3. 통계적 실험과 유의성 검정

실험설계는 사실상 모든 응용 연구 분야에서 통계분석의 토대가 된다. 실험설계는 어떤 가설을 확인하거나 각각하기 위한 목표를 갖고 있다. 특히나 **데이터 과학자들은 사용자 인터페이스나 제품 마케팅 실험과 같이 지속적으로 어떤 실험을 수행해야 하는 상황이 온다.**

이 장에서는 전통적인 실험설계에 대해 알아보고 데이터 과학에도 적용되는 몇몇 어려움에 대해 논의한다. 또한 통계적 추론에서 자주 인용되는 일부 개념들을 다루고 데이터 과학에서의 의미와 관련성(또는 무관련성)을 설명한다.

통계적 유의성, t검정, p값 등에 대한 자료를 찾아보면, 전형적인 통계적 추론이라는 '파이프라인'속에 있음을 알 수 있다.

1) 이 과정은 '약품 A가 기존의 표준 약품보다 낫다', '가격 A가 기존 가격 B보다 수익성이 높다'라는 식의 가설을 세우는 것에서 출발한다. (가설을 세운다.)

2) (A/B 검정과 같은) 실험은 가설을 검정하기 위해 설계되고, 원하는 최종적인 결론을 도출할 수 있도록 설계한다. (실험을 설계한다.)

3)데이터를 수집하고 (데이터를 수집한다,)

4)결론을 도출한다. (추론 및 결론을 도출한다.)

추론 이라는 용어는 제한된 데이터로 주어진 실험 결과를 더 큰 과정 또는 모집단에 적용하려는 의도를 반영한다.

3.1 A/B 검정

A/B 검정은 두 처리 방법, 제품, 혹은 절차 중 어느 쪽이 다른 쪽보다 더 우월하다는 것을 입증 하기 위해 실험군을 두 그룹으로 나눠 진행하는 실험이다. 종종 두 가지 처리법 중 하나는 기준이 되는 기존 방법이거나 아예 아무런 처리도 적용하지 않는 방법이 된다. 이를 **대조군**이라고 한다. 새로운 처리법을 적용하는 것이 대조군보다 더 낫다는 것이 일반적인 가설이 된다.

용어 정리

- **처리 : 어떤 대상에 주어지는 특별한 환경이나 조건(약, 가격, 인터넷 뉴스 제목)**
- **처리군(처리 그룹) : 특정 처리에 노출된 대상들의 집단**
- **대조군(대조 그룹) : 어떤 처리도 하지 않은 대상들의 집단**
- **임의화(랜덤화) : 처리를 적용할 대상을 임의로 결정하는 과정**
- **대상 : 처리를 적용할 개체 대상(유의어 : 피실험자)**
- **검정통계량 : 처리 효과를 측정하기 위한 지표**

A/B 검정은 그 결과를 쉽게 측정할 수 있으므로 웹 디자인이나 마케팅에서 일반적으로 사용된다. A/B 검정의 몇 가지 예는 다음과 같다.

- 종자 발아가 어디에서 더 잘되는지 알아보기 위해 두 가지 토양 처리를 검정한다.
- 암을 더 효과적으로 억제하는 두 가지 치료법을 검정한다.
- 두 가지 가격을 검정하여 더 많은 순이익을 산출하는 쪽을 결정한다.
- 두 개의 인터넷 뉴스 제목을 검정하여 더 많은 클릭을 생성하는 쪽을 결정한다.
- 두 개의 인터넷 광고를 검정하여 어느 것이 더 높은 전환율을 얻을지 판단한다.

제대로 된 A?B 경우에는 둘 중 어느 한 쪽 처리를 할당할 수 있는 **대상**이 주어진다. 대상은 사람이 될 수도 있고, 식물의 씨앗이나 웹 방문자가 될 수도 있다. 핵심은 피험자가 어떤 특정 처리에 노출된다는 것이다. 이상적으로, 피험자는 **무작위**로 어느 처리에 할당된다. 그러면 처리 그룹간의 차이는 다음 두 가지 이유 중 하나 때문이라고 할 수 있다.

- 다른 처리의 효과
- 어떤 대상이 어떤 처리에 배정될지에 대한 경우의 수(즉, 무작위로 배정한 결과 자연스럽게 더 좋은 결과를 보이는 대상들이 A 또는 B 한쪽에 집중됨)

또한 그룹 A와 그룹 B를 비교하는데 사용하는 **검정통계량** 또는 측정 지표에 주의를 기울여야 한다. 데이터 과학에서 일반적으로 사용되는 지표는 클릭/클릭하지 않음, 구매/구매하지 않음, 사기/사기아님 등과 같은 이진 변수이다.

CAUTION R을 포함하여 모든 통계 SW가 디폴트로 어떤 결과들을 보여준다고 해서, 모든 출력이 유용하거나 관련 있다는 말은 아니다. 앞선 예에서 표준편차는 그다지 유용한 정보가 아니다. 딱 봐도, 수익이 음수가 될 수 없는데 결과는 값들이 음수일 수 있다고 제안하고 있다. 이 데이터는 적운 수의 상대적으로 높은 값(전환이 있는 페이지 뷰)과 많은 수의 0(전환이 없는 페이지 뷰)으로 구성된다. 이처럼 데이터의 변동성을 숫자 하나로 요약한다는 것은 참 어렵다. 이 경우에는 평균절대편차(A의 경우 7.68, B의 경우 8.15)가 표준편차보다 합리적이라고 볼 수도 있다.

3.1.1 대조군은 왜 필요할까?

대조군 없이, 관심 있는 처리를 한 그룹에만 적용해서 실험을 하고 그 결과를 단순히 이진 경험과 비교해 보면 안 될까?

