Frequentist and Bayesian Statistics; Different Philosophy, Different Approach

2020 여름방학 베이지안 머신러닝 스터디 1주차 자료

경제학과 13학번, ESC 22기 강경훈

1. Probability Densities와 Likelihood

어떤 확률 변수 x_i가 가질 수 있는 값들을 sample space $\mathcal{X}$라고 하고, 그 값들의 분포는 어떤 모수 θ에 의해 완전히 결정되는 함수 $f(x|\theta)$라고 생각해봅시다(예컨대 이항분포나 분산이 주어진 정규분포 등을 생각해볼 수 있겠습니다). 모수 θ가 가질 수 있는 값들은 parameter space Ω라고 합니다. (이때 모수 θ는 스칼라가 아니라 벡터일 수도 있습니다. 여기서는 스칼라인 경우만 일단 생각해볼게요.)

$$\text{Sampling Density of } x_i\text{:} \quad f(x|\theta) \quad (x \in \mathcal{X}, \theta \in \Omega) \tag{1}$$

이때 함수 f를 probability density라고 합니다. (x가 이산형일때는 probability mass function이라고도 하는데, 여기서는 편의상 그냥 density라고 퉁치겠습니다.) 이와 똑같은 분포를 따르며 독립적으로 샘플링된(i.i.d라고 하지요) 샘플들이 $x_1, x_2, x_3, \dots$ 이렇게 있으면 우리는 데이터의 벡터 $\mathbf{x}$를 생각해볼 수 있겠습니다.

$$\text{Joint Sampling Density of } \mathbf{x} = [x_1, x_2, \ldots, x_N] \quad p(\mathbf{x}|\theta) = \prod_{n=1}^{N} p(x_n|\theta) \qquad (2)$$

앞서 말한 것처럼 density는 모수 θ에 의해 전적으로 결정되는 함수입니다. 그런데 생각해보면, 우리는 모수 θ를 모른 상태에서 오직 하나의 데이터셋 $\mathbf{x}$만 알고 있습니다. 이런 상황에서 density는 그 자체로는 아무 의미가 없어요. 모수의 값에 따라 수없이 많은 density가 가능하니까요. 즉 이처럼 데이터는 알고 모수를 모르는 상태에서는, 위 함수를 **모수에 대한 함수**로 바꿔쓸 수 있습니다. (사실 바꿔쓴다는 것도 아니고 그냥 그대로 똑같은 함수인데, 어떻게 해석하냐의 차이입니다.) 이를 데이터에 대한 모수 θ의 Likelihood라고 합니다.

$$\text{Likelihood of } \theta \quad L(\theta|\mathbf{x}) \qquad (3)$$

해석을 해보자면, Likelihood가 말하는 바는 어떤 모수의 값 θ가 우리에게 주어진 데이터 $\mathbf{x}$를 고려해보면 얼마나 "Likely"한지를 나타내는 것입니다. **이때 정말정말 헷갈리기 쉽지만 명확히 이해해야 할 것은 Likelihood는 θ의 확률이 아니라는 것입니다!!** 함수 $f(x|\theta)$는 함수의 형태가 θ에 의해 결정되는 확률밀도함수입니다. 즉 어느 θ에 대해서도 x에 대해 적분을 하면 1이 되도록 만들어진 함수라는 것이죠. 그러나 Likelihood는 이렇게 인위적으로 고안된 함수를 θ에 대해서 바라본 함수입니다. 때문에 주어진 데이터에 대해서 θ에 대해서 적분을 한다해도 1이 되리라는 보장이 없습니다. (이 부분이 통계학에 대해서 가장 많이 오해를 하기 쉬운 부분이고, 대중적으로 빈도론적 통계추론, 특히 p-value가 크게 오해를 받고 있는 부분입니다. 나중에 자세히 설명하겠습니다.)

예를 들어 이항분포의 pdf를 생각해봅시다. 동전을 3번 던져서 앞면이 나온 횟수는 $B(3, p)$를 따릅니다. 실제로 3번 던져서 1번 앞면이 나왔다고 해봅시다. 이 사건의 확률은 다음과 같이 쓸 수 있습니다.

$$\text{Probability Density of } x = 1 \text{ outcome:} \quad f(x = 1|p) = \binom{3}{1} p^1 (1-p)^2 \qquad (4)$$

그런데 모수 p를 모르면 위 함수는 뭐 써 먹을 수가 없죠. 때문에 위의 함수를 똑같이 쓰되 모수 p에 대한 함수로 생각해볼 수 있습니다.

$$\text{Likelihood of } p\text{:} \quad L(p|x = 1) = \binom{3}{1} p^1 (1-p)^2 \qquad (5)$$

똑같은 식이에요. 근데 해석만 다를 뿐이죠. $p = 1/2, 1/3, 1/4, \ldots$인 경우에 대해서 Likelihood를 계산해보세요. 그 합이 모두 1이 될까요? 아닙니다! 핵심만 정리해보면,

- Likelihood는 Probability Density와 그냥 똑같은 함수인데, 해석만 다르다.
- Probability Density는 모수 θ에 의해 결정되는 "확률"밀도 함수이다. Likehood는 데이터 $\mathbf{x}$에 의해 식이 결정되는 "그냥" 함수이다.
- Likelihood는 θ의 확률이 아니다! 즉 $L(\theta|\mathbf{x}) \neq p(\theta|\mathbf{x})$

2. 통계학의 목적: Inference와 Prediction

이처럼 데이터는 알고 모수는 모르는 상황에서, 우리는 오직 하나의 Likelihood 함수만 관측할 수 있습니다. 이 Likelihood 함수를 가지고 통계학자들이 하고 싶은 일은 두 가지로 요약할 수 있습니다.

1. **Inference:** 데이터 $\mathbf{x}$를 바탕으로 모수 θ에 대해 무엇을 말할 수 있는가?
2. **Prediction:** 데이터 $\mathbf{x}$를 바탕으로 새로운 데이터 x_{new}를 예측해보자.

동전의 예를 생각해보면, inference는 이 동전이 과연 fair한가 아닌가, 즉 앞면이 나올 확률이 무엇인가에 대해 답하고자 하는 것이며, prediction은 그렇다면 다음 시행에서 앞면이 나올지 뒷면이 나올지 예측하는 것입니다.

여기서부터는 사견인데, inference는 전통적으로 통계학의 분야였고, 앞으로도 통계학만의 분야로 계속 남을 거라고 생각합니다. inference는 결국 어떤 모수에 대한 통계적인 추론이기 때문에, 애초에 시작부터 데이터의 형성 과정에 대한 모수를 동반한 확률적인 가정이 들어가야 하니까요. 데이터의 형성 과정에 대한 여러 가정을 반영해 Likelihood 모델을 세우고, 그 가정 하에서 이런 저런 이야기를 하는게 inference입니다. 그러나 Prediction만을 한다면 굳이 데이터 형성 과정에 확률적인 가정이 들어갈 필요가 없어요. inference에서 우리가 가정을 세운다는 것은 다시말해 데이터 형성 과정을 일부로 인간이 알아보기 좋게 단순화한다는 것입니다. 그래야 어떤 모수의 값이 변하면 결과가 어떻게 변하는지 파악하기 쉬우니까요. 하지만 단순히 예측만 잘 하면 된다면 이런 가정이 필요없고, 예측 오차만을 줄일 수 있다면 수많은 다양한 시도를 해볼 수 있습니다. 그 과정에서 우리 인간이 더이상 예측 모델을 해석할 수는 없겠지요.

회귀분석을 예로 들어보면, 회귀분석은 종속변수와 설명변수들 간의 관계를 일부로 선형으로 단순화해 해석을 용이하게 만든 확률 모형입니다. 오차항의 정규 가정을 통해 종속변수의 조건부 분포를 평균이 설명변수의 선형결합으로 단순화된 정규분포로 나타낸 거에요. 그게 제일 만만하니까. 그래서 베타가 0이냐 아니냐 얘기도 하고 베타 계수가 대충 이 정도 구간 안에 있지 않을까 얘기도 하고. 그러나 실제로 데이터의 조건부분포가 완벽히 선형인 경우는 극히 드믈죠. 그런데도 쓰는 이유는 "마케팅 예산을 이정도 늘리면 매출이 얼마나 오를까"와 같은 질문에 대충 답이라도 해볼 수 있기 때문입니다. 그런데 그게 아니라 최대한 정확히 예측을 하는게 목적이면 회귀분석 말고 SVM, Boosting, Neural Network 등 다양한 알고리즘을 쓸 수 있겠지요. 그러나 "아니 우버가 왜 트럭에다가 들이박았냐" 같은 질문을 해결을 할 수 없어요.

3. 빈도통계학과 베이즈통계학: 철학의 차이

이제부터는 inference와 prediction 문제를 해결하는 통계학의 두 가지 접근법을 차례로 살펴보겠습니다. 첫 번째는 빈도통계학 접근법으로, 학부 통계학에서 가장 많이 접해본 내용입니다. 사실 그냥 통입통방 수통1 수통2가 전부다 빈도통계학을 위한 준비 + 논리 이해하기입니다. 그래서 베이즈통계 안 듣고 졸업하면 통계학을 반쪽만 알고 가는거에요. 두 번째는 베이지안 접근법인데, 두 방법의 큰 차이점은 모수에 대한 해석의 차이라고 생각해요. 빈도통계학 접근법에서 추론이란 **알지는 못하지만 단 하나의 상수로 존재하는 참 모수 θ^* 찾기**에요. 우리는 한 번 확률 실험으로 얻은 데이터를 가지고 모수를 찾아야 하는 참 안습한 상황에 처해있지요. 하지만 만일 이 확률 실험을 무수히 반복하면 얻어지는 모든 가능한 결과들의 분포가 있을 것이고, 그 분포를 결정하는 모수는 오직 한나의 참 모수값 θ^*라는 거에요. 쉽게 말하면 "앞면이 나올 확률을 알려면 무수히 많이 던져보면 된다"입니다. 때문에 빈도 통계학 접근법은 **Fixed Parameter, Random Data**라고 요약할 수 있겠습니다.

아니 뭐 당연한 거 아니냐 싶을 수도 있겠는데, 잘 생각해보면 좀 이상한 부분이 있어요. 확률실험이 무수히 반복할 수 있으면 그래 뭐 그 말대로 수많은 데이터들을 관측이야 할 수 있다면 인정하겠는데, 실제로 갖고 있는 데이터는 오직 하나밖에 없습니다. 많아야 100번 정도 던지고 앞면이 나온 횟수만 알 수 있어요. 그런데 모수를 모르는 데이터의 분포를 바탕으로 하나의 참 모수를 찾아내겠다고? 순환 논리죠 이건. 이를 빈도통계학에서는 극한분포로 우회해서 해결합니다. 표본크기 n짜리인 하나의 데이터 D에서 계산한 모수에 대한 추정량을 $\hat{\theta}(D)$라고 하면, 이 추정량의 분포를 sampling distribtion이라고 합니다. 이때 데이터의 크기가 무한으로 늘어날 때 $(n \to \infty)$ 추정량 $\hat{\theta}(D)$의 극한분포를 알 수 있습니다. 중심극한정리를 생각해보면, 모분포 $f(x|\theta)$가 제대로 생겨먹으면 (평균과 분산이 유한하면) 분포가 뭐든 간에 표본 평균의 분포가 (정확히 말하면 모평균과 표본평균의 표본수로 스케일된 편차가) 정규분포를 따르죠. 즉 **참 모수의 값 θ는 몰라도, (표본 크기가 크면) 모수의 추정량 $\hat{\theta}(D)$의 분포는 알 수 있는** 거에요. 신기하지 않나요? 이 원리를 모든 Likelihood 모델에 확장시켜 적용한 것이 MLE의 극한분포입니다.

