EBS 기술인협회 스터디

『넷플릭스 기술 연구회』 3부

EBS 기술인협회 스터디
『넷플릭스 기술 연구회』

지난 9월호에서 넷플릭스 기술 연구회는 '차세대 이미지 코딩을 위한 AVIF'와 '예측 모델링으로 넷플릭스의 콘텐츠 품질관리 최적화'라는 두 가지 기사를 소개했습니다. 이번 3부에서는 'A/B 테스트란 무엇인가?'라는 기사를 파트 1부에서 4부까지 소개해드리도록 하겠습니다. 넷플릭스에서 어떠한 방식으로 의사결정의 근거를 찾고, 어떤 방식으로 통계 자료를 해석하는지 포스팅을 통해 알아보도록 하겠습니다. 해당 포스팅이 이 글을 읽는 독자분들에게 도움이 되는 자료가 되었으면 합니다.

A/B 테스트란 무엇인가?

원본 정보

제목	What is An A/B Test? (Part 1 ~ Part 4) A/B 테스트란 무엇인가? (파트 1 ~ 파트 4)
Link	https://netflixtechblog.com/what-is-an-a-b-test-b08cc1b57962

그림 1. 원문 QR-Code

이번 포스팅에서는 넷플릭스가 A/B 테스트를 활용하여 구독자들에게 더 많은 기쁨과 만족감을 제공할 수 있는 의사결정을 내리는 과정에 대해 설명해 드리고자 합니다. 이후의 포스팅에서는 A/B 테스트의 기본 통계 개념, Netflix 내에서 실험의 역할, 그리고 Netflix 내에서 실험 문화의 중요성에 대해 다룰 것입니다.

Decision Making at Netflix - 넷플릭스의 의사결정 프로세스

넷플릭스는 소비자들의 선택을 기반으로, 소비자들이 가장 즐거워할 수 있는 환경을 만들어 나가고자 하는 목표로 운영되고 있습니다. 그 가치를 계속 유지하기 위해 서비스를 지속해서 발전시키고자 노력하고, 그런 예시 중 하나로 넷플릭스 웹사이트의 대규모 UI 개편이 있었습니다.

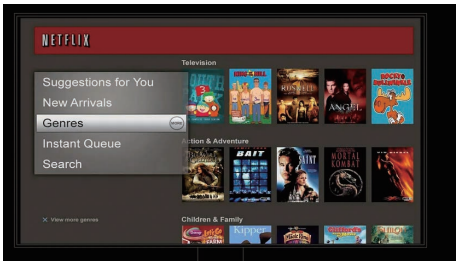
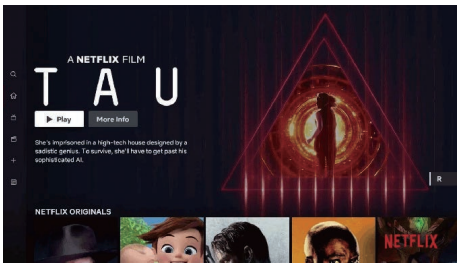
2010년 넷플릭스 UI	2020년 넷플릭스 UI
	
• 정적인 UI • 한정적인 내비게이션 옵션 • 비디오 렌탈샵을 연상하게 함	• 동적인 비디오 중심적 UI • 다양한 내비게이션 옵션

그림 2. 넷플릭스 2010 UI vs. 2020 UI

끊임없이 이어지는 다양한 질문들에 대한 의사결정을 통해 대규모 UI 개편이 성공적으로 이루어질 수 있었습니다. “정적인 이미지보다 비디오를 활용하는 것이 나은지?”, “내비게이션 메뉴는 어디에 위치해야 하며, 어떤 요소를 포함해야 하는지?”, “제약된 네트워크에서 원활한 경험을 어떻게 제공할 수 있을지?” 등등 다양한 고민에 결정을 내려야만 했습니다. 포스팅에선 이렇게 말합니다.

> “Making decision is easy – what’s hard is making the right decisions
> 결정을 내리기는 쉽다 – 옳은 결정을 내리는 것이 어렵다.”

넷플릭스는 이러한 과정에서 옳은 결정을 내릴 수 있게 해주는 방식이 바로 다양한 실험들이고, 그 중 대표적인 실험이 바로 A/B 테스트라고 말합니다.