대조군이 없다면 '다른 것들은 동일하다'는 보장이 없으며 어떤 차이가 처리(또는 우연) 때문인지 확신할 수 없다. 대조군의 경우, 관심 처리를 뺀 나머지는 처리 그룹과 동일한 조건이 적용된다. 단순히 '기준선' 또는 이전 경험과 비교할 경우, 처리 이외의 다른 요소가 다를 수도 있기 때문이다.

NOTE_ 연구를 위한 눈가림

눈가림 연구란 피실험자가 처리 A나 처리 B 중 어느 것을 받고 있는지 알지 못하도록 하는 연구 방식이다. 특정 처리를 받는 것에 대한 인식이 반응에 영향을 줄 수 있기 때문이다. 이중눈가림 연구는 조사자와 진행자 모두가 어떤 대상이 어떤 처리를 받았는지 모르게 하는 연구이다. 물론 '컴퓨터 대 심리학자로부터 받는 인지 치료의 차이'와 같이 처리의 성격이 투명할 때는 눈가림 연구가 불가능하다.

데이터 과학분야에서 A/B 검정은 웹 환경에서 많이 사용된다. 웹 페이지의 디자인. 제품의 가격. 헤드라인의 어감 등 많은 항목이 처리 조건이 될 수 있다. 무작위 원칙을 지키기 위해서는 약간의 아이디어가 필요하다. 이때, 대상은 일반적으로 웹 방문자이며 측정하고자 하는 결과는 클릭 수, 구매 수, 방문 기간, 방문한 페이지 수. 특정 페이지 방문 여부 등이다.

일반적인 A/B 검정 실험에서는 미리 하나의 측정 지표를 결정해야 한다. 여러 행동 유형과 관련된 지표들이 수집 대상이 될 수 있지만, 실험이 결국 처리 A와 처리 B 사이의 결정으로 이어질 경우, 단일 지표 또는 검정통계량을 사전에 미리 정해놓아야 한다.

실험을 수행한 뒤 나중에 검정통계량을 선택한다면 연구자 편향이라는 함정에 빠지게 된다.

3.1.2 왜 하필 A/B일까? C,D가 아니라..

A/B 검정은 마케팅 및 전자 상거래 분야에서 널리 사용되고 있지만 그렇다고 이것이 유일한 통계 실험 유형인 것은 아니다. 당연히 추가적인 처리가 포함될 수 있다. 피실험자를 대상으로 반복 추정을 할 수도 있다. 제약회사의 임상 실험과 같이 대상이 매우 귀하고 비용이 비싸며 측정에 많은 시간이 필요할 경우, 실험을 중간에 중단하고 결론을 얻을 수 있는 장치를 마련해두고 실험을 설계한다. 전형적인 의미의 통계적 실험설계는 특정 처리법의 효과에 대한 정적인 질문에 답하는 데 초점을 맞추었다. 데이터 과학자들은 이러한 질문에는 별로 관심이 없다.

- **가격 A와 가격 B의 차이가 통계적으로 유의한가?**

그 보다는 이러한 질문에 더 관심 있다.

- **가능한 여러 가격 중에서 가장 좋은 가격은얼마일까?**

주요개념

- ***연구 대상을 두 가지 이상의 그룹 중 하나에 할당한다. 여기서 서로 다른 처리 조건을 제외한 나머지 조건들은 정확히 동일하게 처리된다.***
- ***이상적으로 대상들은 그룹에 무작위로 배정된다.***

3.2 가설검정

가설검정 혹은 **유의성 검정**은 지금까지 발표된 대부분의 연구 논문에 등장하는 전통적인 통계분석 방법이다. 목적은 관찰된 효과가 우연에 의한 것인지 여부를 알아낸느 것이다.

용어 정리

- **귀무가설 : 우연 때문이라는 가설 (유의어 : 영가설)**
- **대립가설 : 귀무가설과의 대조(증명하고자 하는 가설)**
- **일원검정 : 한 방향으로만 우연히 일어날 확률을 계산하는 가설검정**
- **이원검정 : 양방향으로 우연히 일어날 확률을 계산하는 가설검정**

A/B 검정을 계획할 때, 일반적으로 가설을 염두에 두고 한다. 예를 들면 가격 B가 더 높은 이익을 산출한다는 가설이 있을 수 있다. 왜 굳이 가설을 세워야 할까? 단순히 실험 결과를 보고 더 나은 치료법을 선택하면 안될까?

이에 대한 답은 임의성을 과소평가하려는 인간의 경향에 있다. 이를 잘 보여주는 예는 '블랙스완' 같은 예외적인 사건을 예상하지 못하는 것이다. 또 다른 예로는, 무작위 사건을 어떤 중요한 의미가 있는 패턴을 갖는 것으로 오해하는 경향이 있다. 통계적 가설검정은 연구자가 랜덤하게 우연히 일어난 일에 속지 않도록 보호하기 위한 방법으로 개발되었다.

cf) 랜덤성에 대한 오해

다음 실험에서 랜덤성을 과소평가하려는 인간의 경향을 볼 수 있다. 몇 명의 친구들에게 동전 던지기를 50번 수행하도록 부탁하자. 먼저 그들에게 임의로 앞면(H)과 뒷면(T)을 예측해서 기록하라고 하자. 그런 다음 동전을 실제로 50번 던져 나온 결과를 기록하도록 한다. 그들에게 실제 동전 던진 결과를 임의로 예측한 결과와 구분하여 서로 다른 더미에 놓게 하자.

어떤 결과가 진짜인지 쉽게 알 수 있다. 실제 결과에서는 H 또는 T가 연속적으로 나오는 경우가 더 길게 나타난다.

