KAIST GSDS Entrance Exam – Sample #1

Statistics. [Hint list] 아래 질문들에 답하는 데에 있어서, 다음의 수학적 사실들은 별도의 증명없이 활용가능하다.

(H1) 임의의 non-negative 확률변수 Y와 자연수 k에 대하여, $\mathbb{E}[Y^k] = \int_0^\infty ky^{k-1} \cdot \mathbb{P}[Y > y]dy$.

(H2) Jensen 부등식: 임의의 non-negative 확률변수 Y에 대하여, $\mathbb{E}[\sqrt{Y}] \leq \sqrt{\mathbb{E}[Y]}$.

(H3) 임의의 확률변수 Y와 실수 y에 대하여, $\mathbb{E}[Y] = \mathbb{E}[Y|Y \leq y] \times \mathbb{P}[Y \leq y] + \mathbb{E}[Y|Y > y] \times \mathbb{P}[Y > y]$.

(H4) $\int_0^\infty e^{-t^2}dt = \frac{\sqrt{\pi}}{2}$.

[문제 상황] 특수목적의 광원으로 사용되는 "SunLight"라는 모델의 전구의 수명에 관심이 있어서, 20개의 전구들에 대하여 수명 테스트를 진행하였고($n = 20$), 각 전구들에 대하여 다음과 같은 데이터 $(x_1, \ldots, x_n)$를 얻게되었다.

수명을 다한 전구들 ($n = 20$)	x_1	x_2	x_3	$\ldots$	x_{n-1}	x_n	$\frac{1}{n}\sum_{i=1}^n x_i$
고장까지 동작한 시간 (hours)	1,164	1,207	1,585	$\ldots$	886	922	1,200

"SunLight" 전구의 수명을 X라는 확률변수로 표현하고, 그 평균 수명을 $\mu = \mathbb{E}[X]$라고 하자. 아래 질문들에 답하시오.

1. 어떤 양의 실수 θ가 존재하여, 모든 양의 실수 x에 대하여 $\mathbb{P}[X > x] = \exp\left(-(\theta x)^2\right)$가 성립한다고 가정하자. (총 35점)

   (a) 확률변수 X의 확률 밀도 함수(probability density function)와 평균(μ) 및 second moment($\mathbb{E}[X^2]$)를 구하시오. (15점)

   (b) 최우추정법(MLE)으로 주어진 데이터셋을 사용하여 θ의 값을 추정하는 방법을 제시하시오. (10점)

   (c) 위 최우추정법으로 추정한 θ의 값을 이용하여 평균 수명을 추정하고자 한다. 이러한 추정량의 기대값과 실제 평균 수명 μ를 비교하여 보시오. (10점)

2. 위 실험을 진행한 담당자와 면담한 결과 다음과 같은 사실을 알게 되었다: "여건상 실험은 2000시간 동안만 진행될 수 밖에 없었고($t = 2000$), 해당 시간 동안 고장나지 않은 5개의 전구들에 대한 데이터는 위의 표에서 누락되었다($r = 5$)". 즉, 실험에 사용된 총 전구의 수는 $n + r = 25$개 였고, 그 중 위의 표에 기재된 $n = 20$개의 전구들은 2000시간 내에 수명을 다하였으며, 나머지 $r = 5$개의 전구들은 2000시간 동안 고장나지 않았다. (총 45점)

   (a) 위 표에 기재된 n개 전구들만의 평균 동작 시간$\left(\frac{\sum_{i=1}^n x_i}{n}\right)$으로 평균 수명을 추정한다면, 이는 불편추정량인가? 직관적인 논리를 제시하시오. (10점)

   (b) 전체 $n + r$개 전구들의 평균 동작 시간($\frac{\sum_{i=1}^n x_i + r \times t}{n+r}$)으로 평균 수명을 추정한다면, 이는 불편추정량인가? 직관적인 논리를 제시하시오. (10점)

   (c) 다음 질문들에 순서대로 답하시오.

      i. $t = 2000$일 때, 주어진 데이터로 부터 $\mathbb{E}[X|X \leq t]$와 $\mathbb{P}[X > t]$의 값을 추정하여 보시오. (10점)

      ii. 모든 양의 실수 x에 대하여 $\mathbb{E}[X|X > x] = \mathbb{E}[X] + x$가 성립한다고 할때, 이 조건을 이용하여 평균 수명을 추정하는 방법을 제시하시오. (15점)