What is an A/B Test? - A/B 테스트란 무엇인가?

A/B 테스트는 두 개의 그룹, A와 B를 비교해 더 나은 성과를 비교하는 실험이며, 4가지의 특징이 있습니다:

1. 통제된 실험(Controlled Experiment)

통제된 환경에서 변수를 조절하여 인과 관계를 확인하려는 실험 방법입니다.

2. 실험 그룹 & 대조 그룹(Experimental Group & Control Group)

- 실험 그룹 : 실험의 변인을 받는 그룹으로, 새로운 아이디어, 기술, 제품 또는 접근 방식을 테스트하는 데 사용됩니다. 실험 그룹의 결과는 일반적으로 대조 그룹과 비교하여 변화의 효과를 파악하기 위해 사용됩니다.

- 대조 그룹 : 실험 그룹과 동일한 조건에서 실험을 수행하지 않고, 기존의 시스템 또는 접근 방식을 계속 사용하는 그룹입니다. 이러한 그룹 간 비교를 통해 실험 결과의 효과를 정량적으로 평가할 수 있습니다.

3. 가설(Hypothesis)

연구나 실험을 통해 검증하고자 하는 주장이나 예측을 나타내는 문장 또는 명제입니다. 가설은 연구의 출발점이며, 연구의 목적과 방향을 제시합니다.

4. 무작위 할당(Random Assignment)

연구 대상을 실험 그룹과 대조 그룹으로 무작위로 나누는 과정을 의미합니다. 이것은 실험에서 내재한 편향을 줄이고 연구 결과를 신뢰성 있게 만드는 중요한 요소 중 하나입니다.

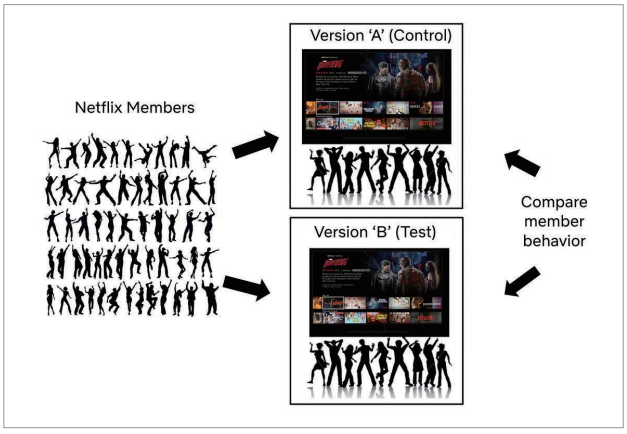


그림 3. 넷플릭스 A/B 테스트 예시

해당 포스팅에서 이해를 쉽게 돕기 위해 가상의 예시를 하나 듭니다. 만약에, 넷플릭스 UI 개편을 한 번 더 진행하여 메인 페이지 박스아트를 거꾸로 뒤집는다면 구독자들의 참여도가 증가할지에 대해 테스트를 해보려고 합니다.

가설(Hypothesis)	거꾸로 뒤집힌 박스 아트 UI가 이용자 참여도를 증가시킨다.

여기서 이용자 참여도가 해당 실험의 **주요 결정 지표***가 됩니다.

* **주요 결정 지표** (Primary Decision Metric) : 실험의 주요 결과 또는 관심사를 나타내는 지표입니다. 실험에서 독립 변수(조작된 변수)의 영향을 측정하거나, 그룹 간 차이를 확인하는 데 사용됩니다.

