이제 몸풀기는 그만하고 본격적으로 통계학의 꽃인 통계적 추정을 공부해봅시다.
여기서부터는 시리어스한 내용이니까 조금 진지하게 갑시다. 앞서 우리가 정리했던 내용을 끌고오면

$$Let\ X_i\ be\ a\ sample\ f\ rom\ X\ where\ mean\ and\ variance\ are\ \mu\ and\ \sigma^2$$

$$and\ \overline{\overline{X}} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$then\ we\ know\ that$$

$$E\left[\overline{\overline{X}}\right] = \mu\ and\ Var\left(\overline{\overline{X}}\right) = \frac{\sigma^2}{n}$$

한글이랑 영어 섞어쓴다고 재수없다고 하지 마세요.. 저도 1학년 때는 수학 전공책을 전부 한글로 번역하고 싶은 꿈이 있었습니다. 근데 안되더라구요. 서울대 교수님들께 이런 면에서는 경의를 표합니다. 근데 영어는 좀 똑바로 쓰라구요? 알게 뭡니까 코쟁이들 문자 따위. 아무튼 다시 진지해집시다.

우리는 표본평균의 평균은 모집단의 평균과 같고, 표본평균의 분산은 모집단의 분산을 n으로 나눠준 값과 같다는 것을 배웠습니다. 이 때 표본평균의 평균이 모집단의 평균과 같다는 사실에 주목해봅시다.

우리가 통계를 배우는 목적은 채집한 어떤 샘플로부터 모집단에 관한 어떤 결과를 도출해내기 위함입니다.
그런데 과연 이 샘플이 얼마나 공정한가? 에 대한 질문은 모집단의 성질에 관한 결과를 도출함에 있어서 상당히 중요한 관점을 제공합니다.

예를 들어봅시다. 정치떡밥을 끌고오는 것은 싫어하지만, 최근 대선이 있었습니다. 이 대선에서 후보자의 지지율을 추론하기 위해서 어떤 샘플을 추출했다고 합시다. 전통적으로 보수적인 성향이 강한 TK에서 샘플을 추출한다면? 아니면 전통적으로 진보 성향이 강한 전남에서 sample을 추출한다면? 우리의 추론은 정확할까요?

이런 관점에서 통계학자들이 비편향추정량이라는 개념을 정의했습니다. 너무 복잡한 내용은 생략하고 핵심만 정리하자면,

$$Let\ \hat{\theta}\ be\ a\ esimator\ of\ a\ sample\ \ and$$
$$\theta\ be\ a\ estimate\ of\ population$$
$$then\ if\ E\left[\ \hat{\theta}\ \right] = \theta, we\ call\ \hat{\theta}\ as\ unbiased\ estimator$$
$$if\ not,\ \hat{\theta}\ called\ biased\ estimator$$

영어는 존잘님들이 좀 수정해주시면 굽신거리면서 멍멍 짖어보겠습니다.

자, 우리가 샘플로부터 모집단의 성질을 추정하기 위해 정의한 어떤 값을 추정량($\hat{\boldsymbol{\theta}}$)이라고 합시다. 이 때 이 추정량의 평균이 모집단의 성질과 같다면, 우리는 이 추정량을 편향되지 않게 추출되었다고 볼 수 있습니다. 왜 그런지는 확률론에서 설명하겠습니다. 평균의 정의로부터 도출되는 직관적 정의입니다.

예를 들면, 어떤 표본이 주어졌을 때, 모집단의 평균에 대한 추정량을 표본의 평균으로 정의한다면, 표본평균의 평균은 모집단의 평균과 같아지기 때문에, 어떤 sample을 가져오더라도, 편향되지 않은 값이라고 간주한다는 것 입니다. 자, 그런데 여기서 한 가지 이슈가 생깁니다. 분산에 대한 이슈입니다.

Let X_i be a sample from X where $(mean, var) = (\mu, \theta)$

and $\overline{X} = \frac{1}{n}\sum_i X_i$.

now we define a estimator for variance of X as $\overline{S}_n^2 = \frac{1}{n}\sum_i \left(X_i - \overline{X}\right)^2$

then

$$E\left[\overline{S}_n^2\right] = E\left[\frac{1}{n}\left(\sum_i \left(X_i - \overline{X}\right)^2\right)\right] = \frac{1}{n}E\left[\sum_i X_i^2 - \sum_i 2 \cdot X_i \cdot \overline{X} + \overline{X}^2\right]$$

$$= \frac{1}{n} \cdot \left(\sum_i E\left[X_i^2\right] - 2 \cdot \sum_i E\left[\overline{X}\right]\sum_i E\left[X_i\right] \cdot + \sum_i E\left[\overline{X}^2\right]\right)$$

we know the fact that $Var(X) = E\left[X^2\right] - E[X]^2$

thus,

$$E\left[X_i^2\right] = E\left[X_i\right]^2 + \sigma^2 = \mu^2 + \sigma^2$$

$$E\left[\overline{X}^2\right] = E\left[\overline{X}\right]^2 + \frac{\sigma^2}{n} = \mu^2 + \frac{\sigma^2}{n}$$

hence

$$\frac{1}{n} \cdot \left(\sum_i E\left[X_i^2\right] - 2 \cdot \sum_i E\left[\overline{X}\right]\sum_i E\left[X_i\right] \cdot + \sum_i \left[\overline{X}^2\right]\right)$$

$$= \frac{1}{n}\left(\sum_i E\left[X_i^2\right] - \sum_i E\left[\overline{X}^2\right]\right) = \mu^2 + \sigma^2 - \mu^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

thus $\overline{S}$ *is not a biased estimator.*

then if we take $\hat{S}_n^2 = \frac{n}{n-1}\overline{S_n}^2$, *as an estimator of variance*,

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_i \left(X_i - \overline{X}\right)^2 \text{ and } E\left[\hat{S}_n^2\right] = E\left[\frac{n}{n-1}\overline{S_n}^2\right] = \sigma^2$$

thus $\hat{S}_n^2$ *is the biased estimator of X*

만약 우리가 모집단의 분산을 추정하기 위한 추정량으로 표본분산을 가져온다면, 표본분산의 평균은 실제 분산보다 (n-1)/n 배 만큼 차이가 나게 됩니다. 그래서 이 추정량을 편향되지 않게 조정해주기 위해 표본분산에 n/(n-1)만큼을 곱해주면, 분산의 비편향추정량을 얻게됩니다. 이게 표본의 분산을 n-1로 나눠주는 이유 입니다.

이상으로 어렵고 지겨운 얘기는 그만하고, 이제 우리가 관심있는 모집단에 대한 추정으로 들어가보겠습니다.

물론 여기서 눈치가 빠르거나, 직관력이 좋거나 하신 분은 n이 충분히 커질 때, 분산을 굳이 n-1로 나누지 않아도, 추청량이 근사한다는 것을 알 수 있을겁니다. 이런 관점에서 출발한 정리가 바로 큰 수의 법칙과, 중심극한정리, 그리고 T 검정과 Z검정에 대한 내용입니다. 자세한건 뭐다? 확률론에 짬을 때리겠다 이마립니다.