실제로 동전 뒤집기를 50번 했을 때, H 또는 T가 대여섯번 연속적으로 나오는 것은 그렇게 이상한 일이 아니다. 그러나 대부분의 사람에게 랜덤한 동전 던지기를 예측하라고 하면, H가 3~4번 연속으로 나온 것을 보곤 무작위이기 위해서는 이제 T가 나올차례라고 스스로에게 주문을 걸게 된다.

동전 던지기 예제의 또 다른 측면은, H가 6번 연속으로 나오는 실제 상황을 대할 때(비슷한 예로, 어떤 헤드라인이 다른 것보다 10% 정도 더 나은 결과를 보일 때) 우리가 그것을 뭔가 의미있고 단순한 우연은 아닐 것이다라고 생각하는 경향이 있다는 점이다.

적절하게 설계된 A/B 검정에서는, A와 B 사이의 관찰된 차이가 다음 원인들로 설명될 수 있도록 A와 B에 대한 데이터를 수집한다.

- **유연한 대상 선정**
- **A와 B의 진정한 차이**

통계 가설검정은 그룹 A와 그룹 B 사이에서 보이는 차이가 우연에 의한 것인지를 평가하기 위해 A/B 검정이나 더 나아가 그 외 여러 무작위 실험을 포함하는 분석을 의미한다.

3.2.1 귀무가설

가설검정은 다음과 같은 논리를 사용한다. '실제로는 우연히 일어난 일이지만, 흔하지 않다는 것에 주목하고, 그것이 뭔가 의미가 있고 우연이 아닐 것이다라고 해석하려는 인간의 경향을 감안할 때, 실험에서 얻은 그룹 간의 차이가 랜덤을 통해 얻을 수 있는 합리적인 수준과는 더 극단적으로 달라야 한다는 증거를 보여야 한다' 그룹들이 보이는 결과는 서로 동일하며, 그룹 간의 차이는 우연에 의한 결과라는 것을 기본 가정으로 설정한다. 이 기본 가정을 **귀무가설**이라고 부른다. 결국, 귀무가설이 **틀렸다**라는 것을 입증해서, A그룹과 B그룹 간의 차이가 우연이 아니라는 것을 보여주는 것이 모두의 희망이다.

이를 위한 한 가지 방법은 재표본추출 순열검정을 통한 방법이 있다. 이는 A와 B 그룹의 결과를 서로 섞어서 비슷한 크기의 그룹들을 반복적으로 만든 다음, 관찰된 차이가 각 경우 발생되는 차이들과 비교했을 때 얼마나 극단적인지를 관찰하는 방법이다. 이에 대한 내용은 3.3 절에서 더 자세히 다뤄보고자 한다.

3.2.2 대립가설

가설검정은 그 성격상 귀무가설뿐만 아니라 그와 대립하는 가설을 포함한다. 여기 몇 가지 예가 있다.

1) 귀무가설: 그룹 A와 그룹 B의 평균에는 차이가 없다.

대립가설: A는 B와 다르다 (더 크거나 작을 수 있다.)

2) 귀무가설: $A \leq B$

대립가설: $A > B$

3) 귀무가설: B는 A보다 X% 더 크지 않다.

대립가설: B는 A보다 X% 크다.

결국, 귀무가설과 대립가설이 모든 가능성을 설명할 수 있어야 한다. 귀무가설의 본질은 가설 검정의 구조를 결정한다.

3.2.3 일원/이원 가설검정

A/B 검정을 통해 기존에 기본으로 사용되던 옵션(A라고 하자)과 비교하여 새 옵션(B라고 하자)이 어떠한지 검증한다고 하자. 새로운 옵션이 완벽히 더 나은 것으로 입증되지 않는 이상, 기본 옵션을 계속 사용한다는 게 가정이다. 이 경우에는 B를 선호하는 방향으로 우연에 의해 속지 않도록 가설 검정을 하기 원할 것이다. B가 확실하게 증명되지 않는다면 A를 계속 고수하면 되기에, 우연에 의해 반대로 속는 경우는 없을 것이다. 따라서 우리는 **방향성을 고려한** 대립 가설이 필요하다. (B는 A보다 낫다) 이 경우 **일원**

(또는 한쪽 꼬리) 가설검정을 사용한다. 즉 우연에 의한 극단적인 결과에 대해 한 방향만을 고려하여 p값을 계산한다는 의미이다.

어느 쪽으로도 속지 않도록 가설검정을 원한다면 대립가설은 **양방향** (A는 B와 다르며 더 크거나 더 작을 수 있음)이 된다. 이 경우 **이원**(또는 양쪽 꼬리) 가설을 사용한다. 우연에 의한 극단적인 결과가 양쪽에서 나타날 p값을 계산한다는 것을 의미한다.

새로운 옵션이 더 좋은 것으로 증명되지 않는 한 일반적으로 원래 옵션이 '기본값'으로 지정되는 상황에서는 의사 결정을 필요로 하는 A/B 검정의 특성상 한쪽 꼬리 가설검정과 잘 어울린다. 그러나 R을 포함해서 여러 SW들은 일반적으로 양쪽 꼬리 검정 결과를 기본적으로 제공하며, 많은 통계 전문가도 논쟁이 야기되는 것을 피하기 위해 좀 더 보수적인 양쪽 꼬리 검정을 선택한다. 한쪽 꼬리 대 양쪽 꼬리는 아직 논란이 있는 주제이나, p 값의 정확성이 그리 중요하지 않은 데이터 과학에서는 그렇게 중요하지 않다.