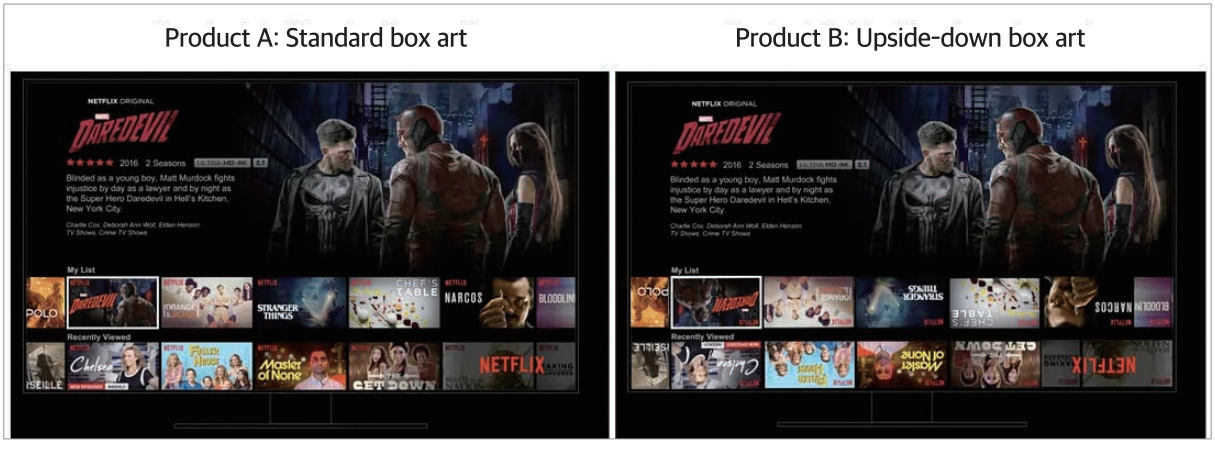


그림 4. 기본 UI vs. 실험 UI

이 예시로 통제된 실험의 중요성을 이해해 볼 수 있습니다. 독립 변수인 거꾸로 뒤집힌 박스 아트, 즉 변화된 UI 외에는 모든 것을 다 동일하게 둔 채로 실험이 진행되어야만 합니다. 그것은 실험이 진행되는 시간과 기간 또한 포함됩니다. 예를 들어, 한 달이라는 실험 기간에 프로덕트 A와 B를 동시에 노출하는 것이 아닌 첫 15일 동안은 프로덕트 A만 노출, 남은 15일 동안은 프로덕트 B만 노출한다면, 아래와 같은 결괏값이 나오게 됩니다.

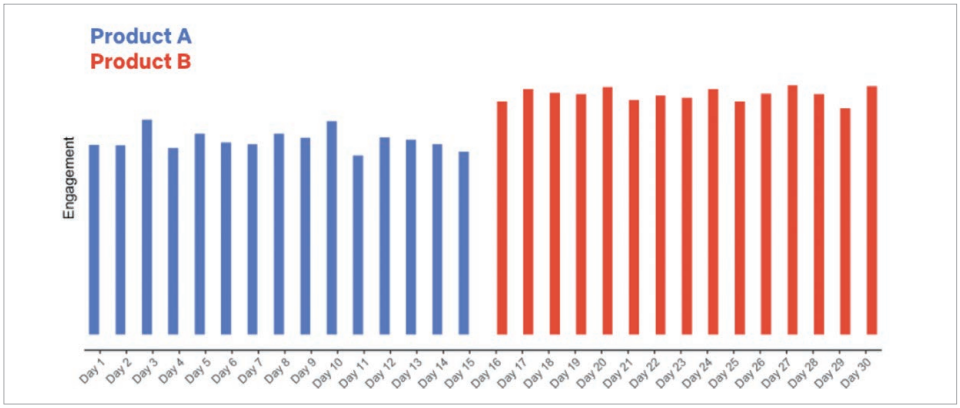


그림 5. 16일째 되는 날에 거꾸로 뒤집힌 UI를 출시했다는 가상의 데이터

해당 데이터를 보았을 땐 거꾸로 뒤집힌 박스 아트 UI인 프로덕트 B가 구독자들의 참여도를 더 높인다는 결론을 낼 수 있습니다. 하지만, 만약에 마침 16일이 되는 날에 <기묘한 이야기>와 같은 블록버스터 작품이 출시되었었다면 얘기는 달라집니다. 16일부터 증가한 참여도가 거꾸로 뒤집힌 UI 때문인지에 대한 확신을 잃게 되기 때문입니다. 그렇기 때문에, 제대로 된 A/B 테스트를 진행하기 위해선 동일한 기간에 두 가지의 UI를 노출해야 합니다. 무작위 할당을 통해 구성한 실험 그룹과 대조 그룹에 30일이라는 동일한 기간에 프로덕트 A와 B를 노출했을 때 결괏값은 다음과 같습니다.

