I. Basic Statistics

Probability : 수학의 영역.

Statistics : Data → Probability → 의사 결정.
기록, 측정. / 비율. / test.

Data type : Categorical / ORdinal / Continuous / Discrete.

Distribution of Random Variable → Emperical Distribution (Sample version)

Graph : Histogram / Boxplot
CDF, Scatter plot

Mean / Variance and the other moments.

Q. 이걸 왜 보는 걸까?
Distribution의 모양을 확인하기 위해서.

Bayesian vs. Frequentist
: 파라미터 가는 목표, 목적을 추정하는 걸로 바라보고 생각해 보자.

Properties from Population.

↓

Sample of population → Estimate & iid → Sample Mean / Variance, Median, Percentiles, IQR, Mode, Range

: Useful
Not perfect. (Caution)

Q. iid 깨지면 어떻게 되나요?
A. MLE를 사용하기 곤란해집니다.
Maximum Likelihood에서 production (Π)을 사용하지 못 하기 때문.

And iid라는 가정 자체가 Sample들이 모두 동일한 성격을 유지하고 있다는 것. 하나 튀어서 평균, 분산이 빠지면 혼란스러워 진다.

→ Expectation으로 평하나 결정하기 쉽다.

Distribution

Bernoulli : 1번 시행

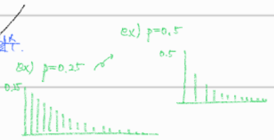
↓ Generalization

Binomial (Success) : n번 시행. k번 성공할 확률.

Hypergeometric
: 비복원 추출
성공 k번 나올 확률

Negative Binomial
: 복원 추출.
성공 k번 나올 때까지

성공 횟수.



Geometric (Fail) : 성공 1번 나올 때까지.
Generalization

Q. 그래프 둘 모두 가면 갈수록 확률이 작아지는 이유는?
A. 성공이든 실패든 확률은 1보다 작으니까.

Continuous

Log Normal : Gaussian을 exponential한 함수.
log transform으로 정규분포 만듦.

Student's t
: Gaussian을 따르는 변수면서 Sample 수가 적을 때

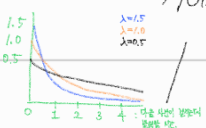
정규화 Continuous

Normal : 종모양.
$p \to 0$ 일 때

포아송 분포의 이항 근사

Poisson : 희박하지만 발생.
$p \to 0$ 일 때
단위 시간 안에 사건이 몇 번 발생하는지는 말은 평균 (상수) λ번 발생하는 사건이 있을 때, 단위 시간 동안 사건이 λ번이나 k번 발생하는지를 묻는 거고 포아송 분포는 λ번을 중심으로 확률을 설명해주는 것.

핵심은 단위 시간 혹은 공간에서 사건이 1번 발생할 확률이 극도로 희박해서 분포가 0에 몰려 있는 것.



Exponential
: 성공 1번 나올 때까지 대기 시간

* Poisson은 성공을 나타내는 Binomial의 연속 버전. Exponential은 실패를 나타내는 Geometric의 연속 버전.

Generalization

chi-squared
: $N(0,1)$의 제곱합.
2nd moment의 동향을 살피는 데 유용.

*Scale 이슈 때문에 Data를 Standardization한다면. $N(0,1)$을 만족하고 Data가 정규 분포를 따른다고 했을 때 우리는 2nd moment 즉, Data가 퍼져 있는 정도를 보면 된다.

$\alpha = \frac{1}{2}$
$\beta = 2$

→ Vector
*Multinomial
: n번 시행해서 각각의 값들이 k번 나타날 확률.
(주사위 결과들이 나올 확률)

Beta ⟷ Gamma
: 성공 k번 나올 때까지 대기 시간

Generalization

Bayesian의 Prior로 활용.

Uniform : 모든 값이 나올 확률이 동일.

Continuous

Dirichlet
: $\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$, $\sum x_i = 1$, $x_i \ge 0$
: 모든 합이 1이 되게 하는 분포.