그런데 생각을 달리 해볼 수 있어요. 모수 θ가 내가 모르는 값인데, 예컨대 θ_1일 수도 있을 것 같고 θ_2일 수도 있을 것 같아요. 그러면 그냥 "θ는 반반의 확률로 θ_1이거나 θ_2일거임" 이라고 말을 할 수 없을까요? 즉 **모수는 모르는 값이니까 모수에 대한 나의 불확실한 믿음을 확률 분포로 표현**하는 것이죠. 이렇게 모수에 대한 Belief를 확률분포 $p(\theta)$로 주고, 우리가 얻은 단 하나의 데이터 D를 바탕으로 이 믿음을 $p(\theta|D)$로 업데이트하며 θ에 대해 추론하는 방법이 베이즈 접근법입니다. 아니 모수는 모르지만 단 하나의 값으로 있는데 니 맘대로 그렇게 해도 돼? 응 그렇게 할 수 있고 그렇게 할 거야! 라고 당당히 말하고

있어요. 한마디로 말하면 **Fixed Data, Random Parameter**라는 것이죠. 이때 믿음을 업데이트하는 방법이 베이즈 정리인데, 나중에 차차 살펴볼게요.

이렇게 베이즈 접근법을 사용하면 빈도 통계학처럼 굳이 "관측은 못 하지만 기필코 어딘가 존재하는 무수히 많은 표본들"같은 공상과학적 망상이 필요가 없어요. 추론의 결과도 빈도통계학에 비해 이해하기가 훨씬 직관적입니다. 그냥 모수에 대한 확률 분포를 딱 주니까요. 다만 베이즈 통계학이 처음에 외면을 받았던 이유 중 하나이자 지금도 이를 유사과학으로 생각하는 분들이 많은 이유는 바로 $p(\theta)$, 즉 데이터를 관측하기 전 모수에 대해 가지고 있는 믿음 때문입니다. 좋게 말하면 데이터 형성 과정에 대한 관찰자의 사전 지식을 반영할 수 있는거긴 한데, 또 어찌보면 관측하는 사람에 따라서 믿음이 다 다를거고 추론 결과도 다를거에요. 베이즈 통계학의 근본적인 한계라고 볼 수도 있고, 무한한 확장성의 발판이라고도 할 수 있는 양날의 검 같습니다. 나중에 또 자세히 살펴볼게요.

4. Frequentist Approach: Fixed θ, Random X

이제 대략적인 소개는 했으니 수식을 사용해 좀 더 자세히 설명해볼게요. 일단 데이터 형성 과정에 대한 우리의 가정을 Likelihood로 표현해봅시다.

$$\begin{aligned}
\text{Data (iid):} \quad & D = [x_1, x_2, \ldots, x_n] \\
\text{Sampling Density of } D\text{:} \quad & f(D|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \quad (x_i \in \mathcal{X}, \theta \in \Omega) \\
\text{Likelihood of } \theta\text{:} \quad & L(\theta|D)
\end{aligned}$$

4-1. 빈도론적 세계관 이해하기

위의 가정은 빈도론과 베이즈 접근법에 상관없이 일반적으로 데이터의 형성과정을 확률 모델로 가정함과 동시에 성립하는 그냥 자명한 사실들입니다. 그러나 빈도론적 세계관에서는 이를 다음과 같이 다시 씁니다.

$$\begin{aligned}
\text{"Ensemble" of Data:} \quad & \{D^{(s)}\}_{s=1}^{\infty} = \{[x_1^{(s)}, x_2^{(s)}, \ldots, x_n^{(s)}]\}_{s=1}^{\infty} \\
\text{Sampling Density of } D^{(s)} \text{ (iid):} \quad & f(D^{(s)}|\theta) = \prod_{i=1}^{n} f(x_i^{(s)}|\theta) \quad (x_i^{(s)} \in \mathcal{X}, \theta \in \Omega) \\
\text{Likelihood of } \theta \text{ given } D^{(s)}\text{:} \quad & L(\theta|D^{(s)})
\end{aligned}$$

이게 무슨 말일까요?

The fundamental assumption of the frequentist approach is that the parameter is "a" fixed value, and the data is random. This assumption leads them to seek for a single value $\hat{\theta}$, so called estimator, which is a specific function of the data X that best approximate the true value θ and an error bound of that estimator given the significance level, say 95%. Interestingly, the error bound itself is also a function of the data, so we end up with an interval $\hat{\theta}(X) \pm err(X)$ that is determined by the data X.

Since the frequentists believe the true parameter is a fixed value, they also conduct a hypothesis test to accept or reject a hypothesis, which is a belief that the specific single value θ_{null} is a true parameter θ. The validation of any hypothesis relies entirely upon the p-value of $\hat{\theta}(X)$ under the the hypothesis in question. p-value quantifies how unlikely the data in hand (more precisely the value $\hat{\theta}(X)$) is under the assumption that $\theta = \theta_{null}$.

The interpretation of the error bound and the p-value is what confounds people outside the circle the most. What the frequentists mean by believing that the data is random is that there exists an infinite array of possible data, $X_1, X_2, X_3, \ldots$, if only we can observe the outcome of the sampling process infinitely many time, all at the same time. A Sci-fi imagination helps here; envisage infinitely many parallel universes in which the same statistical experiment happens

simultaneously. In our universe we observe only a single outcome X_1, but this result will vary across the different parallel universes and all those outcomes $\{X_i\}$ comprise a distribution of the data X, which is in turn dictated by the true parameter θ.

Given the imagined distribution of $\{X_i\}$ and the choice of an estimator (i.e. a choice of a function) $\hat{\theta}(X_i)$, we get the distribution of the estimator $\{\hat{\theta}(X_i)\}$, that is, many different values of the estimator $\hat{\theta}$. In each universe we calculate the interval $\hat{\theta}(X_i) \pm err(X_i)$. 95% confidence interval means that in infinitely many universes, about 95 out of 100 of them has the interval containing the true value θ. p-value of, say 3%, means that if the true distribution of estimators $\{\hat{\theta}(X_i)\}$ is indeed decided by θ_{null}, the result in our universe is so rare that only 3 out of 100 universes can observe something rarer than this.

As we can see, the core of the frequentist inference is to figure out the distribution of $\{\hat{\theta}(X_i)\}$ (**sampling distribution**) to make an inference on the distribution of $\{X\}$, which seems quite an oxymoron. Indeed, in a finite sample with $X_i = [x_1, x_2, \dots, x_N]$, finding the exact distribution of $\{\hat{\theta}(X_i)\}$ is impossible unless we make some bold assumptions about the distribution of $\{X\}$. However, if $N \to \infty$, the distribution of $\{\hat{\theta}(X_i)\}$ can be asymptotically determined. Thus the frequentist inference hinge on the asymptotic sampling distribution of $\{\hat{\theta}(X_i)\}$.

4-2. 예시: 중심극한 정리를 이용한 모평균 추정

중심극한정리를 한 번 다시 써보면 다음과 같습니다.

$$\text{Let } x \sim p(x).\ \text{If } E(x) = \theta, V(x) = \sigma^2 \text{ exist, then} \quad \frac{\bar{x} - \theta}{\sqrt{\sigma^2/n}} \sim N(0,1) \quad as \quad n \to \infty \quad (6)$$

교수님에 따라서는 중심극한정리를 이렇게 쓰시는 분도 있긴 합니다.

$$\bar{x} \sim^A N(\theta, \sigma^2/n) \tag{7}$$

즉 표본평균이 어심토틸컬하게 노말하다... 라는 건데, 사실 오해하기 쉬운 표현이긴 합니다. 중심극한정리가 표본평균이 분포수렴한다는 건 아니거든요. 표본평균은 대수의 법칙에 의해 하나의 값 모평균으로 확률수렴해요. 분포가 그냥 모평균 주위로 엄청 뾰족해지고 나머지는 다 density가 0이 되는 degenerate 분포가 됩니다. 정확히 말하면 "표본평균과 모평균의 차이에 $\sqrt{n}$을 곱해 스케일된 편차가 정규분포 $N(0, \sigma^2)$를 따르더라"입니다. 표본 수가 무한정으로 가면 표본평균의 오차는 0이 되겠지만, 거기에 표본 수를 곱한 편차는 정규분포를 따른다는 겁니다.

그래도 이런 점만 주의하면 위의 표현이 좀더 직관적입니다. 중심극한정리가 말하는거는, 원래 분포가 뭐든 간에 상관이 없이, 어떤 조건만 만족하면 (대강 말하면) 표본평균의 근사적 분포를 알 수 있다는 것입니다. 그러니까 모수를 몰라도, 모수에 대한 추정량의 극한분포는 알 수 있으며, 때문에 빈도론적 추론이 가능한 것입니다. 어떻게 하는지 이제 자세히 살펴보겠습니다.

우리의 관심사가 θ라고 할 때 이 모수의 추정량을 표본평균으로 ($\hat{\theta}(D) = \bar{x}$) 잡을 수 있겠습니다. *(사실 표본평균도 되고 표본 미디안도 되고, 그냥 뭐 아무 값이나 뽑아서 추정량으로 삼아도 됩니다. 어떤 추정량 $\hat{\theta}(D)$을 고를 것이냐의 문제는 frequentist optimality라고도 하는데, 나중에 자세히 볼게요. 결론만 말하면 그냥 MLE가 짱입니다. 표본평균도 MLE 입니다.)* 그러나 중심극한정리를 보면 관심 없는데 끼어있는 다른 모수 σ^2가 있네요. 이런 애들을 nuisance parameter라고 합니다. 그래서 이런 모수들을 제거해주는 방법을 우리가 수통1 5단원에서 배웠죠. 이 nuisance parameter을 이에 확률 수렴하는 표본통계량으로 바꿔도 분포 수렴에 문제가 없다는 정리가 바로 Slutsky's Theorem이었습니다.

$$\text{Sampling Distribtion} \quad \frac{\bar{x} - \theta}{\sqrt{s^2/n}} \sim N(0,1) \quad as \quad n \to \infty \tag{8}$$

$$(s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2)$$

자 이제 깔끔한 통계량의 극한분포가 나왔습니다. 이제부터 빈도통계학자들의 논리를 그대로 따라가보겠습니다.

1) 빈도통계학의 점 추정량과 신뢰구간

빈도통계학의 대전제는 모수 θ는 하나의 참값 θ^*으로 고정되어있다는 것이었습니다. 그러니 자연스럽게 우리의 관심사는 데이터에서 하나의 추정치 $\hat{\theta}(D)$ (pointe estimate)을 정하고, 그 추정량이 얼마나 정확한지를 신뢰구간으로 나타내는 것입니다. 모평균 추정에서 점 추정량는 표본 평균 $\hat{\theta}(D) = \bar{x}$입니다. 신뢰구간을 구하는 방법은 다음과 같습니다. 먼저 sampling distribution으로부터 다음의 "확률"구간을 얻습니다.

$$p[-1.96 \leq \frac{\bar{x} - \theta}{\sqrt{s^2/n}} \leq 1.96] = 0.95 \tag{9}$$

위의 구간을 다시 써보면 다음과 같습니다.

$$p[\bar{x} - 1.96\sqrt{s^2/n} \leq \theta \leq \bar{x} + 1.96\sqrt{s^2/n}] = 0.95 \tag{10}$$

이제 데이터로부터 $\bar{x}$를 계산하고 위의 구간에 대입합니다. 그러면 왼쪽과 오른쪽 구간의 값이 하나의 숫자로 계산이 되겠죠. 그런데 말입니다, 여기에서 확률분포를 가진 확률변수는 $\bar{x}$이며, θ는 그 값이 하나로 고정된 상수입니다. 때문에 **데이터로부터 $\bar{x}$를 계산하고 위의 구간에 대입한 순간 위의 확률구간은 더이상 "확률"구간이 아닙니다.** 가운데 상수가 있고 양 옆에도 상수가 있는 그냥 "구간"이에요. 이 구간을 빈도통계학에서는 95% 신뢰구간이라고 합니다.

$$95\% \text{ Confidence Interval:} \quad [\bar{x} - 1.96\sqrt{s^2/n}, \bar{x} + 1.96\sqrt{s^2/n}] \tag{11}$$

이게 무슨 말일까요? 빈도통계학은 항상 "관측되지 않았지만 무수히 많이 존재할 것인 가상의 데이터의 앙상블"을 전제한다고 했습니다. 신뢰도 $(1-\alpha)\%$인 신뢰구간을 간단히 쓰면 $\hat{\theta}(D_s) \pm err(D_s, \alpha)$으로 나타낼 수 있습니다. 우리가 가진 데이터는 표본 크기 n개짜리 D_1 하나이지만, 무수히 많은 데이터들 $\{D_s\}_{s=1}^{\infty}$들에서 서로 각기 다른 신뢰구간 $\hat{\theta}(D_s) \pm err(D_s, \alpha)$들을 계산할 수 있어요. 여기서 **신뢰도가 95%라는 것은 "대략 100개 중에서 95개 정도의 구간은 그 안에 참 모수 θ^*를 포함하고 있을 것"**이라는 말입니다.