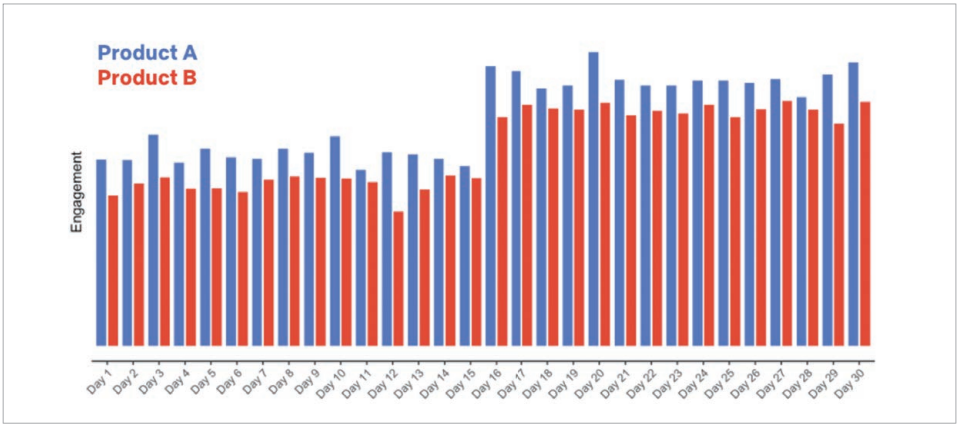


그림 6. 동일 기간 두 개의 프로덕트를 노출 시킨 결과

박스 아트가 거꾸로 뒤집힌 UI는 일반적으로 참여도가 낮고, 두 그룹 모두 인기 작품 출시와 함께 참여도 증가세를 보입니다. 이처럼 통제할 수 있는 부분들을 전부 통제하는 A/B 테스트는 인과적인 결론에 도달할 수 있습니다.

가상의 예시이기에 주요 결정 지표인 '참여도' 한 가지만 두고 결론을 내릴 수 있었지만, 넷플릭스에서 실제 A/B 테스트를 진행할 땐 조금 더 세분화된 지표들도 함께 고려하게 됩니다. 'Top 10' 카테고리를 메인 UI에 추가할지 말지에 대한 A/B 테스트에선, 주요 의사결정 지표인 이용자들의 참여도뿐만이 아니라, **보조 지표***인 (Secondary Metric) Top 10 목록에 표시된 콘텐츠들의 프로그램별 시청률, 그리고 Top 10과 기타 UI 영역에서의 시청률 차이 또한 고려하여 진행하였습니다.

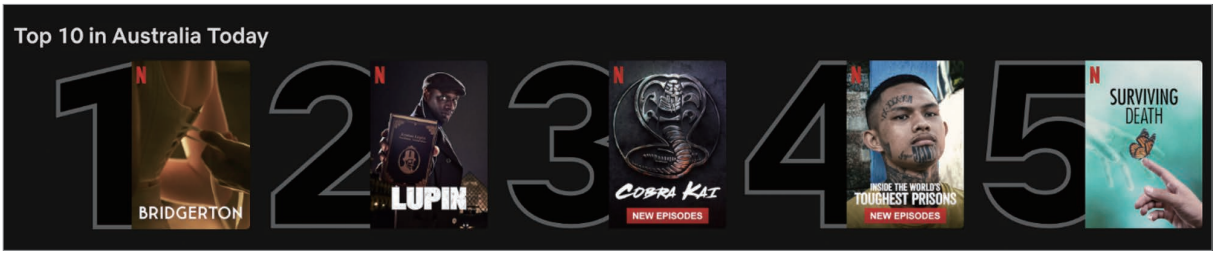


그림 7. 넷플릭스 Top 10 UI

더 나아가 **가드레일 지표*** 또한 함께 고려하여, 변화의 부작용을 최소화할 수 있도록 하였습니다. 예를 들어, 실험 그룹과 대조 그룹의 고객센터 연락률을 비교하여 새로 추가된 'Top 10' 큐레이션 기능이 연락률을 증가시키지 않는지 확인합니다. 해당 지표는 회원들의 혼란 또는 불만을 상징할 수 있기 때문입니다.

***보조 지표** (Secondary Metric) : 주요 결정 지표는 주요 목표를 나타내지만, 보조 지표는 해당 목표를 달성하기 위한 과정 또는 부수적인 영향을 측정하는 데 사용됩니다. 이러한 메트릭은 실험 결과를 보다 심층적으로 이해하고 해석하는 데 도움이 됩니다.