주요 개념

- **귀무가설은 우리가 관찰한 어떤 효과가 특별한 것이 아니고 우연에 의해 발생한 것이라는 개념을 구체화하는 일종의 논리적 구조이다.**
- **가설검정은 귀무가설이 사실이라고 가정하고, '영모형(확률모형)' 을 생성하여 관찰한 효과가 해당 모델로부터 합리적으로 나올 수 있는 결과인지 여부를 검증하는 것이다.**

3.3 재표본추출

통계학에서 **재표본추출** 이란 랜덤한 변동성을 알아보자는 일반적인 목표를 가지고, 관찰된 데이터의 값에서 표본을 반복적으로 추출하는 것을 의미한다. 또한 일부 머신러닝 모델의 정확성을 평가하고 향상시키는 데에도 적용할 수 있다(예를 들면 여러 부트스트랩 데이터 집합을 기반으로 하는 각각의 의사 결정 트리 모델로부터 나온 예측들로부터 **배깅**이라는 절차를 통해 평균 예측값을 구할 수 있다).

용어정리

- *순열검정: 두 개 이상의 표본을 함께 결합하여 관측값들을 무작위로 (또는 전부를) 재표본으로 추출하는 과정을 말한다. (유의어: 임의화검정, 임의순열검정, 정확검정)*
- 복원/미복원 : 표본을 추출할 때, 이미 한번 뽑은 데이터를 다음번 추출을 위해 다시 제자리에 돌려놓거나/다음추출에서 제외하는 표집 방법

3.3.1 순열검정

순열과정에는 두 개 이상의 표본이 관여되며 이들은 통상적으로 A/B 또는 기타 가설검정을 위해 사용되는 그룹들이다. **순서를 바꾼다**라는 의미의 영어 표현은 말 그대로 어떤 값들의 집합에서 값들의 순서를 변경한다는 의미가 있다. 순열검정의 첫 단계는 그룹 A와 그룹 B(더 필요하다면 C,D, ...) 의 결과를 하나로 합치는 것이다. 이것은 그룹들에 적용된 처리의 결과가 다르지 않다는 귀무가설을 논리적으로 구체화한 것이다. 그런 다음이 결합된 집합에서 무작위로 그룹을 뽑아 그 가설을 검정하고 서로 얼마나 다른지 살핀다. 순열 절차는 다음과 같다.

1. **여러 그룹의 결과를 단일 데이터 집합으로 결합한다.**
2. **결합된 데이터를 잘 섞은 후, 그룹 A와 동일한 크기의 표본을 무작위로 (비복원) 추출한다.**
3. **나머지 데이터에서 그룹 B와 동일한 크기의 샘플을 무작위로 (비복원) 추출한다.**
4. **C, D 등의 그룹에 대해서도 동일한 작업을 수행한다.**
5. **원래 샘플 (예를 들면 그룹 비율의 차이)에 대해 구한 통계량 또는 추정치가 무엇이었든 간에 지금 추출한 재표본 에 대해 모두 다시 계산하고 기록한다. 이것으로 한 번의 순열 반복이 진행된다.**
6. **앞선 단계들을 R번 반복하여 검정통계량의 순열 분포를 얻는다.**

이제 실험을 통해 관찰했던 그룹 간의 차이점으로 돌아가서 순열 과정에서 얻은 집합에서의 차이와 비교해보자. 관찰된 차이가 순열로 보이는 차이의 집합 안에 잘 들어 있다면, 우리는 어떤 것도 증명할 수 없다. 즉, 관찰된 차이가 우연히 일어날 수 있는 범위 안에 있다는 말이다. 하지만, 관찰된 차이가 대부분의 순열 분포 바깥에 있다면, 우리는 이것이 우연 때문이 **아니라고** 결론 내릴 수 있다. 전문적인 표현으로, 이 차이는 **통계적으로 유의미하다.**

3.3.2 예제 : 웹 점착성

상대적으로 고가의 서비스를 제공하는 한 회사에서 두 가지 웹 디자인을 두고 어느 쪽이 더 나은 판매 효과를 가져올지를 검증하려고 한다. 판매되는 서비스가 고가이다 보니 판매가 자주 있지 않으며 판매 주기가 상당히 길다. 실제 매출 데이터를 충분히 얻는 데는 너무 오랜 시간이 걸려, 이를 통해 프레젠테이션의 우수성을 검증하기가 어렵다. 이런 이유로 이 회사는 서비스를 상세히 설명하는 내부 페이지의 이용을 대리변수로 사용하여 그 결과를 측정하기로 결정한다.

TIP 대리변수란 참된 관심 변수를 대신하는 변수를 말한다. 이 관심 변수를 직접 얻을 수 없거나, 측정하는 데 많은 비용이나 시간이 소요될 경우 이를 대체하여 사용된다. 예를 들어 기후 연구에서 고대 빙하 중심부의 산소 함량을 당시 온도의 대체제로 사용하고 있다. 관심 가는 진짜 변수에 대한 실제 데이터가 있다면 소량이라고 할지라도 유용하게 사용할 수 있다. 최소한 이를 통해 대리변수가 실제 변수를 대신해서 사용할 만한지 그 상관성이 얼마나 있는지 평가할 수 있다.

이 회사의 잠재적 대리변수 중 하나는 상세한 랜딩 페이지에 대한 클릭 수이다. 더 좋은 방법은 사람들이 페이지에 머문 시간을 측정하는 것이다. 사람들의 관심을 더 오래 끌 수 있는 웹 디자인이 더 많은 매물을 만들거라고 생각하는 것은 합리적이다. 따라서 측정 지표를 페이지A 와 페이지B에서의 평균 세션 시간을 비교하는 것으로 정할 수 있다.

특별한 목적의 내부 페이지이므로 많은 방문객을 받지는 못한다. 또한 세션 시간을 측정하기 위해 많이들 사용하는 구글 웹 로그 분석으로는 사용자가 마지막으로 방문한 세션의 시간을 측정할 수 없다는 점도 고려해야 한다.