2) 빈도통계학의 가설유의수준검정 (Null Hypothesis Significance Test)과 한계

가설 검정은 통입 시간떄부터 정말 주구장창 배웠죠. 뭐라뭐라 말은 많이 하는데 결국 결론은 "p-value가 0.05보다 적으면 기각해야하는 구나"라고 기억할 것입니다. 잘 이해가 안 됐을 거에요. 왜냐면 애초에 논리가 직관적이지 않아서 그래요. 지금부터 한번 빈도론자들의 가설 검정에 대한 논리를 찬찬히 따라가볼게요.

앞서 우리는 모수의 값이 몰라도 표본평균과 모수의 스케일된 편차는 극한적으로 표준정규분포를 따른다고 배웠습니다. **모수 θ에 대한 점 추정과 신뢰구간을 만드는 작업은 극한분포를 바탕으로 "모수 θ가 무엇이냐"**에 대한 물음을 해결하는 것이 목적이었습니다. 이와 달리 **빈도론적 가설검정은 그렇다면 과연 θ가 0이냐, 즉 "모수가 이 값이 맞냐"**에 답을 하는 것이 목적입니다. 빈도론자들은 그 물음에 대한 답을 데이터에서 찾습니다.

가설이라는 것은 결국 모수 공간의 분할입니다. 귀무가설이 $\mathcal{H}_0$, 대립가설이 $\mathcal{H}_1$이라고 할 때 다음과 같이 쓸 수 있어요.

$$\Omega = \mathcal{H}_0 \cup \mathcal{H}_1 \quad (\mathcal{H}_0 \cap \mathcal{H}_1 = \emptyset) \tag{12}$$

예컨대 위에서 모평균 θ를 추정하는 문제에서, 귀무가설과 대립가설을 다음과 같이 세울 수 있겠지요. (단측검정도 있지만 실제로 통계분석에서 관심이 있는 경우는 "두 변수가 독립인지", "이 변수의 회귀계수가 0인지" 등 양측검정인 경우가 많습니다. 단측검정에 대한 논의도 살짝 다르긴 한데 큰 흐름은 똑같습니다.)

$$\mathcal{H}_0 : \theta = 0$$
$$\mathcal{H}_1 : \theta \neq 0$$

이는 전체 모수공간을 하나의 점(귀무가설)과, 하나의 점을 제외한 모든 실수 구간(대립가설)로 나눈 것입니다. 때문에 $\mathcal{H}_0$을 귀무가설 영역으로, $\mathcal{H}_1$을 대립가설 영역이라고 부르겠습니다.

빈도론적 가설검정이란 결국 두 영역 중 하나를 데이터를 바탕으로 선택하는 것입니다. 영역을 선택하는 알고리즘은 요약하면 다음과 같습니다.

<빈도론적 가설 검정 알고리즘>

1. 일단 귀무가설 $\theta = 0$ 이 맞다고 해보자.
2. 귀무가설이 맞다고 치면 검정통계량 Z_{null}의 (극한) sampling distribution은 $f(Z_{null}|\theta = 0)$이 될 거다.
3. 실제 분포가 $f(Z_{null}|\theta = 0)$일때 내가 얻은 데이터에서 계산한 $z(D_s)$가 얼마나 "말이 안 되는지"를 보자.
4. 더 말이 안 되는 경우가 나올 확률이 $\alpha\%$보다 낮으면 귀무가설이 틀린거다.

앞서 든 모평균 θ에 대한 검정을 예로 들어봅시다. 먼저 귀무가설 하에서, 그러니까 진짜로 모수가 0일 때에 우리의 추정량 $\hat{\theta}(D_s) = \bar{x}$의 (극한) sampling distribution는 다음과 같을 것입니다.

$$\text{Sampling Distribtion of } \bar{x} \text{ (given any } \theta\text{)}: \quad Z = \frac{\bar{x} - \theta}{\sqrt{s^2/n}} \sim N(0,1)$$
$$\text{Sampling Distribtion if } \mathcal{H}_0 \text{ is true } (\theta = 0): \quad Z_{null} = \frac{\bar{x} - 0}{\sqrt{s^2/n}} \sim N(0,1)$$

이처럼 **귀무가설이 맞다고 칠 때의 추정량 $\hat{\theta}(D_s)$의 함수를 검정통계량** Z_{null}이라고 합니다. 여기서 실제 데이터를 넣고 ($\bar{x}, s^2$을 계산해서 넣으면 숫자가 나오겠죠) 검정통계량의 값 $z(D_s)$을 계산하고 나면 우리는 다음과 같이 p-value를 계산할 수 있어요.

$$\text{p-value}: \quad p(|Z_{null}| \geq z(D_s)) \tag{13}$$

p-value에 대한 빈도통계학자들의 설명은, "만일 귀무가설이 맞다고 가정했을때, 내가 가진 데이터보다 더 귀무가설과 반대되는 방향으로 (귀무가설과 동떨어진) 데이터가 관측될 확률"입니다. 즉 **내 귀무가설에서 정말 상상할 수도 없는 일이 일어날 확률**이라고 생각할 수 있어요. 만일 실제로 귀무가설이 맞다면, 우리의 데이터가 그렇게 희박하지 않고 충분히 있을만하겠죠? 즉 p-value가 높을 것입니다. 그러나 실제로 귀무가설이 틀렸다면, 우리의 데이터가 "아 이건 우연이라고 보긴 힘들거같은데" 정도 수준으로 희박할 것입니다. 그 우연이라고 보기 힘들다는 수준을 유의수준 $\alpha\%$으로 결정합니다. 즉 귀무가설 하에서 내가 가진 데이터보다 더 희박한 데이터가 관측될 확률이 10%면 흠 뭐 충분히 그럴만하지 하고 넘어가는데, 5%도 안 되면, 이건 귀무가설이 잘못된거다! 라고 생각하는게 합리적이라는 것입니다. 이때 유의수준 $\alpha\%$를 1종 오류의 확률이라고 얘기합니다.

$$\text{Type I Error}: \quad \alpha = p[\text{Reject } \mathcal{H}_0 | \theta^* \in \mathcal{H}_0] \tag{14}$$

$\theta = 0$인지 아닌지를 알고 싶을 때, 가설검정 알고리즘을 따라서 결정을 내리면, 원래 $\theta = 0$이 맞는데 하필 데이터가 이상하게 나와서 엉뚱하게 $\theta \neq 0$이라고 결론내릴 확률은 5%으로 제한할 수 있다는 것입니다.

(이거 말고도 검정력이라는 개념도 나오고, 기각역을 어떻게 잡아야 α를 유지하면서 검정력 β가 제일 높냐는 논의도 있습니다. 만일 귀무가설과 대립가설이 각각 하나의 값일 때 Likelihood Ratio의 기각역을 어떻게 잡아야 "Best"하냐에 대한 논의가 수통2 에서 Neymann-Pearson Theorem 이라고 해서 자세히 나옵니다. 그리고 대립가설 영역이 하나의 점이 아닐 때에도 "uniformly best"한 기각역을 어떻게 잡을 수

있는가라는 논의도 UMP라고 나오는데, 양측검정는 해당이 안 됩니다. 나중에 설명할 이유 때문에 저는 중요하게 생각 안 해서 그냥 넘어갔는데, 수통2에서 자세히 배웁니다.)

자 여기까지만 보면 그럴듯합니다. 정말 그럴듯해서 20세기 초중반 이 알고리즘이 나온 이후부터 2020년이 되는 지금까지 사람들이 계속 쓰고 있습니다. 그러나 완벽한 방법은 아닙니다. 빈도론적 가설검정을 쓰는 많은 연구에서 연구자들이 하고 싶은 것은 귀무가설의 기각입니다. 귀무가설이 대표하는 바는 "효과 없음", "의미 없음" 이런 것들인데, 연구자들이 비싼 돈 들여서 표본을 열심히 모았는데 귀무가설을 기각을 못했다는 것은 헛짓거리 했다는 것이니까요. 그런 연구자들에게 빈도론적 가설검정는 p-value 하나만 보면 "효과 있음"이라는 결론을 내릴 수 있게 해줍니다. 때문에 빈도론적 가설검정을 쓰는 연구자들은 다음과 같은 유혹과 오해에 빠지기 쉽습니다.

1. p-value가 낮을 때까지 데이터를 계속 수집하거나, p-value가 이쁘지 않은 데이터는 무시한다.
2. 하나의 데이터에 대해 p-value가 낮은 결론이 나올 때까지 이런 저런 가설 검정을 계속 한다.
3. p-value가 낮으면 낮을수록 $\theta \neq 0$의 확률이 (즉 약발이 좋을 확률이) 더 높은 것으로 착각한다.

빈도론적 가설검정 자체는 문제가 없지만, 이런 유혹과 오해 때문에 빈도론적 가설검정 자체의 신뢰성에 큰 문제가 생기고 있는 실정입니다. (사실 제가 통계학 공부를 처음 시작한 이유도 이에 대한 의구심 때문이었습니다.) 통계적으로 유의한 결과를 보고한 논문들을 다시 재현해봤더니 보고된 대로 유의한 결과가 다시 나온 논문이 거의 없다고 하네요. 자세한 사항은 Replication Crisis를 한번 검색해보면 알 수 있습니다. 빈도론적 추론의 병폐에 대해서는 이후에 자세히 알아보겠습니다.

(제 생각이지만 이런 문제가 발생한 이유는 애초에 전제를 모 아니면 도, 즉 무적권 백퍼센트 모수가 $\mathcal{H}_0$ 이거나 $\mathcal{H}_1$ 이라고 전제를 했기 때문입니다. 사실 자료를 모으는 사람이 관심이 있는 거는 "신약의 효과가 없을 때 데이터가 얼마나 레어하냐"같은 에둘른 말보다는 "그래서 신약 효과가 없을 확률이 얼마나 되냐"가 아닐까 싶습니다. 이를 확률로 쓰면 $p(\mathcal{H}_0|D)$로 쓸 수 있겠죠. 그러나 빈도론적 세계관에서는 이 확률을 구할 방법이 없습니다. 애초에 모수가 확률변수가 아니기 때문입니다. 이런 게 불가능하고 오로지 "채택", "기각" 같은 극단적인 선택만 해야하는데, 그 선택의 근거가 p-value이고, 결과에 따라 내 모가지가 왔다리갔다리하면, 당연히 p-value를 손보고 싶은 유혹이 생기지 않을까요? 이런 이유로 빈도론적 가설검정을 "trigger-happy"하다 (그러니까 깊이 고민 안 해보고 p-value 낮으면 무조건 무조건 응 기각~ 해버린다), 그리고 "p-hacking"이다라는 비판이 있습니다.)

지금까지 본 예시는 검정통계량을 중심극한정리로 얻은 경우였습니다. 모평균에 대한 추정만 예시를 들었는데, 이 외에도 모평균의 차이에 대한 검정에도 적용이 가능하고, 이항분포의 정규근사를 이용하면 모비율 차이 검정이나 분할표의 독립성 검정 문제에서도 중심극한정리로 검정통계량을 얻을 수 있습니다. 하지만 많은 경우 중심극한정리로는 검정통계량을 얻을 수 없는데, 다음에 살펴볼 MLE와 LRT는 데이터 형성 과정을 Likelihood로 세운 모든 경우에 해당하는 빈도론적 추정 방법입니다. 중심극한정리의 일반화라고 볼 수 있겠습니다.

4-3. Frequentist Optimality: 어떤 추정량을 쓸 것인가?