*분포에 대해 알려는 이유? 우리가 Control할 수 없는 요인들을 control 하려고 하지 말고 (Random이니까?)

최소한의 노력으로 평균적으로 최소한만 틀리도록 하라는 것. (Expectation을 보는 이유 중 하나)

C.I & Test.

- Estimator : $\boxed{\text{Mean}}$ & Estimator. ⟶ Confidence Intervals:

$-se \cdot z \quad + se \cdot z$ (Mean)

Z : critical value at α.

- CLT

: Mean 1, 2, 3, ⋯ from Sample.

$\boxed{\rightarrow N(\mu, \frac{\sigma^2}{n})}$ → Hypothesis testing.

Variable : $X_i \rightarrow E(\bar{X}) = \frac{\Sigma}{n}\mu = \frac{n \cdot \mu}{n} = \mu.$

$Var(\bar{X}) = \frac{\Sigma}{n^2}\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

※ Sample set # vs. # of one sample set data.
(Mean) (Variable)

Poisson ver. t-test (Mean = Variance)

$H_0 : \lambda = \theta$
$H_1 : \lambda < \theta$ ($>$, $\neq$)

Reject H_0 if $P(X \le x) \le \alpha$ (nominal significance level) $\Longleftrightarrow$ $|p| = \left|\frac{\hat{\lambda} \pm \theta}{\frac{\theta}{\sqrt{n}}}\right| \le z_\alpha$ (one-tail), $z_{\alpha/2}$ (two-tail)

$\ge$: $P(X \le x) \le \frac{\alpha}{2}$ or $P(X \ge x) \le \frac{\alpha}{2}$

⇒ Test at the nominal significance level.

t-test

*Why? 정규분포를 가정으로 small sample일 때, t dist^{n}이 나오는 거니까.

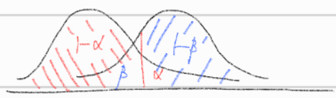
$t = \frac{\hat{\mu} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$, $H_0 : \mu = \mu_0$

if σ is unknown,
use $\hat{\sigma} = \frac{\Sigma(\ )^2}{n-1}$

A/B Tests (Two-sample tests)

: pooled estimation으로 sample간 같은 분산을 가지게 해줬다. (Homoskedastic)

Type I error : H_1 선택해서 틀림.
vs
Type II error : H_0 선택해서 틀림.



α : H_0 참일 때, H_1이 틀릴 확률.

$1-\alpha$: H_0 참일 때, H_1이 맞을 확률.

β : H_1 참일 때, H_0이 틀릴 확률.

$1-\beta$: H_1 참일 때, H_0이 맞을 확률.

↳ 대수의 법칙 기억하기.

*Q. 대체 α가 낮아지면 $1-\beta$가 낮아지는 것인가?

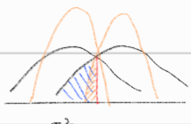
α↓는 H_0가 참일 때 H_1틀릴 확률이 낮아지는 것이므로 H_1가 참일 때 관점으로 보면 H_0의 틀릴 확률(β)이 높아지는 것으로 생각할 수 있고(β↑) 맞을 확률이 낮아진다고 생각할 수 있다($1-\beta$↓)

H_0를 바탕으로 우리의 estimation이 진행됐고 test-statistics이 나옴.

☆ 결국 H_1과 H_0 둘 중 하나는 맞아야 하고 서로의 가정 속에서 누군가 맞을 확률이 높아지면 반대쪽은 틀릴 확률이 높아지는 것.

* Data 1000개 짜리 Data set 하나.
 " 30개 짜리 " 30개
둘은 뭐가 다를까?

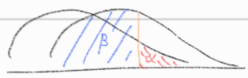
변형 cases : ① 분산 & α가 변할 때, $1-\beta \propto \frac{\alpha}{\text{분산의 크기}}$



⇒ $\frac{\sigma^2}{n}$니까 데이터가 커지면 됨 (길쭉해지는 data set)

*Tip. α와 β는 trade off 관계니까 model의 power를 높이기 위해서는 data를 많이 모아야 값을 알 수 있다.