구글 웹 로그 분석은 마지막 세션 정보를 삭제하는 대신 이 세션을 -0으로 기록한다. 따라서 해당 세션 정보를 제거하기 위해 추가적인 처리가 필요하다. 결과적으로 두 가지 서로 다른 디자인에 대해 총 36세션, 페이지 A는 21, 페이지 B는 15가 기록됐다. ggplot을 이용한 상자그림을 통해 세션 시간을 시각적으로 비교해보자.

```
ggplot(session_times, aes(x=Page, y=Time)) +
  geom_boxplot()
```

위의 상자그림을 통해 페이지 B가 방문객들을 더 오래 붙잡은 것으로 나타낸다.

각 그룹의 평균은 다음과 같이 확인할 수 있다.

```
> mean_a <- mean(session_times[session_times['Page']=='Page A', 'Time'])
> mean_b <- mean(session_times[session_times['Page']=='Page B', 'Time'])
> mean_b - mean_a
[1] 35.66667
```

페이지 B 는 페이지 A와 비교하여 세션 시간이 평균 35.66초 더 길다. 문제는 이 차이가 우연에 의한 것인지 아니면 통계적으로 중요한 것인지를 판단하는 일이다. 이에 대한 한 가지 대답은 순열검정을 적용하는 것이다. 모든 세션 시간을 결합한 다음, 잘 섞은후 21개의 그룹 (A 페이지의 경우 n=21)과 15개의 그룹 (B 페이지의 경우 n = 15)으로 반복하여 표본을 추출한다.

순열검정을 적용하려면 36개의 세션 시간을 21(페이지 A)와 15(페이지 B)개의 그룹에 랜덤하게 할당하는 기능이 필요하다.

```
perm_fun <- function(x, n1, n2)
{
  n <- n1 + n2
  idx_b <- sample(1:n, n1)
  idx_a <- setdiff(1:n, idx_b)
  mean_diff <- mean(x[idx_b]) - mean(x[idx_a])
  return(mean_diff)
}
```

cf) setdiff 함수: 차집합

이 perm_fun 함수는 비복원추출 방식으로 n2개의 표본을 추출하고 그룹 B에 할당한다. 그리고 나머지 n1개는 그룹 A에 할당한다. 이때 두 평균의 차이를 결과로 반환한다. n2 = 15, n1 = 21로 지정한 후, 이 함수를 1,000번 호출한다. 이렇게 얻은 세션 시간의 차이를 히스토그램으로 표시해보자.

```
perm_diffs <- rep(0, 1000)
for(i in 1:1000)
  perm_diffs[i] = perm_fun(session_times[,'Time'], 21, 15)
par(mar=c(4,4,1,0)+.1)
hist(perm_diffs, xlab='Session time differences (in seconds)', main='')
abline(v = mean_b - mean_a)
```

위의 히스토그램을 통해, 무작위 순열로 구한 평균 세션 시간의 차이가 가끔 실제 관찰된 세션 시간의 차이 (수직선)를 넘어가는 것을 볼 수 있다. 이는 페이지 A와 B 사이의 세션 시간의 차이가 확률분포의 범위 내에 있음을 의미하고, 따라서 차이는 통계적으로 유의하지 않다.

3.3.3 전체 및 부트스트랩 순열검정

앞서 살펴본 랜덤 셔플링 절차를 **임의순열검정** 또는 **임의화검정** 라고 부르며, 이외에도 순열검정에는 두 가지 변종이 있다.

- **전체순열검정**
- **부트스트랩 순열검정**

전체순열검정에서는 데이터를 무작위로 섞고 나누는 대신 실제로 나눌 수 있는 모든 가능한 조합을 찾는다. 이것은 샘플 크기가 비교적 작을 때만 실용적이다. 셔플링을 많이 반복할수록, 임의순열검정 결과는 전체순열검정의 결과와 거의 유사하게 근접한다. 전체순열검정은 영모형이 어떤 유의수준 이상으로 더 '유의미하다'라는 식의 다소 애매한 결론이 아닌 보다 더 정확한 결론을 보장하는 통계적 속성 때문에 **정확검정**이라고도 한다.

부트스트랩 순열검정에서는, 무작위 순열검정의 2단계와 3단계에서 비복원으로 하던 것을 **복원추출**로 수행한다. 이런 식으로 리샘플링 과정에서 모집단에서 개체를 선택할 때 임의성을 보장할 뿐만 아니라, 개체가 처리 그룹에 할당될 때에도 임의성을 보장한다. 두 과정 모두 통계학에서 자주 접하게 된다. 하지만 이들 사이를 구별하는 일은 다소 복잡하고, 데이터 과학의 입장에서는 별로 실용적이지 않다.

3.3.4 순열검정 : 데이터 과학의 최종 결론

순열검정은 랜덤한 변이가 어떤 역활을 하는지 알아보기 위해 사용되는 휴리스틱한 절차이다. 이는 상대적으로 코딩하고, 해석하고, 설명하기 쉽다. 그리고 수식에 기반한 통계학이 빠지기 쉬운 형식주의와 '거짓 결정론'에 대한 유용한 우회로를 제공한다.

수학적 접근과 달리 리샘플링의 장점 중 하나는 추론에서 '모두에게 맞는' 접근 방법을 제공한다는 점이다. 데이턴느 숫자형 또는 이진형일 수 있다. 샘플 크기는 같을 수도 다를 수도 있다.

데이터가 정규분포를 따라야 한다는 가정도 필요없다.

주요 개념

- **순열검정에서는 여러 표본을 결합한 다음 잘 섞는다.**
- **그런 다음 섞인 값들을 이용해 재표본추출 과정을 거쳐, 관심 있는 표본통계량을 계산한다.**
- **이 과정을 반복하고 재표본추출한 통계를 도표화한다.**
- **관측된 통계량을 재표본추출된 분포와 비교하면 샘플 간에 관찰된 차이가 우연에 의한 것인지를 판단할 수 있다.**

3.4 통계적 유의성과 p값

통계적 유의성이란, 통계학자가 자신의 실험 (또는 기존 데이터에 대한 연구) 결과가 우연히 일어난 것인지 아니면 우연히 일어날 수 없는 극단적인 것인지를 판단하는 방법이다. 결과가 우연히 벌어질 수 있는 변동성의 바깥에 존재한다면 우리는 이것을 통계적으로 유의하다고 말한다.