빈도론자들의 세계관을 다시 한번 복기해봅시다. 데이터의 sampling density를 모수 θ로 결정되는 확률분포함수로 가정하였고, θ를 모를 때 이 sampling density를 데이터에 의해 정해지는 θ의 식인 Likelihood로 해석합니다. 비록 우리가 가진 샘플은 $D^{(s)}$ 하나이지만 내가 모르는 수많은 평행우주에 똑같은 확률실험의 결과들의 앙상블인 $\{D^{(s)}\}_{s=1}^{\infty}$가 있다고 믿어봅시다.

$$
\begin{aligned}
\text{"Ensemble" of Data:} \quad & \{D^{(s)}\}_{s=1}^{\infty} = \{[x_1^{(s)}, x_2^{(s)}, \ldots, x_n^{(s)}]\}_{s=1}^{\infty} \\
\text{Sampling Density of } D^{(s)} \text{ (iid):} \quad & f(D^{(s)}|\theta) = \prod_{i=1}^{n} f(x_i^{(s)}|\theta) \quad (x_i^{(s)} \in \mathcal{X}, \theta \in \Omega) \\
\text{Likelihood of } \theta \text{ given } D^{(s)}: \quad & L(\theta|D^{(s)})
\end{aligned}
$$

우리는 데이터를 보고 모수를 추정하고자 합니다. 앞서 배운 빈도론적 방식을 다시 쓰면 다음과 같습니다.

$$
\begin{aligned}
\text{True (fixed) Parameter:} \quad & \theta^* \\
\text{Estimator of } \theta^* \text{ given } D^{(s)}: \quad & \delta(D^{(s)}) = \hat{\theta}(D^{(s)}) \\
\text{(Limiting) Sampling Distribution of } \delta(D^{(s)}): \quad & \delta(D^{(s)}) \sim p(.|\theta^*)
\end{aligned}
$$

앞서 살펴본 예시는 관심 모수가 모평균이고 추정량이 표본평균인 경우였습니다. 중심극한정리 덕분에 모수를 몰라도 추정량의 극한분포, 즉 asymptotic sampling distribution을 알 수 있었고, 덕분에 모수에 대한 점 추정치와 신뢰구간을 제시할 수 있었으며, 모수가 귀무가설의 영역에 있을 때의 추정량의 극한분포 (즉 검정통계량의 극한분포)를 사용해 나름의 가설 검정도 할 수 있었습니다.

그런데 왜 표본평균으로 했는지에 대한 설명은 따로 하지 않았습니다. 즉 $\delta(D^{(s)})$의 선택에 대한 이야기를 하지 않았습니다. 자 여러분이 $\delta(D^{(s)})$를 선택해야하는 빈도통계학자라고 생각해봅시다. 관찰은 못하는데 분명 우주 어딘가에 존재하는 데이터들이 $\{D^{(s)}\}_{s=1}^{\infty}$이 있습니다. 각각의 데이터에서 추정량 δ의 선택에 따라 하나의 모수에 대한 각기 다른 추정치를 얻겠죠. 그렇다면 나는 $\delta(D^{(s)})$를 결정할 때, 모든 데이터에 대해서 계산한 "오차"를 최대한 줄이는 방향으로 하고 싶습니다. 이때 **각각의 데이터에서 계산한 오차를 "Loss"라고 하며, 모든 데이터에 대해서 이 Loss를 계산하여 평균을 내린 값을 "Risk"라고 합니다.** Loss는 어떻게 정의하기 나름인데, 여기서는 예시로 squared loss를 쓰겠습니다.

$$
\begin{aligned}
\text{Loss of } \delta \text{ in } D^{(s)}: \quad & L[\theta^*, \delta(D^{(s)})] = (\theta^* - \delta(D^{(s)}))^2 \\
\text{Risk of } \delta \text{ over } \{D^{(s)}\}_{s=1}^{\infty}: \quad & R(\theta^*, \delta) \\
& = \mathbb{E}_{D^{(s)}|\theta^*}\{L[\theta^*, \delta(D^{(s)})]\} \\
& = \int L[\theta^*, \delta(D^{(s)})] p(D^{(s)}|\theta^*) dD^{(s)}
\end{aligned}
$$

이 Risk를 가장 최소화하는 δ가 가장 Optimal한 추정량이겠지요. 이걸 어떻게 구할 수 있을까요? 못 구해요. 일단 "무한 개의 평행우주에서의 데이터" 같은 것도 없고, 애초에 θ^*를 모르니 $p(D^{(s)}|\theta^*)$ 이것도 몰라요. 그래서 빈도통계학에서 "최적의 추정량"같은 것은 없습니다. 그렇다고 아예 아무거나 고를 수는 없지요. 그래서 빈도통계론자들은 추정량이 가지면 참 좋을 것 같은 여러 기준을 제시하는데, 무한 데이터 $\{D^{(s)}\}_{s=1}^{\infty}$에 걸친 추정량 δ의 "행태"에 대한 기준이라고 생각해보면 되겠습니다. (때문에 빈도주의보다는 행태주의(Behaviorism)가 더 어울리는 이름이라는 얘기도 있긴 합니다.)

일단 데이터의 크기가 엄청 늘어나면 추정량이 모수에 근접해야겠지요. 이를 **Consistency**라고 하는데, 가장 기본적인 성질입니다. 상식적으로 생각했을 때 확률실험을 무수히 반복했는데도 모수에 근접하지 않으면 그 추정량은 아무 소용이 없는 것입니다. 이렇게 기본적으로 일치는 해주는 예의를 갖춘 후에 고려할 사항은 **Unbiasedness**와 **Efficiency**입니다. 즉 모든 평행우주 데이터에 걸친 δ의 분포가 가급적이면 모수를 중심으로 하면 좋겠고, 분포의 폭도 좁으면 좋겠다는 것입니다.

Bias도 없으면서 모든 추정량 중에서 분산도 가장 작으면 좋은 추정량이라고 할 수도 있습니다. 하지만 어느정도 편차가 생겨도 Unbiased한 추정량보다 더 좋을 수도 있습니다. 앞서 우리는 좋은 추정량을 Risk가 작은 추정량으로 이야기했습니다. 이때 squared loss를 가정하면 Risk를 다시 써보면 다음과 같이 쓸 수 있습니다. (squared loss으로 정의된 risk를 Mean Squared Error라고 합니다.) 아마 수통1 시간때 배우셨을 겁니다.

$$
\begin{aligned}
\mathbb{E}_{D^{(s)}|\theta^*}[\theta^* - \delta(D^{(s)})]^2 &= \mathbb{E}_{D^{(s)}|\theta^*}(\delta - \mathbb{E}_{D^{(s)}|\theta^*}[\delta])^2 + (\mathbb{E}_{D^{(s)}|\theta^*}[\delta] - \theta^*)^2 \\
&= \mathbb{V}_{D^{(s)}|\theta^*}(\delta) + Bias^2(\delta) \\
\therefore MSE &= Variance + Bias^2
\end{aligned}
$$

위 식을 보면 추정량을 정할 때 만일 Biased하더라고, 그 편차가 크지 않으면서 줄어든다면 오히려 Unbiased Estimator보다 MSE가 더 적을 수 있음을 알 수 있으며, 나중에 그런 예를 한번 살펴보겠습니다.

(지금까지 말한 논리는 Supervised Learning에도 그대로 적용됩니다. 어떤 연속형 확률변수 t의 값을 예측하는 설명변수들의 함수 $\hat{f}(\mathbf{x})$, 즉 예측 모델을 만드는 문제를 생각해봅시다. 이때 실제 함수는 f이겠지만 이를 알 수 없으니 회귀분석 등 갖가지 방법을 이용해 $\hat{f}$를 추정할 수 있습니다. 이 때 f와 $\hat{f}$의 관계도 위의 논의와 똑같습니다.)

4-4. Maximum Likelihood Theory

지금까지의 논의를 종합해보면 다음과 같습니다.

1. 빈도통계학 추론은 평행우주 데이터 $\{D^{(s)}\}_{s=1}^{\infty}$에서의 Sampling Distrubtion $\delta(D^{(s)}) \sim p(.|\theta^*)$에 달렸다.
2. 모수 θ에 대한 추정량 $\hat{\theta} = \delta(D^{(s)})$의 결정은 다음의 사항을 고려해야 한다.
    1. 일단 δ의 sampling distribtion을 근사적으로나마 알아야 한다.
    2. 가급적이면 δ의 평행우주 데이터 $\{D^{(s)}\}_{s=1}^{\infty}$에서의 행태가 "이쁘면" 좋겠다. (Consistent, Unbiased, Efficient)
3. Sampling distribution $\delta(D^{(s)}) \sim p(.|\theta^*)$만 알면 점 추정, 구간 추정, 가설 검정 다 할 수 있다!

빈도통계학 추론에서 제일 "빡센" 부분은 바로 2-1. 입니다. 데이터에 대한 Likelihood 모형을 세우고 이를 바탕으로 모수에 대한 추정량 δ을 결정했는데, 여기서 δ의 극한 분포를 유도하는 과정이 여간 힘든게 아니라고 합니다. (해석학이 필수인 이유 중 하나입니다.) 이를 구하는 방법은 크게 다음과 같습니다.

1. **Plug-in Principle**: 우리가 매일 처음 봤던 예시가 중심극한정리와 슬러츠키 정리인데, 이처럼 추정량 안에 있는 nuisance parameter를 표본에서 계산한 값으로 때려넣는 것을 plug-in한다고 합니다.
2. **Taylor-series Approximation**: 수통1에서 배운 delta-method입니다. 어떤 모수 θ의 추정량 $\hat{\delta}$의 극한분포를 이미 알고 있을 때, 그 모수의 함수 $g(\theta)$의 극한분포를 구하는 방법입니다. 함수 g를 $\hat{\theta}$ 근처에서 선형근사 (테일러 1계 근사)하여 다음과 같이 나타냅니다.
$$g(\theta) \approx g(\hat{\theta}) + g'(\hat{\theta})(\theta - \hat{\theta}) \tag{15}$$
그러면 (슬러츠키 정리도 적용하고 난 후) 다음과 같이 쓸 수 있습니다.
$$\text{If } \sqrt{n}(\hat{\theta} - \theta) \to^D N(0, \sigma^2), \text{ then } \sqrt{n}(g(\hat{\theta}) - g(\theta)) \to^D N(0, \sigma^2\{g'(\hat{\theta}\}^2) \tag{16}$$
제일 많이 쓰이는 방법입니다. 범주형자료분석에서 배우실 로그오드의 극한분포도 이렇게 구합니다. 그도 그럴 것이 솔직히 이거 말고는 방법이 없어요. 만일 데이터 크기가 크면 $\hat{\theta}$가 얼추 θ 근처에서 얼쩡거릴 거니 (consistent하니까) 저런 선형 근사도 봐줄 만한 정도의 근사라고 생각합니다.
3. **Maximum Likelihood Theory**: 이제 얘기할 내용입니다. 결론부터 스포하자면 모수가 있는 분포함수족에 대해서 Likelihood를 최대화하는 모수를 $\hat{\theta}_{mle}$라고 할건데, 다음과 같은 극한분포를 가집니다.
$$\hat{\theta}_{mle} \sim^A N(\theta^*, \frac{1}{nI_\theta}) \tag{17}$$
신기하지 않나요? 나중에 자세히 살펴보겠습니다.
4. **(Non-parameteric) Bootstrap**: 컴퓨터 연산력도 썩어나겠다 내가 그냥 $\{D^{(s)}\}_{s=1}^{\infty}$를 손수 만들어서 추정량 δ의 분포를 직접 보겠다는 방법입니다. 쉽게 말하면 데이터의 크기가 충분히 크면 데이터의 경험분포 (그러니까 히스토그램)이 실제 분포 $f(D|\theta^*)$를 솔차니 잘 근사할 것이니, 데이터에서 표본 크기만큼 다시 **복원추출**하여 S개 만큼의 인위적인 데이터의 앙상블 $\{D^{(s)}\}_{s=1}^{S}$을 만들겠다는 이야기입니다. 극한분포를 진짜 죽어도 못 구해서 어쩔 수 없을 때 울며 겨자먹기로 논문에 쓰는 방법이라고 들었습니다.

그럼 지금부터 3번 Maximum Likelihood Theory에 대해 알아보겠습니다.