Programming. 일변수 함수 $f(x)$는 구간 $[a, b]$에서 연속이며 단조 증가한다. 방정식 $f(x) = 0$의 해가 구간 $[a, b]$에 존재하는지 판별하고, 해가 존재할 경우 오차 범위 e내에서 해를 효율적으로 찾는 Python code를 작성하고, 그 **시간 복잡도**를 분석하시오 (Python 문법에 익숙하지 않다면 pseudo-code로 작성하시오). (20점)

보다 구체적으로, 다음 조건들을 만족하는 find_root(f, a, b, e) 함수를 완성하시오.

- 입력으로 함수 f (연속, 단조증가), 두 실수 a, b ($a < b$), 허용 오차 e이 주어진다.
- $f(x) = 0$의 해가 $[a, b]$ 내에 존재하지 않는 경우 None을 리턴한다.
- 해가 하나라도 존재한다면 e 범위 이내에 해가 존재하는 실수값을 리턴한다 (즉, 어떤 실수 x^*를 리턴했다면, $[x^* - e, x^* + e]$ 구간 내에 $f(x) = 0$을 만족하는 x가 존재하여야 한다).
- f는 함수이며, $f(a)$와 같이 하나의 실수입력을 파라미터로 제공하여 Python code에서 호출할 수 있다.
- f 함수에 대한 gradient 계산은 알려져있지 않으며, 오로지 f에 대한 계산만을 수행할 수 있다.

```
def find_root(f, a, b, e):
    # write your code here
```

사용 예시는 다음과 같다.

```
>>> def f1(x):
...     return x*x-2
>>> def f2(x):
...     return x*x+10
>>> find_root( f1, 0, 5, 0.1 )
1.484375
>>> find_root( f1, 0, 5, 0.00001 )
1.4142131805419922
>>> find_root( f2, 0, 5, 0.00001 )
None
```

KAIST GSDS Entrance Exam – Sample #2

Statistics 1. 농구 자유투를 던지는 실험을 생각해보자. 자유투의 성공여부가 매번 독립적이라 가정하자. 독립적인 자유투를 실패할 때까지 던져서, 성공한 갯수를 확률변수 X라 하자. 이때, x번 성공할 확률을 $\mathbb{P}(X = x) = \theta^x(1-\theta)$로 표현하자.

1. 이 분포의 이름은 무엇이고, parameter θ의 의미는 무엇인가?
2. $G(x) = \mathbb{P}(X \geq x)$를 구하시오.
3. 한 선수의 자유투 실험 결과가 다음과 같다.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	$\bar{x}$
4	0	1	2	6	0	3	1	10	3	3.0

   (a) θ를 추정하기 위한 직관적인 방법을 제시하시오.

   (b) θ를 최우추정법(MLE)으로 구하시오.

4. 자유투 실험을 할 때, 10회 연속 성공하면 $X = 10$이 되고 끝내는 규칙(S_{10} rule)이 있다면, 위의 데이터로부터 θ를 어떻게 추정할 수 있는가에 대한 다음의 질문에 답하시오.

   (a) S_{10} rule을 감안하면 θ의 추정치는 앞서 이를 감안하지 않고 추정한 값에 비해 커질까 작아질까?

   (b) S_{10} rule 하에서 θ를 추정하기 위한 직관적 방법을 제시하시오.

   (c) MLE를 이용한 θ 추정법을 제시하시오.

5. 관찰자 A가 선수 B의 자유투 연습을 지켜보고 있다. 이 선수 B의 성공률 $\theta_B = 0.8$ 정도로 추정된다. 선수 B가 운좋게 8개의 자유투를 연속으로 성공하였을 때, 선수 B가 갑자기 A에게 다음과 같이 제안하였다. "내가 이번에 자유투 두개를 더 성공하면 10개를 달성(=S_{10})합니다. S_{10} 달성 여부에 만원씩 걸고 내기할까요?" A는 이 제안을 받아들여야 할까?