용어정리

- **p값 : 귀무가설을 구체화한 기회 모델이 주어졌을 때, 관측된 결과와 같이 특이하거나 극단적인 결과를 얻을 확률**
- **알파 : 실제 결과가 통계적으로 의미 있는 것으로 간주되기 위해, 우연에 의한 기회 결과가 능가해야 하는 '비정상적인' 가능성의 임계 확률**
- **제 1종 오류 : 우연에 의한 효과가 실제 효과라고 잘못 결론 내리는 것**
- **제 2종 오류 : 실제 효과를 우연에 의한 효과라고 잘못 결론 내리는 것**

결과	가격A	가격B
전환	200	182
전환되지 않음	23539	22406

가격 A는 가격 B에 비해 약 5% 정도 우수한 결과를 보였다. 물량이 큰 사업에서는 충분히 의미가 있는 차이다.

'빅데이터'라고도 볼 수 있는 45,000개 이상의 많은 데이터를 가지고 이 결과를 얻다 보니, 통계적 유의성 검정 (주로 작은 표본에서 표본의 변동성을 설명하기 위해 사용)이 필요 없다고 생각할 지 모르겠다. 하지만 전환율이 너무 낮아, 실제 필요한 표본크기를 결정하는 데 매우 중요한 값 (전환 횟수)은 정작 200개 정도에 불과하다. 재표본추출 절차를 사용하면 가격 A와 B 간의 전환 차이가 우연에 의한 것인지 검정할 수 있다. 여기서 우연에 의한 차이란 곧 두 전환율 사이에 차이가 없다는 귀무가설의 확률 모형을 가지고 생성한 데이터의 랜덤 변이를 의미한다. 다음 순열 절차는 '두 가격이 동일한 전환율을 공유하는지, 이 랜덤 변이가 5%만큼의 차이를 만들어낼 수 있는지'를 묻는 질문에 대한 답을 준다.

1. 모든 표본 정보가 담긴 항아리가 있다고 생각해보자. 그러면 전체 전환율은 45,945개의 0과 382개의 1이므로 0.008246 = 0.8246%라고 할 수 있다.
2. 크기 23,739(가격 A)의 표본을 섞어서 뽑고 그중 1이 몇 개인지 기록하자.
3. 나머지 22,588개(가격 B)에서 1의 수를 기록하자.
4. 1의 비율 차이를 기록하자.
5. 2~4 단계를 반복한다.
6. 이 차이가 얼마나 자주 >=0.0368인가?

3.3.2 절에서 소개한 perm_fun 함수를 다시 사용하여, 무작위로 순열 추출한 전환율 차이에 대한 히스토그램을 그릴 수 있다.

```
obs_pct_diff <- 100*(200/23739 - 182/22588)
conversion <- c(rep(0, 45945), rep(1, 382))
perm_diffs <- rep(0, 1000)
for(i in 1:1000)
  perm_diffs[i] = 100*perm_fun(conversion, 23739, 22588 )
hist(perm_diffs, xlab='Conversion rate (percent)', main='')
abline(v = obs_pct_diff, lty=2, lwd=1.5)
text("  Observed\n  difference", x=obs_pct_diff,  y=par()$usr[4]-20, adj=0)
```

위 그림은 1,000개의 재표본추출 결과를 보여주는 히스토그램이다. 이 경우, 관찰된 0.0368%의 차이는 랜덤 변이의 범위 내에 있다.

3.4.1 p값

그래프를 눈으로 보는 것보다는, **p값**과 같이 통계적 유의성을 정확히 측정하기 위한 지표가 필요하다. 이것은 확률모형이 관측된 결과보다 더 극단적인 결과를 생성하는 빈도라고 할 수 있다. 다시 말해, 순열검정으로 얻은 결과 중에서, 관찰된 차이와 같거나 더 큰 차이를 보이는 경우의 비율로 p값을 추정할 수 있다.

```
> mean(perm_diffs > obs_pct_diff)
[1] 0.341
```

p값은 0.341이다. 즉, 우연히 얻은 결과의 30% 정도가 관찰한 것만큼 극단적이거나 그 이상 극단적인 결과를 얻을 것으로 기대된다.

사실 이 경우에는 p값을 얻기 위해 순열검정을 할 필요가 없다. 가설이 이항분포를 따르기 때문에, 대신 정규분포를 사용하여 p값을 근사할 수도 있다. R에서 prop.test 함수를 사용하여 구할 수 있다.

```
> prop.test(x=c(200,182), n=c(23739,22588), alternative="greater")

    2-sample test for equality of proportions
    with continuity correction

data:  c(200, 182) out of c(23739, 22588)
X-squared = 0.14893, df = 1, p-value =
0.3498
alternative hypothesis: greater
95 percent confidence interval:
 -0.001057439  1.000000000
sample estimates:
     prop 1      prop 2
0.008424955 0.008057376
```

인수 x는 각 그룹의 성공 횟수이고 인수 n은 시행 횟수이다. 정규근사법을 통해, 순열검정에서 얻은 p값과 비슷한 0.3498을 얻었다.