1) Maximum Likelihood Theory 수식으로 보이기

자 먼저 빈도론자들 세계관부터 다시 써봅시다.

$$\text{"Ensemble" of Data:} \quad \{D^{(s)}\}_{s=1}^{\infty} = \{[x_1^{(s)}, x_2^{(s)}, \dots, x_n^{(s)}]\}_{s=1}^{\infty}$$
$$\text{Sampling Density of } D^{(s)} \text{ (iid):} \quad f(D^{(s)}|\theta) = \prod_{i=1}^{n} f(x_i^{(s)}|\theta) \quad (x_i^{(s)} \in \mathcal{X}, \theta \in \Omega)$$
$$\text{Likelihood of } \theta \text{ given } D^{(s)}: \quad L(\theta|D^{(s)})$$

여기서 모수의 추정량을 결정할 때 다음과 같은 추정량을 생각할 수 있습니다.

$$\text{True (fixed) Parameter:} \quad \theta^* \tag{18}$$
$$\text{Estimator of } \theta^* \text{ given } D^{(s)}: \quad \delta(D^{(s)}) = \arg\max_{\theta} L(\theta|D^{(s)}) \tag{19}$$

즉 데이터로 결정되는 θ의 함수인 Likelihood에서, 함수의 값 $L(\theta|D^{(s)})$이 가장 큰 지점을 δ로 정하는 것입니다. 이 추정량을 $\hat{\theta}_{mle}$이라고 합니다. 뭔가 직관적으로 크게 이상하지는 않아요. 비록 $p(D^{(s)}|\theta)$는 모르지만, 적어도 우리가 가진 데이터에서는 $\theta = \hat{\theta}_{mle}$일 때 $p(D^{(s)}|\theta)$의 값이 가장 크니까, 데이터에 가장 잘 들어맞는 값이라고 생각할 수 있으니까요. Maximum Likelihood Theory는 현대통계학의 천재 (그러나 본인이 애연가시라 흡연과 폐암의 관계는 한사코 부정하신(사실 Randomized된 실험 결과가 없으니까)) 로날드 피셔께서 그냥 혼자서 뚝딱 다 만들었는데, 피셔 이전에도 대충 이런 생각으로 MLE를 쓰자고 제안한 사람은 많았지만, MLE의 극한분포를 제시한 분이 바로 피셔입니다.

사실 MLE는 Likelihood를 최대화하는 값이므로 단조변환인 로그변환을 하면 log-likelihood를 써도 됩니다. 그렇게 되면 최대화 문제가 훨씬 더 쉬워집니다.

$$\hat{\theta}_{mle} = \arg\max_{\theta} L(\theta|D^{(s)}) = \arg\max_{\theta} l(\theta|D^{(s)}) = \arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i^{(s)}|\theta) \tag{20}$$

이제 MLE의 행태를 살펴보면서 MLE 사용에 대한 정당화 근거를 간단히 알아보겠습니다. 그 전에 몇 가지 개념을 짚고 넘어갈게요.

- **Log Likelihood:** 말 그대로 Likelihood에 로그를 취한 함수입니다. (로그 우도)

$$l(\theta|x_i) = \log f(x_i|\theta) \tag{21}$$

- **Score Function:** Log likelihood을 모수 θ에 대해 미분한 함수입니다. 모수가 벡터일 경우 로그 우도의 gradient로 정의합니다. score function이 말하는 바는 "로그 우도의 기울기"입니다.

$$\begin{aligned} s(\theta|x_i) &= \frac{d}{d\theta} l(\theta|x_i) \quad (\theta \in \mathbb{R}) \\ &= \nabla_\theta \, l(\theta|x_i) = [\frac{\partial \, l(\theta|x_i)}{\partial \theta_1}, \frac{\partial \, l(\theta|x_i)}{\partial \, \theta_2}, \dots, \frac{\partial \, l(\theta|x_i)}{\partial \theta_d}] \quad (\theta \in \mathbb{R}^d) \end{aligned}$$

  n개의 iid 데이터로 이뤄진 $D^{(s)}$의 로그 우도는 개별 데이터의 로그 우도의 합으로 이뤄짐을 보았습니다. 때문에 $D^{(s)}$의 score function은 다음과 같습니다.

$$s(\theta|D^{(s)}) = ns(\theta|x_i) \tag{22}$$

- **Fisher Information**: Score function도 또한 데이터의 함수 $s(\theta|x_i)$입니다. 때문에 모수값이 어떤 θ로 주어지면 $x_i \sim f(x_i|\theta)$의 분포를 따르며, 데이터의 통계량인 $s(\theta|x_i)$의 평균과 분산을 계산할 수 있습니다. 이때 score function의 평균은 0이며, 분산을 Fisher Information $I(\theta)$이라고 합니다. 데이터가 n개인 경우도 위와 마찬가지로 쓸 수 있습니다.

$$\begin{aligned} &\mathbb{E}_{x_i|\theta}[s(\theta|x_i)] = \mathbb{E}_{x_i|\theta}[\frac{d \, l(\theta|x_i)}{d\theta}] = \int \frac{f'}{f} f dx = \frac{d}{d\theta} \int f dx = 0 \quad (\theta \in \mathbb{R}^1) \\ &\mathbb{E}_{x_i|\theta}[s(\theta|x_i)] = \mathbf{0} \in \mathbb{R}^d \quad (\theta \in \mathbb{R}^d) \end{aligned}$$

  Fisher Information은 다음과 같이 score function으로 나타낼 수 있음을 어렵지 보일 수 있습니다. (이런 걸 다 수통 2때 합니다)

$$\begin{aligned} I(\theta) &= \mathbb{V}_{x_i|\theta}[s(\theta|x_i)] = -\mathbb{E}_{x_i|\theta}[s(\theta|x_i)^2] = -\mathbb{E}_{x_i|\theta}[s'(\theta|x_i)] \\ I_n(\theta) &= \mathbb{V}_{D^{(s)}|\theta}[s(\theta|D^{(s)})] = n\mathbb{V}_{x_i|\theta}[s(\theta|x_i)] = nI(\theta) \end{aligned}$$

Fisher Information이 말하는 바는 "로그 우도의 peak한 정도"입니다. 로그 우도 함수가 뾰족하다는 것은 그만큼 데이터가 θ_{mle}에 많이 쏠려있다는 것이고, 여기에서 즉 모수의 추정량으로서 MLE의 분산이 작을 것을 기대할 수 있습니다.

(모수가 벡터인 경우 Fisher Information Matrix을 다음과 같이 정의할 수 있으며, 아래의 등호가 성립함을 어렵지 않게 보일 수 있습니다.)

$$\mathbf{I}(\theta)_{k,j} = \mathbb{C}\text{ov}_{x_i|\theta}(\frac{\partial\, l(\theta|x_i)}{\partial\theta_k}, \frac{\partial\, l(\theta|x_i)}{\partial\theta_j}) = \mathbb{E}_{x_i|\theta}[-\frac{\partial^2 l(\theta|x_i)}{\partial\theta_k\partial\theta_j}]$$

이제 단일 모수인 경우에 대해 간단하게 MLE의 극한분포를 증명해보겠습니다. 우선 score function에 대하여 다음과 같이 테일러 1계 근사 식을 쓸 수 있습니다.

$$0 = s(\hat{\theta}_{mle}|x_i) \approx s(\theta^*|x_i) + s'(\theta^*|x_i)(\hat{\theta}_{mle} - \theta^*) \tag{23}$$

score function은 정의상 MLE에서 0입니다. 이때 score function을 MLE에 대한 함수로 보고, MLE 근처에 있을 것만 같은 참 모수값 θ^*에서 이 함수 $s(\hat{\theta}_{mle}|x_i)$를 선형근사해본 것입니다. 그러면 위 식을 다시 아래처럼 쓸 수 있습니다.

$$\hat{\theta}_{mle} \approx \theta^* + \frac{s(\theta^*|x_i)/n}{-s'(\theta^*|x_i)/n} \tag{24}$$

- $s(\theta^*|x_i)/n$ 이 식은 score function의 평균입니다. score function의 평균은 0이며, 분산은 $I(\theta^*)$입니다. 때문에 중심극한정리에 의해 $s(\theta^*|x_i)/n \sim^A N(0, I(\theta^*)/n)$으로 쓸 수 있습니다.
- 위에서 이미 $I(\theta) = -\mathbb{E}_{x_i|\theta}[s'(\theta|x_i)]$임을 보였습니다. 때문에 대수의 법칙에 의해 $-s'(\theta^*|x_i)/n$은 $I(\theta^*)$으로 확률 수렴합니다.

이걸 다 종합해보면, 그리고 슬러츠키 정리를 사용해 θ^*을 $\hat{\theta^*}_{mle}$으로 plug-in 하면, 우리는 다음과 같은 MLE의 극한 분포를 얻을 수 있습니다.

$$\text{Single parameter} \quad \hat{\theta}_{mle} \sim^A N(\theta^*, \frac{1}{nI_{\hat{\theta}_{mle}}}) \tag{25}$$

$$\text{Multi-parameter} \quad \hat{\theta}_{mle} \sim^A N(\theta^*, \frac{1}{n}\mathbf{I}^{-1}(\hat{\theta}_{mle}))$$

즉 모든 Likelihood 모델에 대하여, 여기에서 말하지는 않았지만 이 함수들이 어떤 정규성 가정들을 만족한다면 (지수분포족이면 충분히 만족하는), 우리는 Likelihood를 최대로 하는 추정량의 극한분포가 정규분포라는 것을 알 수 있습니다. 이 놀라운 결과를 바탕으로 아까 봤던 빈도통계학의 추론을 똑같이 다 할 수 있는 것입니다. (Likelihood를 이용한 검정 방법은 따로 소개하지 않겠습니다. https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-are-the-likelihood-ratio-wald-and-lagrange-multiplier-score-tests-different-andor-similar/)

MLE는 점근적으로 Minimum Variance인 Unbiased Estimator입니다. MLE 극한분포의 분산은 Rao-Cramer lower bound에 의하면 모든 불편추정량이 가질 수 있는 분산의 하한입니다. 또한 MLE는 항상 평균이 모수와 일치하지는 않지만, 그 편차가 점근적으로 0이 되므로 점근적으로 Unbiased라고 볼 수 있습니다. 이처럼 MLE는 굉장히 훌륭한 빈도통계 추정량 성질을 가지고 있으며, Likelihood만 있으면 Test를 위한 검정통계량의 극한분포도 알려져 있기에, 사실상 거의 모든 빈도통계 추론은 MLE와 LRT로 이뤄진다고 봐도 과언이 아니겠습니다.

4-5. 빈도론적 추론의 병폐

자 이제 빈도통계론의 논리도 이해했고 그 끝판왕인 MLE와 LRT도 봤습니다. 지금부터는 빈도론적 통계추론, 그 중에서도 검정 (NHST)이 가진 "병폐"들에 대해서 살펴보겠습니다. 앞서 잠깐 봤는데, 여기서는 좀 더 자세하게 다뤄보겠습니다.

1) Trigger Happy: $p(D|H_0)$만 보고 H_0을 기각함

$p(D|H_0)$이 굉장히 작아야지만 귀무가설을 기각하는 것이 얼핏 보면 굉장히 보수적으로 보이지만, 사실 이런 식으로 세팅을 해놓으면 "귀무가설에 반대되는 evidence"만 반영하게 되지, 귀무가설에 좋은 evidence는 절대 반영을 못함. 실제로 사람들이 관심 있는 것은 $p(H_0|D)$에 더 가까움. 그러나 NHST는 이런 거를 반영을 못 함. 이에 대하여 Sellke et al.(2001) 논문은 실제로 p-value가 0.05 미만이지만 $p(H_0|D)$는 무려 0.3이 넘는 예시도 제시함. 때문에 귀무가설이 아니다, 즉 어떤 효과가 있다고 결론을 내리기에는 p-value가 굉장히 부적합해서, 어떤 의학 저널에는 p-value만 쓰는거를 금지하기도 했음.

2) Stopping Rule: 같은 데이터라도 수집 환경에 따라 결론이 다름

동전을 던져 앞면이 나온 개수를 세는 실험을 생각해보자.

- 실험 1. 딱 6번 던져서 앞면이 나오는 개수를 센다.
- 실험 2. 뒷면이 나올 때까지 계속 던지고 난 후 앞면을 셈.

두 실험에 의해 똑같은 데이터 HHHHHT가 나왔다고 하자. 상식적으로 p-value는 같아야할 거 같은데, 다르다. 왜 그럴까? 실험 1에서의 귀무가설은 $p = 1/2$ 하에서 이항분포이지만, 실험 2에서 귀무가설은 $p = 1/2$하에서의 음이항분포이기 때문이다. 때문에 Likelihood가 다르므로 신뢰구간도 다르고, p-value도 다르게 나온다.