***가드레일 지표** (Guardrail Metric) : 제품 또는 서비스 변경 사항의 부작용을 모니터링하고 예기치 않은 부정적인 영향을 방지하기 위해 사용되는 지표입니다.

Interpreting A/B test results: false positives and statistical significance
A/B 테스트 결과 해석하기: 거짓 양성과 통계적 유의성

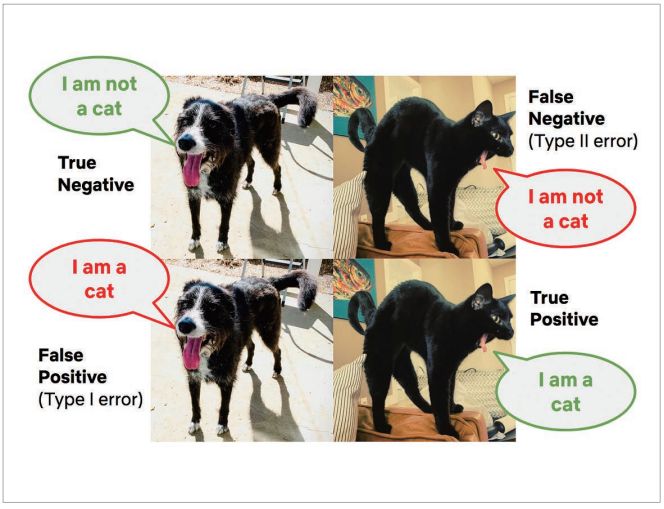


그림 8. 거짓 양성 vs. 거짓 음성

이번 파트에서는 A/B 테스트의 결과를 분석 및 해석하기에 앞서, 통계적 오류에 관해서 설명합니다. 넷플릭스는 통계적 결과엔 항상 불확실성과 실수가 존재하며, 그 가능성을 완전히 없앨 수 없다는 것을 인지하는 것이 중요하다고 얘기합니다. 다양한 통계적 오류를 이해하는 프레임워크를 사용하여, 불확실성을 신중하게 양적화하는 과정이 필요하다고 합니다.

테스트 결과에 관한 결정을 내릴 때 저지르는 실수엔 두 가지 유형이 있습니다: 거짓 양성과 거짓 음성입니다.

1. 거짓 양성 (1종 오류)

거짓 양성은, [그림 8]에서 보았을 때, 결괏값은 "I am a cat"이라고 하지만, 실제론 고양이 사진이 아닌 개 사진이 있는 경우입니다. 일상적 예시로는 코로나에 걸리지 않았지만, 결과만 양성으로 나왔을 때 거짓 양성이라는 통계적 오류가 있다고 볼 수 있습니다. A/B 테스트에 적용한다면, 테스트 결과가 유의미하다고 데이터가 말하고 있지만, 실제론 유의미한 결과가 없는 경우입니다.

2. 거짓 음성 (2종 오류)

거짓 음성은, 그림에서 보았을 때, 결괏값은 "I am not a cat"이라고 하지만, 실제론 고양이 사진이 있는 경우입니다. 일상적 예시로는 코로나에 걸렸지만, 결과가 음성으로 나왔을 때 거짓 음성이라는 통계적 오류가 있다고 볼 수 있습니다. A/B 테스트에 해당 오류를 적용한다면, 테스트 결과가 유의미하지 않다고 데이터가 말하지만, 실제론 유의미한 결과가 있는 경우입니다.

위 두 가지의 통계적 오류를 완전히 없앨 순 없다고 넷플릭스는 말합니다. 실제로, 두 가지 오류는 서로 상충합니다: 테스트에 거짓 양성의 비율이 거의 없게끔 설계를 하게 되면, 거짓 음성의 비율은 반드시 증가하게 되고, 반대로 테스트에 거짓 음성의 비율을 줄이려면 거짓 양성의 비율이 증가할 수밖에 없다고 합니다. 그렇기 때문에, 넷플릭스는 더욱 오류의 원인을 수량화하고 이해하며 통제하기 위해 노력해야 한다고 주장합니다.