3.4.2 유의수준

통계학자는 어떤 결과가 우연히 발생한 것인지 아니면 '진짜 특별한' 것인지 결정하는 것을 연구원의 재량에 맡기는 관행에 난색을 표한다. 오히려 '우연히 얻은 (귀무가설) 결과의 5%보다 더 극단적인' 결과와 같이 어떤 임계값 (5%) 미리 지정하는 것을 선호한다. 이 임계값을 보통 유의수준(알파) 이라고 한다. 많이 사용되는 유의수준은 5%와 1%이다. 이 값은 임의로 선택된다. 올바른 x% 값을 보장하는 프로세스는

없다. 이는 확률 문제가 '우연히 일어날 확률은 무엇인가?'가 **아닌** '랜덤 모델이 주어졌을 때, 극단적인 결과가 나올 확률은 어느 정도인가?' 이기 때문이다.

즉 랜덤 모델의 적합도에 관해 역으로 추적하는 것이고, 그에 대한 판단은 어떤 확률로 나타나지 않는다. 이런 점에서 많은 혼란을 가져온다.

p값의 의미

최근 몇 년 사이에 p값의 의미를 두고 상당한 논란이 있었다. 한 심리학 저널에서는 그동안 p값만을 가지고 논문 출판을 결정한 것이 저널의 질을 낮추었다는 고민 끝에, 논문을 제출할 때 p값의 사용을 '금지'하기까지 했다. 너무 많은 연구자가 p값이 무엇인지 어렴풋이만 아는 상태로, 논문이 통과될 수 있도록 유의미한 p값이 나올 때까지 온갖 가설검정을 수행한다.

진짜 문제는 사람들이 p값을 통해 실제 의미하는 것보다 더 많은 의미를 찾으려 한다는 것이다. 우리가 p값을 통해 전달하고자 하는 의미는 다음과 같다.

- **결과가 우연에서 비롯될 확률**

우리는 더 낮은 p값을 원하는 결국 뭔가를 증명했다고 결론을 내릴 수 있기를 바란다. 많은 저널 편집자들이 p값을 이런 식으로 해석한 것이다. 그러나 **실제** p값이 나타내는 것은 다음과 같다.

- **랜덤 모델이 주어졌을 때, 그 결과가 관찰된 결과보다 더 극단적일 확률**

이 차이가 미묘해 보이지만, 이게 사실이다. p값이 유의미하다고 해서 그것이 기대처럼 바로 '증거'가 되는 것은 아니다. p값의 진짜 의미를 이해하면 '통계적으로 유의미하다'는 결론에 대한 논리적 뒷받침이 다소 약하다는 것을 알게 된다.

2016년 3월 미국통계협회는 내부 심의를 거쳐, p값의 사용에 대한 경고를 촉구하는 성명서를 통해 p값에 대한 오해들을 밝혔다.

미국통계협회의 성명서는 연구자들과 저널 편집자들에게 아래 6가지 원칙을 강조했다.

1. **p값은 이 데이터가 특정 통계 모델과 얼마나 상반되는지 나타낼 수 있다.**
2. **p값은 연구 가설이 사실일 확률이나, 데이터가 랜덤하게 생성되었을 확률을 측정하는 것이 아니다.**
3. **과학적 결론, 비즈니스나 정책 결정은 p값이 특정 임계값을 통과하는지 여부를 기준으로 해서는 안 된다.**
4. **적절한 추론을 위해서는 완전한 보고와 투명성이 요구된다.**
5. **p값 또는 통계적 유의성은 효과의 크기나 결과의 중요성을 의미하지 않는다.**
6. **p값 그 자체는 모델이나 가설에 대한 증거를 측정하기 위한 좋은 지표가 아니다.**

3.4.3 제 1종과 제 2종 오류

통계적 유의성을 평가할 때는 두 가지 유형의 오류가 발생할 수 있다.

- **1종 오류 : 어떤 효과가 우연히 발생한 것인데, 그것이 사실이라고 잘못 판단하는 경우**
- **2종 오류 : 어떤 효과가 실제로 있는 것인데, 그것이 우연히 발생한 것이라고 잘못 판단하는 경우**

실제로 2종 오류는 어떤 오류라기보다 표본크기가 너무 작아서 효과를 알아낼 수 없다고 판단하는 것과 같다. p값이 통계적 유의성에 미치지 못하는 경우 (예를 들면 5% 초과), 실제 의미는 '효과가 아직 입증되지 않았다'는 뜻이다. 표본크기가 더 클수록 p값이 더 작아진다.

유의성 검정 (가설검정)의 기본 기능은 어쩌다 우연히 일어난 일에 속지 않도록 하는 것이다.

따라서 보통은 1종 오류를 최소화하도록 가설을 설계한다.

3.4.4 데이터 과학과 p값

데이터 과학자들이 하는 일은 일반적으로 과학 저널에 게재하기 위한 일거리가 아니다. p값의 가치에 대한 논쟁은 다소 학문적이다. 데이터 과학자에게 p값은 관심 있고 유용한 모델의 결과가 일반적인 랜덤 변이의 범위 내에 있는지를 알고 싶을 때 유용한 측정 지표이다. p갑을 모든 실험에서 의사 결정을 좌우하는 도구로서 간주해서는 안된다. p값은 어떤 결정에 관련된 정보의 일부일 뿐이다.

예를 들어 p값은 일부 통계 또는 머신러닝 모델에서 중간 입력으로 사용되기도 한다. p값에 따라, 어떤 피처를 모델에 포함하거나 제외하기도 한다.

주요개념

- **유의성 검정은 관찰된 효과가 귀무가설 모형에 대한 무작위 변이의 범위 내에 있는지 결정하는 데 사용된다.**
- **p값은 귀무가설로부터 나올 수 있는 결과가 관찰된 결과만큼 극단적으로 나타날 확률이다.**
- **유의수준(알파), 귀무가설 모델에서 '비정상'이라고 판단할 임계값을 말한다.**
- **유의성 검정은 데이터 과학보다는 좀 더 공식적인 연구 보고와 관련이 있다(그러나 공식적인 연구 보고의 경우에도 최근에는 중요성이 희미해지고 있다).**

3.5 t검정

데이터가 횟수나 측정값을 포함하는지, 표본이 얼마나 큰지, 측정 대상이 무엇인지에 따라 다양한 유형의 유의성 검정 방법이 있다. 가장 자주 사용하는 것은 **t검정**으로, 스튜던트 t분포의 이름을 따서 붙인 것이다. 이것은 원래 윌리엄 고셋이 단일 표본평균의 분포를 근사화하기 위해 개발한 것이다.