이때 실험 1과 2의 차이는 Stopping Rule이다. 실험 1은 미리 수집할 데이터 개수를 정한 경우고, 실험 2는 특정한 기준이 만족 될때까지 (대개 p-value가 이쁘게 나올 때까지) 데이터를 수집하는 경우다. 이러면 실험 1과 실험 2에서 관측된 결과는 같아도 p-value는 다르게 나온다.

때문에 빈도론적 추론을 위해서는 데이터 자체 뿐만 아니라 데이터의 수집 환경까지 고려해야 한다. 이러한 아이러니함을 꼬집은게 바로 voltmeter story인데, 결론은 NHST는 Likelihood Principle을 만족하지 않는다는 것. 이 부분은 나도 아직 자세히 이해는 못했지만 충분히 생각해볼 만한 지점인듯. https://psychology.wikia.org/wiki/Likelihood_principle

3) 가설 검정을 많이 하다보면 몇 개가 얻어 걸리게 되어있음

유의수준 5%라는게 한번 했을 때 귀무가설을 제대로 기각할 확률이 95%라는거지, 가설검정을 계속 하다보면 귀무가설을 제대로 기각할 확률도 계속 내려가서, 나중에는 거의 무조건 겉으로는 p-value가 유의미하게 나오는 경우가 생김. 이런 문제를 인지하고 가설검정을 에컨대 수 천개 할 때 가설검정 방법으로 제시된게 false-discovery rate인데, 시간이 없어서 아직 공부를 못함

4-6. (참고) 수통 2가 어려울 때 읽어보는 고전빈도통계학의 역사

학교 통계학과 커뮤니티에 게재한 글인데, 본 문서와 같이 읽으면 좋을 것 같아 첨부합니다

사실 데이터 분석하려고 통계학을 공부하는 입장에서는 베이지안이나 빈도론 접근을 둘 다 아는게 좋죠! 둘 다 똑같은 가정에서 출발해 가정만 다른건데, 그 차이는 결국 철학과 세계관의 차이로 귀결되고 그 차이를 아는게 적어도 데이터에 대학 확률적 가정을 동반한 데이터 분석에는 큰 도움이 되는 것 같아요. 그리고 대부분의 통계학 교수님들도 뭐 어느쪽으로 기울인 정도는 있겠지만 두 방법 다 알고 계시고 상황에 따라 그때그때 데이터에 더 맞는 접근법을 쓰는 것 같아요. 저도 겉으로는 베이즈 베이즈 하지만 스스로 연구할 준비가 되기 전까지는 앞서간 분들이 닦아놓은 이론들은 가리지 않고 공부하는게 맞다고 생각합니다.

아무튼!

그런 점에서 우리가 학교에서 배우는 고전 빈도론 통계학의 맥락을 아는게 좋을 것 같아, 괜찮은 책을 소개하고자 합니다. 이것도 학회 친구들에게 쓴 글인데 좀 다듬고 내용 추가해서 올립니다. 이걸 읽고 나서 수통 2(a.k.a 빈도통계학)를 들으면 더 재밌을 거에요!

천재들의 주사위 (The lady tasting tea, David Salsburg, 2001)

본인이 처음 통방듣고 수통1 들을때 도대체가 수업 내용들이 뭔 맥락이란게 없어서 그 맥락을 찾고 싶어서 책을 뒤지다가 찾은 책인데, 20세기 빈도통계학의 발전과정을 위인전처럼 인물 중심의 서술로 재밌게 엮어낸 책이라, 우리가 배우는 내용에 어떤 역사적인 맥락이 있었는지를 수식같은거 하나도 없이 이야기처럼 재밌게 읽을 수 있는 책이라고 느꼈음. 물론 지금은 읽은지 좀 꽤 돼서 다 까먹었긴 한데, 약간 경제학계에서 "세속의 철학자들(by 로버트 하일브로너)"라는 책이랑 굉장히 비슷하다고 느꼈음. 칼 피어슨 - 에곤 피어슨 부자와 로날드 피셔의 대를 이으는 자강두천이 하이라이트인거같음. 그리고 피셔께서 흡연과 폐암에 대한 논문들을 "Randomize 안 했자나!"라며 모조리 까버리며 평생 애연하신게 정말 인상깊었음.

원래 칼 피어슨이 바이오메트리카 저널 만들면서 먼저 통계학 일짱 먹었는데 (이분은 데이터를 노가다로 계속 모으면서 1,2,3,4차 모멘트를 계산해나가면 결국 당신께서 pearson family라고 부르는 분포함수들 중 하나의 분포를 따를거라고 주장하신거로 본인은 이해함. 그 피아슨패밀리 분포로 퉁치는거를 정당화하는게 카이제곱 피어슨 검정인데, 피어슨 패밀리가 묻히고 피어슨도 학계에서 아싸가 된 후에도 살아남아 당당히 통입책에 등장하는 불멸의 빈도론 검정) 칼 피어슨의 방법론에 정면으로 반기를 든게 피셔라서 처음에 엄청 찍혀서 피셔가 고생했지만 결국 피셔의 너무 우아하고 고귀하고 깔쌈한 방법론에 결국 칼 피어슨이 퇴물됨.

피셔는 정말 어렸을 때부터 천재였는데 통계학 박사까지 따고도 하필 그때 통계학의 대가인 피어슨 면전에다가 "개소리ㄴㄴ"하니 어디 머리에 피도 안 마른게 덤빈다면서 학계에서 자리가 없어짐. 진짜 공부 많이 하신 어른들이 똥고집은 대단한 듯. 하필 박사 따고 나서 대공황마저 겹쳐서 실업자에 방황하다가 저어어기 한적한 시골동네 농장에서 통계분석하는 자리를 겨우겨우 얻었는데, 거기서 개판으로 모인 데이터와 주먹구구식 일처리를 보고 경악하며 데이터를 모으는 방식과 통계적 추론을 하는 방식을 본인이 처음부터 끝까지 만들었는데 그게 우리가 배우는 통입 통방 실계라고 봐도 과언이 아님. ㄹㅇ 입지전적인 인물...

그 외 피셔의 업적을 잘 모르는 본인이 이해한대로 정리해보자면 크게

1. CLT를 확장해 MLE의 극한분포를 증명함. 물론 그전에도 막연히 최대값 쓰면 되지 않을까~ 싶었지만 그 통계량의 극한에사의 노말리티를 보인 것은 정말 기념비적인 업적. 지금까지도 빈도통계론자들의 No.1 통계량임 그냥 무적권 MLE써야함
2. 귀무가설과 p-value를 처음으로 도입하였으며, 극한분포를 통해 처음으로 "통계적 거리"를 나타낼 수 있는 거리함수를 제공함. 통계학자들이 해결해야할 문제들은 대개 데이터만 보면 애매한 것들임. 예컨대 모평균비교를 보면, 박스플럿 그려버면 저놈이 이놈보다 높은거 같은데 애매하네~ 이랬을 때 그 차이가 유의미한지 아닌지, 큰지 작은지를 말해야하는데, 이전에도 그런 시도는 있었지만 처음으로 체계적으로 통계적 거리를 제시한 사람이 피셔라거 생각함. 정확히 말하자면 차이가 없다는 "귀무가살" 하에서 데이터가 나타내는 이 차이가 얼마나 먼지를 말하는게 바로 귀무가설 하에서의 통계량의 분포이며, 데이터의 값을 넣었을 때 이 귀무가설이 얼마나 말이 안 되는지를 말하는게 p value임.
3. 사견: 피셔의 p-value는 데이터만 보고 모수에 대해 말하겠다는 세계관에서 통계학자가 낼 수 있는 최선임. 그말은 즉 p-value보다 더 한 얘기는 함부로 할 수 없다는 거임. 피셔는 살아생전에 피밸루가 0.05보다 낮으면 귀무가설이 틀린거고 0.050001이면 귀무가설이 맞는거야~ 라는 정신나간 소리는 한번도 할 수 없으며, 오히려 그런 해석을 정말 극도로 경계하거 혐오하며 극렬히 깠음. 피셔에게 피밸류는 그냥 말그대로 "evidence against the null hypo"일 뿐임. 피밸류가 높은데 어떡할까요? 라고 물으면 피셔는 데이터 더 모아 ㅂㅅ아 이러고 말았을 인물임. 제한된 데이터를 가지고 "기다 아니다" 판단하는 것에 대해 극도로 경계하고 주의를 준게 피셔임. 피셔에게 p-value는 오로지 "inferential"할 뿐이지, 절대로 어떤 행동방침, 결정알고리즘 이런거를 제공하는 의미가 아님.

그러나 칼 피어슨의 아들인 에곤 피어슨이 네이만과 함께 또 피어슨의 방법론에 반기를 들면서 (정작 아버지는 퇴물이라고 거리둠.. 나중에 피어슨께서는 연구소는 정말 아무도 안 오고 대학교 내에서도 혼자 밥 먹고, 말년이 정말 외로웠을 텐데 아들은 아버지 자존심을 아니까 뭐라 하지는 못하고 피한 것 같음.. 아들아 아빠 연구소 와서 일 도와야지? 앗 아버지 제가 친구랑 논문쓰느라 바빠서^^; 나중에 연락 드릴

게요!) 이어지는 드라마를 알고나면 수리통계학 2 공부가 더 재밌을거임. 베이즈 서술은 끝에 약간 반란이라는 식으로 잠깐 서술되고 맘

첨언:
지금 우리가 통입 통방에서 배우고 사람들이 많이 쓰는 귀무가설유의성검정 (NHST)의 토대를 제공한게 네이만 -(아들) 피어슨 이 두 분임. 이분들은 피셔가 다 만들어놓은 귀무가설 p발루 프레임에다가 어떤 결정알고리즘을 제시하신 분임. 피셔가 귀무가설만 말했다면 네이만 피어슨은 대립가설이란거를 명시적으로 끌고온 분들이며, 진리는 귀무가설 혹은 대립가설에만 있다! 라고 실질적으로 주장하신 거임. 그래서 피셔 철학을 그냥 따로 Fisherian이라고 구분하기도 하는듯.

쉽게 말하면 피발류가 낮을 때 피셔는
"음 귀무가설의 증거가 낮군." 이러고 말았다면 네이만 피어슨은
"그러니까 귀무가설이 틀.린.거.이고 대립가설이 맞다!" 라고 말할 수 있다는 것을 주장하신 거라고 생각함.

이렇게 대립가설이란 거를 끌고오니까 이분들은 검정의 power라는거를 제시함. 데이터의 진리는 귀무가설 혹은 대립가설이니, 우리가 진리를 밝히는 검정을 설계할때는 (데이터를 수집해 검정통계량을 만들때에는) 먼저 "귀무가설이 맞는데 에쿠 실수로 기각해버리는 알파" 확률을 일단 고정하고("유의"수준) 그러고 나서 "대립가설이 맞을 때 제대로 결정할 확률"인 검정력, 검정의 파워! 를 최대화해야 한다고 주장하심. 피셔의 검정통계량이 뭐 MLE를 쓰긴 하지만 그건 추정량으로서 착한 통계량이지 검정에 있어서 최선인지 아닌지는 모르고 임의적이었다면, 당신들은 최고의 검정방법과 최고의 검정통계량을 만드는 알고리즘을 제시했다고 의미를 두신 분들이고 수통2 7장부터 우리를 괴롭히시는 분들이지만... (LRT를 만드신 분들)

문제는 저 알고리즘이 참 정말 제한적인 상황에서만 "uniformly most powerful"하지 대부분의 경우에서는 못 쓴다는거. 정확히 말하면 likelihood가 모노톤하고 (이건 뭐 지수분포족이면 ㅇㅋ한다 쳐도) 단측검정에만 쓸 수 있으니.. 배우긴 배우는데 약간 현실성은 없고 그냥 빈도론자들의 논리는 (이상적인 상황에서는) 고결하다!는 걸 보여주려는 것 같음.

네이만 피어슨 방식의 이분법은 진짜 진리가 기다 아니다 (방구 꼈냐 안 꼈냐) 이런 상황에서는 정말 좋은데, 모수 공간이 연속적인 경우에서는 대립가설이라는 거 자체가 존나 우스워짐. 무한대와 무한대 사이에서 0이라는 존나 미미한 점 빼고 모두! 라는 가설이 뭔 의미가 있을까?

