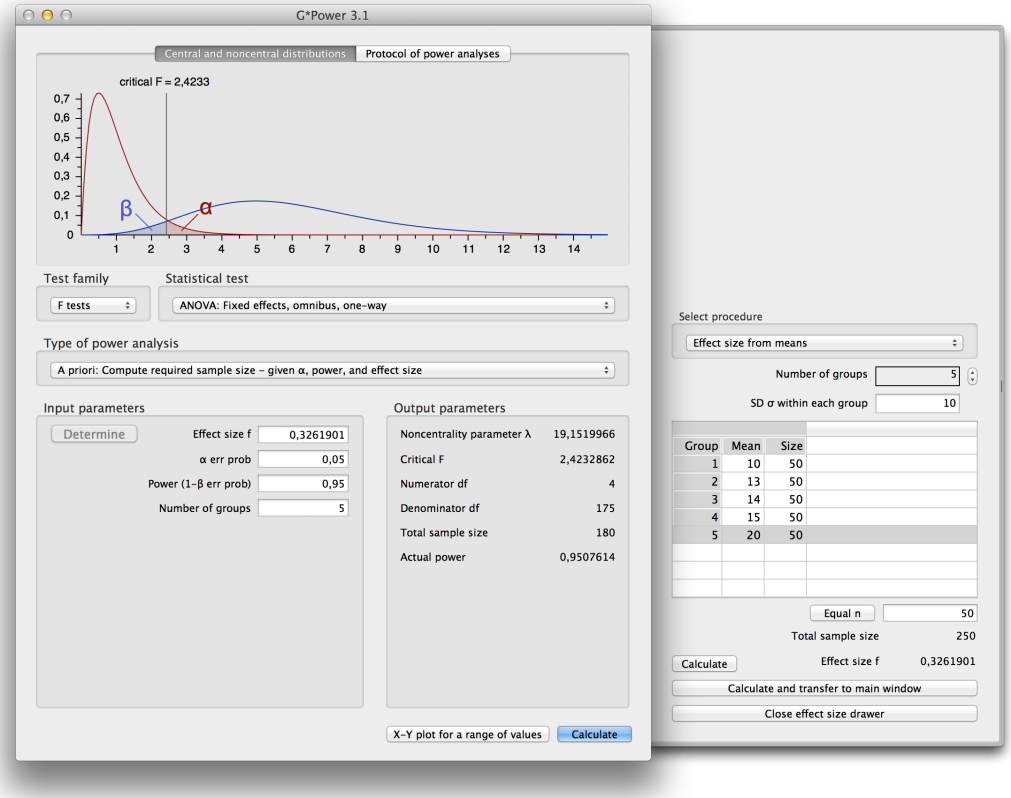
CODE: 04_SampleSize.py

- https://github.com/iklee99/StatCode
- 다음 Hypothesis Test 를 위한 최소 Sample size 계산
  - One-Sample t-Test
  - Independent Samples t-Test
  - Pared t-Test
  - Binomial Test
  - Chi-Square Test
  - One-Way ANOVA
  - *Two-Way ANOVA
  - *Two-Way ANOVA with Repeated Measures
  - *Mann-Whitney U Test
  - *Wilcoxon Test
  - *Friedman Test
  - Kruskal-Wallis Test
  - Pearson Correlation
  - Spearman Correlation
  - Point-Biserial Correlation
  - Linear Regression
  - Logistic Regression

  * 표시된 것은 불완전하게 계산됨.

UTIL: G-Power

- https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower



Sample size는 클수록 좋을까?

- 과학이 아닌 공학을 전공한 많은 연구자들은 피험자가 많을수록, 항상 더 좋다고, 잘못 생각합니다. 그러나 이는 사실이 아닙니다.
- User test의 피험자 수는 **Power Analysis**에서 구해진 수와 같아야 하며 그 이상도 이하도 아니어야 하며, 그 이유는 다음과 같습니다.
- 과거에는 과학 분야의 많은 연구자 (그리고 현재 공학 분야의 연구자?) 가 20명의 참가자를 대상으로 실험을 실행하고 그 결과 데이터를 분석했습니다.
- 원하는 결과가 유의미한 (significant한) 결과에 도달하지 못하면 5명을 더 실행하고 다시 분석합니다.
- 그래도 여전히 유의미하지 않으면 유의미해 질 때까지 5명을 더 실행하는 등의 과정을 반복한 다음 거기서 멈춥니다. 예를 들어, 30명의 참가자로 significant 하다는 결과가 나오면, 거기에서 중단합니다.
- 그러나 참가자를 5명 더 늘리면 35명에서 또다시 유의미하지 않으며, 40명에서 또다시 유의하지 않음을 발견할 수 있을 것입니다.
- 즉, 30명이라는 특정 표본 수는 뭔가 특이한 것이었으며, 30명 위아래의 다른 모든 표본 크기에서 볼 수 있듯이, 이 가설은 (이런 식으로 실험을 했다면) **전체적으로는 유의미하지 않은 것으로 보아야 하는 것입니다.**
- 문제는 30명을 대상으로 한 잘못된 '유의미한' 결과가 논문으로 발표된다는 것입니다.
- 이 때문에 대부분의 심리학 학술지(예: 심리학 저널)에서는 Power analysis를 포함하지 않거나 피험자 수가 Power analysis에 명시된 것과 다른 경우 원고를 데스크에서 거부합니다.
- 즉, 너무 많은 피험자를 대상으로 하는 것은 연구자의 조작을 나타낼 수 있으므로 부적절합니다.
- 이 요건은 공학 분야에서는 덜 일반적이지만, power analysis가 부족하거나 따르지 않은 경우 검토자가 직접 나서서 거부를 권고하기 시작했습니다.
- 실제로 가상 현실 연구에서는 장비 고장 등으로 인해 피험자 몇 명의 데이터를 사용하지 못하는 경우가 흔합니다. 따라서 연구가 약간 부족하거나 약간 초과된 상태로 끝나더라도 양해를 해 줄 수 있을 것입니다. 연구자가 데이터 손실에 대비하여 몇 번 더 실행할 계획을 세웠다면 괜찮을 것입니다.
- 따라서 Power analysis 에 표시된 정확한 숫자를 기대하는 것은 너무 엄격할 수 있지만 그 숫자는 비슷해야 합니다.