A/B 테스트 결과가 통계적 유의성을 확보하고 있는지에 대한 기준점 설정은 일반적으로 허용 가능한 거짓 양성 비율을 설정함으로써 시작됩니다. 관례로, 이 거짓 양성 비율은 5%로 설정됩니다. 실험 그룹과 대조 그룹 사이에 실제론 유의미한 차이가 없지만, 통계적으로 유의미한 차이가 있다는 잘못된 결론에 도달하게 될 확률이 5%라는 뜻입니다. 이럴 경우, 테스트는 "5%의 유의수준에서 실행되었다"라고 말합니다.

거짓 양성 비율은 실험 그룹과 대조 그룹 사이의 지표 값들 차이의 통계적 유의성과 밀접한 관련이 있으며, 이는 P-값을(P-value) 사용하여 측정합니다.

3. 통계적 유의성(Statistical Significance)

주어진 데이터와 가설 간의 관련성 또는 차이를 평가하는 과정에서 사용됩니다. 통계적 유의성은 주어진 가설 검정에서 얻은 결과가 우연에 의한 것이 아니라 실제로 관찰된 패턴 또는 차이의 정도를 나타내는 지표입니다. 이 지표는 주로 P값(P-value)을 통해 확인됩니다. P값이 작을수록, 즉 0.05 또는 이하인 경우, 결과는 통계적으로 유의미하다고 판단됩니다.

4. P값(P-value)

P값은 '귀무가설(null hypothesis)'을 검정하는 데 사용됩니다. 귀무가설은 주로 "효과가 없다." 또는 "두 그룹 간에 차이가 없다"와 같이 어떤 가정을 나타냅니다. 예를 들어, 넷플릭스 'Top 10' 카테고리 추가 여부에 대한 A/B 테스트를 진행했을 시, 귀무가설은 "Top 10 큐레이션과 이용자 참여도는 상관관계가 없다"일 것입니다.

P값이 작을수록 귀무가설이 기각되고, '대립 가설(alternative hypothesis)'이 받아들여질 가능성이 커집니다.

일반적으로 P값이 0.05 또는 이하인 경우, 통계적으로 유의미한 차이가 있다고 판단됩니다.

귀무가설	Top 10 큐레이션과 이용자 참여도는 상관관계가 없다.
P-value<0.05	Top 10 큐레이션과 이용자 참여도는 상관관계가 있다
P-value>0.05	Top 10 큐레이션과 이용자 참여도는 상관관계가 없다

Ways to increase power - 검정력을 높이는 방법

A/B 테스트를 설계할 때, P-값을 위와 같이 설정하고, 거짓 양성 및 음성 확률을 통제합니다. 가짜 양성 및 음성의 확률을 최소화하고, 검정력을 높이기 위한 주요 방법은 세 가지가 있습니다.

1. 효과 크기(Effect Size) : 변수들 사이의 관계가 얼마나 의미 있는지, 즉 A와 B 그룹 사이의 차이가 얼마나 의미 있는지를 나타내며, 효과 크기가 클수록 오류 발생 가능성이 낮아집니다.

2. 표본 수(Sample Size) : 실험 단위가 많을수록 검정력이 높아져 오류 발생 가능성이 낮아집니다. 예를 들어, 동전 던지기를 했을 시, 동전을 20번 던져서 나온 결괏값보다 동전을 100번 던져서 나온 결괏값이 더 신뢰도가 높을 것입니다. 제품 개발의 맥락에서 봤을 때, 실험 그룹에 더 많은 이용자/구독자를 할당하면 할수록 오류 발생 가능성을 낮출 수 있습니다.

3. 피실험자 산포도(Variability) : 실험 대상들 사이의 동질성이 얼마나 높은지를 나타내는 지표이며, 실험 대상이 서로 비슷할수록 오류 발생 가능성이 낮아집니다.

마치며

지금까지 넷플릭스에서 어떠한 방식으로 의사결정의 근거를 찾고, 어떤 방식으로 통계 자료를 해석하는지에 대해 소개하였습니다. A/B 테스트는 다양한 영역에서 활용될 수 있지만, 특히 UI/UX 개선과 온라인 광고 및 마케팅적 요소에서 보다 손쉽고 효과적으로 응용할 수 있다고 생각합니다. 추후 기회가 된다면 EBS에서 진행하는 다양한 온라인 사업에도 적용하여 더 나은 서비스를 이용자들에게 제공할 수 있게 되길 바랍니다.

그림 9. PT 토론