또 본인이 생각하기에 네이만 피어슨 논리의 "신성모독"은 감히 진리공간을 멋대로 잘라내서 그 안에 정답이 있다고 단언하는 것임. 이게 뭐냐면, 실제 진리공간이 끝없이 넓다면 통계적 추론은 여기서 모수로 결정되는 함수공간을 인위로 그려서 그 안에서만 생각하는 거고, 그게 모수적 가정임. 피셔는 "적어도 모수공간의 한 점은 진리일 증거는 희박하다"라고만 말햇는데, 네이만피어슨 논리는 "이 점이 아니면 모수공간의 다른 모든 부분 안에 진리가 있다"고 말한것으로, 이는 정말 대담하고 건방진 결론이라고 생각함 본인은.

네이만은 피셔의 이론을 보고 "worse than useless"라고 깠는데, 거기다가 피셔는 이렇게 답했죠.
피셔를 인용하자면,
"네이만 피어슨의 "가설검정이론"은 이를 따르는 사람들을 잘못된 길로 이끌어 헛된 노력 끝에 실망을 안겨줄 진데, 정작 그 저자들은 이런 위험을 알려주려고 하지 않는다는 점은 실로 두려운 점이다.
It is to be feared, therefore, that the principles of Neyman and Pearson's "Theory of Testing Hypothesis" are liable to mislead those who follow them into much wasted effort and disappointment, and that its authors are not inclined to warn students of these dangers.(Fisher, 1956)"

물론 네이만 피어슨 당신들께서는 그렇게까지 생각을 안 하셨을 거고, 그리고 빈도통계학 전공한 통계학자들도 그런 극단적인 사고방식에 동조하지 않을 거임. 그러나 그게 중요한 게 아님. 가장 큰 문제는 적어도 그 논리를 보고 실제 데이터를 다루는 practitioner들이 만들고 교육한 NHST는 암묵적으로 그런 사고를 조장했던게 지금까지의 현실임. 지금 우리가 배우는 NHST는 피셔의 피밸류도 쓰고 네이만피어슨의 검정논리도 쓰는데, 피셔와 네이만 피어슨의 진리공간에 대한 철학의 차이에 대한 사고와 설명 없이 정말 기계적으로 가르쳐지며 숱한 오해를 양산하고 있는게 지금의 현실임.

본인은 통계학의 사명은 사람들에게 올바른 귀납적 사고 방식의 틀을 제공해주는 것이라고 생각함. 이 점에서 피 밸루에 대한 통계학자들과 통계학을 쓰는 연구자들 간의 크나큰 인식의 괴리와, 그로 인해 생기는 오용과 오해와 그로 인한 일반 대중들의 피해는 빈도통계학자들이 해결해야하는 가장 큰 숙제라고 생각함. 때문에 통계학 써클 이외 대중들에게의 통계학 교육에 실패했다는 것이 옛날부터 나오는 통계학자들의 한탄이었음

그러니까 빈도통계론자들은 일반 대중에게 NHST라는 "쓸 만한" 기계를 줬는데, 사용설명서를 아무도 읽어보지 못 하게 써서, 기계를 잘못 써서 손가락 다치거나 중독되는 사람들이 생기는 사태. 이런 사태에서 회사는 당연히 리콜을 해야하는데, 딱히 대체품이 없어서 "그...그렇게 쓰면 안돼여!" 라고 말하는 것 외에는 뭐 대안이 없는 것 같음

여기 나오는 내용에 대해 더 공부하고 싶으면

1. Probability Theory and Statistical Inference: Econometric Modeling with Observational Data (Spanos, 1999) 이 책의 14장 가설검정 부분
   Lady tasting tea 이 책에서 나온 피셔-네이만 논쟁을 내용을 수식을 사용해 설명한다고 보면 됨. 좀 어려워서 수통 1 2 다 듣고 읽어야 읽힐 것임.
2. Fisher, Neyman, and the Creation of Classical Statistics (Lehmann, 2011)
   최근에 알게 된 책인데, 저자가 네이만의 제자였다고 합니다. 위의 14장의 내용을 100장으로 풀어낸 책인거같은데 목차만 보고 아직 읽어보지 않음. 시간이 된다면 정말 읽어보고싶은데... 근데 또 어차피 우리가 수통 2에서 배우는 고전통계학은 사실 모수가 무진장 많은 현대에서는 못써서 읽을 시간이 있을지 모르겠다.
3. Machine Learning: a Probabilistic Perspective (Murphy, 2012) 이 책의 6.6절 pathologies of frequentist statistics 이 부분에서 p-value의 오용 등 빈도적 가설검정의 단점을 신나게 까는데, 머신러닝 전공하신 분이다 보니 빈도통계가 상당히 마음에 들지 않는 모양이심. 여기에 나온 내용의 reference를 공부하면 큰 도움이 될건데, 어려워요.... 논문들이 진짜 넘 어려워서 안 읽혀요....ㅠㅠㅠ

그 외 구글링에 frequentist vs bayesian compariosn 검색하면 나오는 MIT opencourse 수학과 교수님들이 쓴 10장짜리? 강의노트 너무 좋아요 간단히 요약으로 정리됨 굳굳 그리고 stack exchange나 cross validated 이런 사이트에 열린 스레드 보면 ㄹㅇ 능력자들의 혜안이 담긴 글들이 많아서 보면서 링크 따라가고 참조문헌 공부해보면 이해가 더 깊어질거임!

5. Bayesian Approach: Fixed D, Random θ

5-1. Bayes Rule: "Inverse" Probability

먼저 베이즈 정리에 대해 간략히 소개하고, 이게 왜 혁신적인 발상의 전환인지 느껴보자.

베이즈 정리 자체는 product rule과 조건부 분포의 정의를 알면 바로 나온다.

$$p(C|E) = \frac{p(C,E)}{p(E)} = \frac{p(C)p(E|C)}{p(E)} = \frac{p(C)p(E|C)}{\sum_{C'} p(C')p(E|C')} \tag{26}$$

여기서 분모의 $p(E)$ 개별 C에 의존하지 않는 상수이다. 때문에 다음과 같이 쓰기도 한다.

$$p(C|E) \propto p(C)p(E|C) \tag{27}$$

어떤 사건 E가 발생했을 때 그 원인이 C일 확률은, 애초에 C가 발생활 확률과, 그 C가 발생했을 때 E의 확률의 곱에 비례한다는 것. 만일 $p(C)$만 알고 있으면 **"사건 발생 후 원인의 확률을 묻는 문제"가 "원인이 주어졌을 때 사건의 확률를 묻는 문제"로 바뀐다는 것.** 인과관계가 역전된 것이 보이나? 이 때문에 이를 Inverse Probability라고도 한다.

이걸 가설 $\mathcal{H}$에 한번 적용해볼까? 모수공간의 분할 $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3, \dots$ 을 생각해보자. 그러면 D가 주어지면 다음과 같이 쓸 수 있다.

$$p(\mathcal{H}|D) = \frac{p(D|\mathcal{H})p(\mathcal{H})}{p(D)} \tag{28}$$

- $p(\mathcal{H})$: 데이터를 보기 전에 모든 가설이 가지는 확률
- $p(D|\mathcal{H})$: 각각의 가설에서 데이터가 얼마나 Likely한지. 즉 데이터가 제공하는 가설의 Evidence
- $p(\mathcal{H}|D)$: 데이터를 바탕으로 업데이트된, 모든 가설 각각에 대한 확률
- $p(D)$: 모든 가능한 가설에서 주어진 데이터가 발생할 확률

여기에서 논란이 되는게 $p(\mathcal{H})$, 즉 prior이다. prior가 존재하지 않으니 인위적으로 prior를 만들면 어떻게든 나의 주관이 들어가게 되는거고, 그 순간 확률보다는 분포로 나타낸 가설에 대한 믿음이라는 해석이 더 적절하다. 그럼에도 불구하고 prior를 만들어서 계속 베이즈 정리를 쓰겠다는 것이 베이지안이고, 데이터를 보기 이전에 인위적으로 prior를 만든다는 발상 자체에 거부감을 느껴, "어떻게 하면 prior가 없이 통계 분석을 할 수 있을까"라는 물음에서 시작되어 Likelihood만 쓰는 방법이 빈도주의이다.

5-2. 베이즈 세계관 이해하기

이제 아까 빈도론적 추론에서 본 것과 똑같은 세팅을 가져오자.

$$\begin{aligned}
\text{Data (iid):} \quad & D = [x_1, x_2, \dots, x_n] \\
\text{Sampling Density of } D\text{:} \quad & f(D|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \quad (x_i \in \mathcal{X}, \theta \in \Omega) \\
\text{Likelihood of } \theta\text{:} \quad & L(\theta|D)
\end{aligned}$$

베이즈 추론에서 모수 θ는 "unknown thus uncertain". 즉 모르니까 확률변수다. $p(\theta)$는 데이터 D를 보기 전(prior) 나의 믿음이고, $p(\theta|D)$는 데이터를 보고 난 후(posterior)의 나의 믿음이다. 그렇다면 $p(\theta|D)$ 를 어떻게 얻는가? 베이즈 정리를 사용한다. **베이지안은 모든 통계분석을 베이즈정리로 한다. 베이즈정리가 알파이자 오메가이다!**

$$\begin{aligned}
\text{(Prior) Belief in } \theta \quad & p(\theta) \\
\text{Likelihood of } D \text{ at each } \theta \quad & L(\theta|D) = p(D|\theta) \\
\text{Updated (Posterior) Belief in } \theta \quad & p(\theta|x) = \frac{p(D|\theta)p(\theta)}{\int_\theta p(D|\theta)p(\theta)d\theta}
\end{aligned}$$

빈도론 추론과 가장 큰 차이점은 θ에 대한 추론의 결과를 분포로 제시한다는 것이다.

새로운 데이터 $\tilde{x}$에 대한 빈도론적 예측 분포는 다음과 같다.

$$\tilde{x} \sim p(x_i|\hat{\theta}) \tag{29}$$

새로운 데이터 $\tilde{x}$에 대한 베이지안 예측 분포는 다음과 같다.

$$\begin{aligned}
\text{Prior Predictive:} \quad & p(\tilde{x}) = \int p(x_i|\theta)p(\theta)d\theta \\
\text{Posterior Predictive:} \quad & p(\tilde{x}) = \int p(x_i|\theta, D)p(\theta|D)d\theta
\end{aligned}$$

5-3. 베이지안 논리의 직관적인 이해

이 부분은 이전에 만들어놓은 강의자료로 대체합니다!

BAYESIAN PROBABILITIES

- 확률의 빈도론적 정의는 확률시행을 무수히 반복할 때의 빈도, "long-run frequency"이다. 근데 이렇게 정의해버리면 "내일 비가 올 확률", "올해 여자친구가 생길 확률" 이런거는 확률이란 말을 쓰기가 애매해진다. 일평생 내일은 단 한번만 오고, 올해 연애 시도를 해봐야 뭐 한 번은 하겠나. 확률보다는 '믿음'이 더 맞겠다.
- 그러나 믿음, belief도 공교롭게도 확률의 세 가지 공리로 표현할 수 있다. 즉 사건의 발생할 확률을 그 사건이 발생할 것이라는 나의 믿음의 강도로 볼 수 있으며, 이런 식으로 불확실성을 직접적으로 정량화할 수 있다. 여기에다가 **Bayes Theorem**을 활용하면 이 믿음을 업데이트할 수 있다!
- 이는 우리의 직관과 상당히 일치한다. 올해 연애나 할 수 있겠어 하고 있는데, 어쩌다가 호감이 있는 상대와 데이트를 한 번 했다고 하자. 그렇다면 연애에 대한 희망이 생기지 않는가? 심지어 대여섯 번 더 만났다고 하자. 희망이 확신이 된다! 베이지언은 이 직관을 고스란히 반영한다.
  - 빈도론은 직관에 호소하기 위해 큰 고생을 해야한다. 통입에서 신뢰구간을 제대로 이해하는 수강생이 있기나 한가? 어려워서가 아니라 애초에 말이 안 돼서 그렇다. 현실에서는 표본평균이 오직 하나만 있기 때문이다. "수 만개의 평행우주에서 구한 표본평균들"과 같은 SF적 개념을 가져와야 이해를 할 수 있다.