Statistics 2. 신체검사에서 양팔의 길이를 측정하였다. (팔길이는 어깨부터 손끝까지 길이이고 mm 단위로 측정함.) 표와 같은 데이터를 얻었을 때 아래 질문에 답하시오.

i (피검자 번호)	1	2	3	$\cdots$	18(= n)	표본평균	표본표준편차
ℓ_i (단위:mm)	750	741	762	$\cdots$	730	$\bar{\ell} = 743.8$	$s_l = 3.4$
r_i (단위:mm)	751	740	764	$\cdots$	731	$\bar{r} = 744.5$	$s_r = 3.4$

1. "오른팔이 왼팔보다 길다"라는 가설을 통계적으로 검정하는 다음의 과정을 살펴보고 어떤 오류가 있는지 찾아보시오.

   (a) $H_0 : \mu_r = \mu_\ell$ vs $H_1 : \mu_r > \mu_\ell$, 유의수준 $\alpha = 0.05$ (여기서 μ_ℓ은 왼팔길이의 모평균, μ_r은 오른팔길이의 모평균)

   (b) 통합표본표준편차 $s = s_\ell = s_r = 3.4$

   (c) $df = 2n - 2 = 34$인 t-분포 사용: $t_{34,\alpha} = 1.69$ (혹은, $t_{34,\alpha} \cong z_\alpha = 1.645$ 사용)

(d) 검정통계량 $t_0 = \frac{\bar{r}-\bar{\ell}}{s\sqrt{2/n}} = \frac{0.7}{3.4/3} = \frac{2.1}{3.4} \cong 0.618 \ll t_{34,\alpha} = 1.69$.
(혹은 p-value $= \mathbb{P}(T_{34} > t_0|H_0) \cong \mathbb{P}(Z > t_0|H_0) \cong 0.27 \gg \alpha = 0.05$)

(e) 결론: 유의수준 $\alpha = 0.05$에서 H_0를 기각할 수 없음. 즉, $\mu_r > \mu_\ell$이라고 말할 수 없음.

2. 위의 과정에서 오류를 수정하여 올바른 검정절차를 제시하시오. (추가로 필요한 정보가 있으면, 명시하고 사용하시오)

Programming. 데이터 $D = [x_1, x_2, \ldots, x_n]$을 input으로 받아서, k-th smallest element를 효율적으로 찾는 code 또는 pseudo-code를 완성하고, 제시한 방법의 Worst Case Complexity를 말하시오. (단, sorting을 사용하지 마시오.)

KAIST GSDS Entrance Exam – Sample #3

Statistics 1. A factory operates with two machines of type I and one machine of type II. The weekly repair cost, X, for each of type I machines is normally distributed with mean μ_1 and variance σ^2. The weekly repair cost, Y, for a type II machine is also normally distributed, but with mean μ_2 and variance $3\sigma^2$. Therefore the expected repair cost per week for the factory is $2\mu_1 + \mu_2$. Suppose you are given a random sample $X_1, X_2, \ldots, X_n$ on the repair costs of type I machines, and an independent random sample $Y_1, Y_2, \ldots, Y_n$ on the repair costs of type II machines. That is, $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n$ are all independent.

1. Show how you would construct a 90% confidence interval for $2\mu_1 + \mu_2$.
   (a) if σ^2 is known.
   (b) if σ^2 is not known.
2. Explain how you would test whether the mean repair costs of type I and II machines are the same or not. Write appropriate hypotheses, specify the test statistic, and its probability distribution under the null hypothesis.

Statistics 2. A married couple, say Parks, have their child in daycare. They sometimes arrive late to pick up their child from daycare, which has a strict policy requiring parents to be punctual. To enforce this policy, the daycare charges 50 cents per minute for tardiness. Assume that the daily delay in picking up their child, represented by variable X, follows an exponential distribution with a mean of 6 minutes.

1. Assuming that their child will attend daycare for 100 days this year, what is the probability distribution of the total duration (in minutes) of their tardiness during this period? What assumption is needed for this conclusion?
2. Approximately calculate the probability that they will pay more than $315 in late fees during 100 days. Use Central Limit Theorem.
3. Suppose that the daycare changes the rate to $(x^2 + x)/10$ dollars for x minutes of tardiness. On average, how much will Parks pay in late fees each day?
4. Another married couple, say Kims, also enroll their child in the same daycare, where their daily delay in picking up the child, represented by variable Y, follows an exponential distribution with a mean of 5 minutes, independent of Parks' delay represented by X. What is the probability that Parks will pay more late fees than Kims on any given day?