② 분포 함수의 모양이 달라질 때,



정규분포를 바탕으로 한 가정들이 깨지며 $1-\beta$ 또한 얘기하기 힘들어짐.

(3rd, 4th moment까지 고려해야 한다.)

QnA

Q. $y \leftarrow x_1$으로 설명하는 게 부족. 여러 개의 variable로 설명된다?

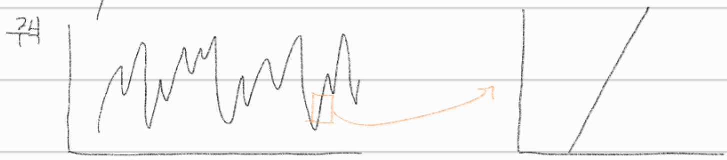
A. ① 설명력이 부족하다. R^2 or R^2_{adj}

② x_1으로 설명하고 남은 error 부분이 x_1에 따라 움직인다.

↳ Regression 적용 후 $E(\varepsilon)=0$ 이 되어야 한다. ($y = X\beta + \varepsilon$) But expectation을 쓴다는 것 자체가 이미 error가 random 이라는 것.

그런데 x_1에 따라 error가 결정되는 것 같다? error가 random이라는 기본적인 가정이 깨진 것.

③ Population의 특정 부분만 본 것.(Selection Bias)



이 주식에 투자하시면 정말 떼돈 버실겁니다?

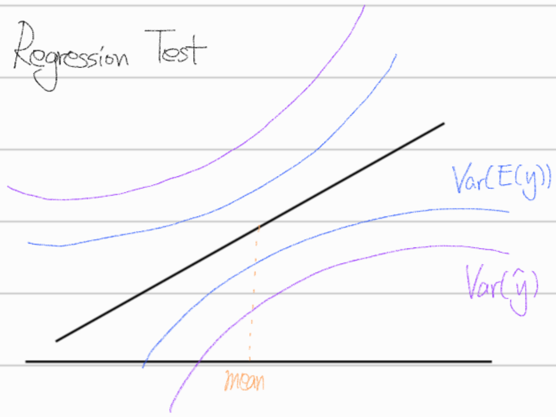
④ 시간이 지나면서 Data shape이 변하는 경우.

④-1. 단순히 평균, 분포만 바뀌는 경우. → 시간이라는 변수를 통해 control하거나 test를 따로 진행. *왜? 설명력이 일정하지 않아서.

④-2. 분포가 달라짐. → 평균은 CLT 개념하에 t-test 진행.

평균 아니면? test 자체가 바뀌어야 함. 지금은 y를 x_1으로 설명하고 싶은 상황.

Regression Test



*Variance는 왜 더 커질까?

Mean(expectation)에서 멀어지면서

$\hat{\varepsilon}$이 커진다. Regression은 평균에 맞춘 계산이니. 평균에서 멀어질수록 편차가 커져서 분산이 커짐.

$\hat{y}$를 예측하는 것과 $E(y)$를 예측하는 것의 차이는 ε의 유무다.

$E(y|X) = E(X\beta|X) + E(\varepsilon|X) = E(X\beta|X)$: 검은 선을 의미. 검은 선만 결과로 나옴.

$Var(y|X) = \underbrace{Var(X\beta|X)}_{E(X\beta|X)\text{의 variance}} + \underbrace{Var(\varepsilon|X)}_{\text{얘가 여기에 추가된 것.}}$

∴ 단지 ε의 유무로 분산에 차이가 생긴 것.

*추가 설명 degrees of freedom

Data가 주어져 있을 때 아무 값이나 될 수 있는 특권. $\bar{x}$라는 평균이 고정되어 있기 때문에 Data들의 자유도 1 감소 ex) $\bar{x}=2$ $x_1=1$ $x_2=2$ $x_3=?$

x_3은 아무 값이나 될 수 있는가?