BAYESIAN PROBABILITY

Bayes Theorem

- 베이즈 정리 자체는 product rule과 조건부 분포의 정의를 알면 바로 나온다.

$$p(C|E) = \frac{p(C,E)}{p(E)} = \frac{p(C)p(E|C)}{p(E)} = \frac{p(C)p(E|C)}{\sum_{C'} p(C')p(E|C')}$$

- 여기서 분모의 $p(E)$ 개별 C에 의존하지 않는 상수이다. 때문에 다음과 같이 쓰기도 한다.

$$p(C|E) \propto p(C)p(E|C)$$

어떤 사건 E가 발생했을 때 그 원인이 C일 확률은, 애초에 C가 발생활 확률과, 그 C가 발생했을 때 E의 확률의 곱에 비례한다는 것. 만일 $p(C)$만 알고 있으면 **"사건 발생 후 원인의 확률을 묻는 문제"가 "원인이 주어졌을 때 사건의 확률를 묻는 문제"로 바뀐다는 것.**

- 인과관계가 역전된 것이 보이나? 이 때문에 이를 Inverse Probability라고도 한다. 20세기까지만 해도 대부분의 통계학자는 베이지언을 "불경한 것"으로 혐오했다. 왜 그들은 베이지안을 싫어했을까?

BAYESIAN PROBABILITY

Bayes Theorem: Example

- 올해 연애를 할 확률(믿음)은 $p(Luv) = 0.1$, 못 할 확률은 $p(Sad) = 0.9$라고 하자. 올해 연애를 한다면 그 전에 데이트를 좀 할 것이다. 때문에 $p(Date|Luv) = 0.8$, $p(Date|Sad) = 0.3$이라고 하자(둘이 합쳐서 1이 안된다? 조건이 다르면 다른 분포니까!). 어쩌다보니 눈이 마주친 그대와 데이트를 했다. 그렇다면 내가 올해 연애할 확률은 어떻게 됐을까?

$$p(Luv|Date) = \frac{p(Date|Luv)p(Luv)}{p(Date|Luv)p(Luv) + p(Date|Sad)p(Sad)} = \frac{0.8 \times 0.1}{0.8 \times 0.1 + 0.3 \times 0.9} \approx 0.2$$

- 데이트를 한 번 더 했다면 어떻게 될까? 이 경우 $p(Luv) = 0.2$이다. 똑같은 방식으로

$$p(Luv|Date) = \frac{p(Date|Luv)p(Luv)}{p(Date|Luv)p(Luv) + p(Date|Sad)p(Sad)} = \frac{0.8 \times 0.2}{0.8 \times 0.2 + 0.3 \times 0.8} = 0.4$$

공고해지는 믿음을 보아라. 따스한 한 해가 만들어지고 있다.

BAYESIAN PROBABILITY

Bayesian Inference

- 모수 추정의 문제에서 빈도론적 추론과 베이지언 추론을 비교해보자. $X \sim p(x|\theta)$에서 θ를 추정하는 문제다. 표본 x가 주어졌을 때 $p(x|\theta)$는 Likelihood로 볼 수 있음을 배웠다.
- **Frequentist MLE:**
  주어진 Likelihood를 최대화하는 θ를 추정량으로 삼는다.

$$\theta_{ML} = \arg\max_{\theta} p(x|\theta)$$

  이는 근본적으로 **데이터가 주어졌을 때 데이터가 나올 확률 $p(x|\theta)$을 극대화**하는 방법이다. 뭐 극한에서는 말이 되긴 한다. 그러나 제한된 표본에서는? 동전 3개 던져서 다 앞면 나오면 뒷면이 나올 확률은 0인가? **MLE는 태생적으로 Overfitting을 하게 된다.**
  - 빈도통계학에서 MLE가 정말 중요한데, 왜냐면 MLE는 CLT처럼 그 극한 분포가 알려져 있기 때문이다. 때문에 추정량의 극한 분포로 신뢰구간, 가설검정 등의 빈도론적 추론을 할 수가 있다. 그러나 머신러닝에서 다루는 대부분의 문제는 $n >> p$가 아니다. 때문에 MLE를 썼다가는 과적화가 되기 마련. 때문에 이를 보완하는 다양한 방법이 있다.

BAYESIAN PROBABILITY

Bayesian Inference

- **Bayesian MAP**:
  좀 더 자연스러운 생각은 **데이터가 주어졌을 때 모수의 확률 $p(\theta|x)$을 극대화**하는 것이다.

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|x)$$

  모수의 확률? 18~20세기 주류 통계학계가 거품을 문 포인트다. 아니 감히 전지전능한 하나님만이 아는 자연의 섭리를, 미천한 인간은 모르지만 엄연히 존재하는, 고정된 모수를 감히 "확률변수"로 취급하다니? 여기서 베이지안과 빈도론자의 철학에 큰 차이가 나타난다.
- 데이터의 생성에 대한 Sampling Density, Likelihood를 $p(x|\theta)$로 본다고 하자.
- **빈도론자:** 모수 θ는 "unknown but certain". 모르지만 고정된 상수이다. 데이터는 하나밖에 없지만 수만 개의 평행우주에는 똑같은 분포를 따르는 수만 개의 다른 데이터가 있을 것이다. 그 데이터들 모두를 가장 잘 설명하는 하나의 최대점이 참 모수값이다. 그러니 지금 가지고 있는 하나의 데이터만을 가장 잘 설명하는 추정치 θ_{ML}를 써도 되지 않겠나!

BAYESIAN PROBABILITY

Bayesian Inference

- **베이지언:** 모수 θ는 "unknown thus uncertain". **모르니까 확률변수야!** 하나님만이 아는 정답같은 건 잘 모르겠고 내가 그 정답에 대해 얼마나 잘 모르냐는 알겠다. 그러니 나는 θ에 대한 나의 믿음을 확률로 표현할거다. 내 맘이다.
  $p(\theta)$는 데이터를 보기 전(prior) 나의 믿음이고, $p(\theta|x)$는 데이터를 보고 난 후(posterior)의 나의 믿음이다. 그렇다면 $p(\theta|x)$ 를 어떻게 얻는가? 베이즈 정리를 사용한다!

$$\begin{aligned} p(\theta) &\quad \text{Belief in each value of } \theta \text{ prior to data} \\ p(x|\theta) &\quad \text{Likelihood of the data per each value of } \theta \\ p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int_\theta p(x|\theta)p(\theta)d\theta} &\quad \text{Belief in each value of } \theta \text{ posterior to data} \end{aligned}$$

베이지안은 모든 통계분석을 베이즈정리로 한다. 베이즈정리가 알파이자 오메가이다!

BAYESIAN PROBABILITY

Bayesian Inference

이전에 든 올해 연애 예시를 들어보자. 이 경우 표본은 데이트 여부이며, 모수는 올해 연애를 할 여부이다. **연애를 하는지 안 하는지에 따라 데이트의 분포가 달라진다.** 즉

$$\textbf{Sampling Density: } \begin{cases} p(Date|\theta = 1) & = 0.8 \\ p(Date|\theta = 0) & = 0.3 \end{cases}$$

- 먼저 빈도론자처럼 생각해보자. θ는 0 혹은 1 둘 중 하나이며, 그 값은 운명의 세 여신 모리아이 자매만 알고 있다. 미천한 인간은 데이트 한 번 하고 나서 이 θ가 0인지 1인지를 결정해야 한다. MLE 원칙에 충실한 빈도론자는 "0.8이 0.3보다 크네"하고 올해 연애를 한다고 결론을 내린다.

$$\therefore \ \theta_{ML} = 1$$

BAYESIAN PROBABILITY

Bayesian Inference

$$\textbf{Sampling Density: } \begin{cases} p(Date|\theta = 1) & = 0.8 \\ p(Date|\theta = 0) & = 0.3 \end{cases}$$

- 베이지안은 이렇게 말한다. **애초에 너가 연애를 할 확률이 굉장히 낮지 않을까?** 아니 뭐 물론 올해 연애한다면 데이트는 당연히 하겠지. 그렇지만 어쩌다 한번 데이트한거 가지고 설레발치는게 아닐까?
- 즉 만일 θ의 값을 하나 골라야한다면, Likelihood 뿐만이 아니라 Prior도 고려해야 한다는 것이다. 이것이 MAP이다. 베이즈 정리에서 분모는 θ의 값에 상관없이 똑같다. 때문에 분자만 고려해보면,

$$\begin{cases} p(Date|\theta = 1)p(\theta = 1) & = 0.8 \times 0.1 = 0.08 \\ p(Date|\theta = 0)p(\theta = 0) & = 0.3 \times 0.9 = 0.27 \end{cases}$$

$$\therefore \ \theta_{MAP} = 0$$

.png" alt="CH01-17" style="zoom:67%;" />

BAYESIAN PROBABILITY

Bayesian Inference

- 하지만 찐 베이지언은 MAP를 하지 않는다. 사실 MLE나 MAP나 똑같다. 전자는 Likelihood라는 함수의 최댓값을 뽑는 거고, 후자는 Likelihood와 Prior까지 같이 고려해 최댓값을 뽑는거니, 결국은 하나의 θ 추정치를 쓴다는 것에서는 똑같다.
- 그러나 베이지안의 철학은 모수도 확률변수라는 게 아닌가! 확률변수를 감히 하나의 값으로 표현할 수 있는가? 확률변수를 표현하는 가장 완전한 방법은 그 분포를 온전히 그려내는 것이다! 즉

$$\textbf{Posterior Belief in } \theta \begin{cases} p(Date|\theta = 1)p(\theta = 1)/p(Date) & = 0.8 \times 0.1 \approx 0.2286 \\ p(Date|\theta = 0)p(\theta = 0)/p(Date) & = 0.3 \times 0.9 \approx 0.7714 \end{cases}$$

이 분포를 드러내기 위해 평균 0.22을 쓸 수도 있고, 극빈값 0을 쓸 수도 있다. θ가 연속일 경우 95% **확률**구간을 쓸 수도 있다. 중요한 것은 θ에 내재한 불확실성 구조를 그대로 가져간다는 것!

예시: 지금 신촌에 비가 올까?

(미국 Facebook 면접문제 변용) 오늘 아침 일기 예보를 보니 신촌에 비가 올 확률이 25%이라고 합니다. 아침을 먹고 현관문을 나서기 전, 당신은 자취하는 친구 세 명에게 지금 비가 오는지 따로따로 물어봅니다. 모두 지금 비가 온다고 합니다. 하지만 당신의 친구들은 당신을 곯리기 위해 세 번 중 한 번은 거짓말을 합니다. 그렇다면 당신은 우산을 챙길 것인가요? 그 이유와 함께 설명해주세요.

인터넷에 떠도는 페이스북 면접 문제를 가져와서 변용하였다. 이 문제에서 친구들의 답변 YYY를 Data, 비가 내리는 여부를 모수 θ로 생각하면, 다음과 같은 두 풀이가 가능하다.

1. **MLE 접근**. 비가 내려는 분포에서 YYY의 확률은 $p(YYY|rain) = (2/3)^3 = 8/27$, 비가 내리지 않는 분포에서 YYY의 확률은 $p(YYY|rain) = (1/3)^3 = 1/27$. 비가 내릴 때의 표본의 Likelihood가 더 크니 우산을 챙긴다.
2. **Bayes Rule 접근**. 비가 내리는 사건의 사전 확률(믿음)은 0.25이다. 데이터를 바탕으로 이 믿음을 업데이트하면

$$\begin{aligned} p(Rain|YYY) &= \frac{p(YYY|Rain)p(Rain)}{p(YYY|Rain)p(Rain) + p(YYY|NoRain)p(NoRain)} \\ &= \frac{8/27 \times 1/4}{8/27 \times 1/4 + 1/27 \times 3/4} \\ &= \frac{8 \times 1}{8 \times 1 + 1 \times 3} = 8/11 > 0.25 \end{aligned}$$

   즉 친구들의 대답으로 인해 나의 믿음이 0.25에서 0.72로 올라갔으니 우산을 챙긴다는 것.

결론은 똑같지만 사고 과정이 다르다.

References

1. Probability Theory and Statistical Inference: Econometric Modeling with Observational Data (Spanos, 1999)
2. Machine Learning: a Probabilistic Perspective (Murphy, 2012)
3. Computer Age Statistical Inference (Efron, Hastie, 2016)
4. Calibration of p Values for Testing Precise Null Hypotheses (Sellke et al, 2001)
5. https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading20.pdf