Standard Normal Probabilities

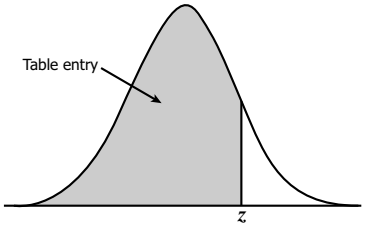


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

KAIST GSDS Entrance Exam – Sample #4

Smart기기를 이용하여 숙면에 빠지기까지 걸리는 시간 (확률변수 X)을 조사하여 다음과 같은 data를 얻었다.

Table 1: Time to Sleep

i	1	2	3	4	5	6	$\cdots$	$10,000(=n)$	$\bar{x}_n$
x_i (단위:분)	7.5	3.1	20	5.5	60.0	1.2	$\cdots$	16.0	10.0

아래의 질문에 답하시오.

1. 데이터 $D = [x_1, x_2, \ldots, x_n]$을 input으로 받아서, 최소값과 최대값 사이를 k개로 나눈 bin에 포함되는 관측치의 빈도수를 count하는 histogram을 계산하는 함수를 완성하시오.

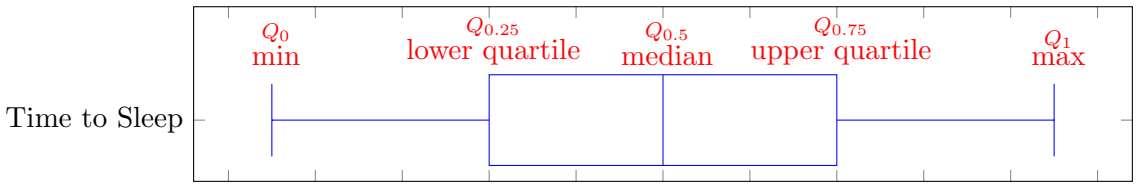
> `HISTOGRAM` (D, k)
> `Input` D: array, $D[i] = x_{i+1}$, $i = 0, \ldots, n-1$
> k: # of bins, $k > 0$.
> `Output` H: array $H[i]$: count for bin i, $i = 0, 1, \ldots, k-1$

2. 확률변수 X의 확률밀도함수를 $f(x) = \lambda e^{-\lambda x}$라고 할 때, 위의 데이터로 부터 λ를 점추정(point estimation)하고자 한다.
   (a) 최우추정법(MLE)에 의한 추정량 $\widehat{\lambda}_n$을 유도하시오.
   (b) $\widehat{\lambda}_n$과 다른 추정량을 하나 제시하고, 어느 추정량이 좋은지와 그 판단의 근거를 말해보시오.

3. 앞서 MLE로 구한 $\widehat{\lambda}_n$에 대한 다음의 질문에 답하시오.
   (a) $E[\widehat{\lambda}_n]$를 구하시오.
   (b) $\widehat{\lambda}_n$는 biased되어 있는가?
   (만일 $E[\widehat{\lambda}_n]$를 구하지 못했다면, Jensen's Inequality를 이용하여 bias여부를 확인해도 됨)
   (c) $\widehat{\lambda}_n$는 consistent한가?
   (d) $\widehat{\lambda}_n$이 biased되어 있다면, bias를 보정한 불편추정량(unbiased estimator)을 제시하시오.

4. 이 기기는 battery절약을 위해서 60분이 지나면 $x_i = 60$으로 기록하고 숙면 monitoring을 멈춘다. 이를 고려하여 λ에 대한 추정량을 MLE (Maximum Likelihood Estimation)로 유도하시오.

5. 위의 data를 box plot으로 나타내고자 한다.



   (a) X의 분포가 앞서 가정한 지수분포가 맞다면, median과 mean중 어느 것이 더 클까?
   (b) X의 분포가 앞서 가정한 지수분포가 맞다면, $Q_{0.25} - Q_0$과 $Q_1 - Q_{0.75}$ 중 어느 것이 더 클까?

(c) 데이터 $D = [x_1, x_2, \ldots, x_n]$ 으로 부터 box plot을 계산하기 위한 함수를 구현하기 위한 방법을 제시하시오. (아래의 예시 답안과 다른 효율적인 방법 제시 요망.)

```
def box_plot(D):
    N = len(D)
    sort D by ascending order
    Q_0=D[N*0]
    Q_0.25=D[round(N*0.25)]
    ...
    Q_1=D[N-1]
    return Q_0, Q_0.25, Q_0.5, Q_0.75, Q_1
```