용어 정리

- **검정통계량 : 관심의 차이 또는 효과에 대한 측정 지표**
- **t통계량 : 표본화된 형태의 검정통계량**
- **t 분포 : 관측된 t 통계량을 비교할 수 있는 (귀무가설에서 파생된) 기준 분포**

모든 유의성 검정은 관심 있는 효과를 측정하기 위해 **검정통계량**을 지정하고, 관찰된 효과가 정상적인 랜덤 변이의 범위 내에 있는지 여부를 판단하는 데 도움을 준다. 재표본 검정에서 데이터의 척도는 큰 문제가 되지 않는다. 데이터로 부터 기준(귀무가설) 분포를 생성하고 같은 검정통계량을 그대로 사용하면 된다.

통계적 가설검정이 한창 개발 중이던 1920년대와 1930년대에는, 재표본 검정을 위해 무작위로 데이터를 수천 번 섞는다는 것은 거의 불가능했다. 통계학자들은 순열(섞인) 분포에 대한 좋은 근사가 고셋의 t 분포에 기초한 t검정이라는 것을 발견했다. 이는 데이터가 수치형인 아주 일반적인 2표본 비교(A/B 검정)에 주로 사용한다. 그러나 척도에 상관없이 t 분포를 사용하려면, 표준화된 형태의 검정통계량을 사용해야 한다.

고전적인 통계학 참고서들은 고셋의 분포를 통합한 다양한 수식들을 보여주며 데이터를 표준화하여 표준 t분포와 비교하는 방법을 보여준다. 여기에서 이를 다루지는 않겠지만, R이나 파이썬뿐 아니라 모든 통계 SW들은 이미 이 수식들을 구현한 명령을 제공한다. R에서 이 함수는 t.test이다.

```
> t.test(Time ~ Page, data=session_times, alternative='less' )

    Welch Two Sample t-test

data:  Time by Page
t = -1.0983, df = 27.693,
p-value = 0.1408
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 19.59674
```

```
sample estimates:
mean in group Page A
            126.3333
mean in group Page B
            162.0000
```

대안가설은 페이지 A에 대한 평균 세션 시간이 페이지 B보다 작다는 것이다. 이는 순열검정을 통해 얻은 p=0.124과 매우 가깝다.

이렇게 재표본추출을 설명할 때, 데이터가 수치형인지 또는 이진형인지, 표본크기가 균형 잡혀 있는지, 표본분산이 얼마나 큰지 등 다양한 다른 요인에 대해 걱정하지 않고, 관측된 데이터와 검증할 가설만들 가지고 답을 구했다. 수식적으로 접근하는 방법에는 고려해야 할 수 많은 변형들이 존재하며 아주 복잡한 형태가 된다. 통계학자는 그 세계를 탐색하고 지도 보는 법도 배워야 하지만, 데이터 과학자들은 그럴 일 ㅇ벗다. 논문 발표를 준비하는 학자 처럼 가설검정과 신뢰구간 분석을 위한 세부 사항 때문에 진땀 흘리는 일은 하지 않는다.

주요개념

- **컴퓨터가 널리 보급되기 전에, 재표본 검정은 실용적이지 않았으며, 대신 통계학자들은 표준적인 분포를 참고했다.**
- **이렇게 하면 검정통계량이 표준화되어 참고할 분포와 비교할 수 있다.**
- **널리 사용되는 표준화된 통계량 중 하나가 t 통계량이다.**

3.6 다중검정

앞서 언급했듯 통계학에서는 '데이터를 충분히 오래 고문하다 보면 언젠가 뭐든 털어놓을 것이다'라는 말이 있다. 다양한 관점으로 데이터를 보고 충분한 질문을 던지다 보면 거의 항상 통계적으로 유의미한 결과가 나온다.

용어정리

- **제 1종 오류 : 어떤 효과가 통계적으로 유의미하다고 잘못 결론 내린다.**
- **거짓 발견 비율(FDR) : 다중검정에서 1종 오류가 발생하는 비율**
- **p값 조정 : 동일한 데이터에 대해 다중검정을 수행하는 경우에 필요하다.**
- **과대적합(오버피팅) : 잡음까지 피팅**

예를 들어 20개의 예측변수와 1개의 결과변수가 모두 임의로 생성되었다고 하자.

유의수준 0.05에서 20번의 일련 유의성 검정을 수행하면 적어도 하나의 예측변수에서 통계적으로 유의미한 결과를 (실수로) 초래할 가능성이 꽤 있다. 앞에서 설명한 것처럼 이것을 **제 1종 오류**라고 한다. 0.05의 유의수준에서 항상 유의미하지 않는다는 올바른 검정 결과가 나올 확률을 먼저 계산해서, 이를 1에서 빼면 확률을 구할 수 있다.

무의미하다고 정확하게 검정할 확률이 0.95 이므로, 20번 모두 무의미하다라고 올바른 검정 결과를 보일 확률은 $0.95^{20} = 0.36$ 이다. 적어도 하나의 예측값이 유의미하다고 검정 결과가 나올 확률은 이 확률의 나머지, 즉 1 - (**모든 것이 무의미하다는 결론이 나올 확률**) = 0.64이다.

이 문제는 데이터 마이닝에서 '모델이 잡음까지 학습하는' **오버피팅** 문제와 관련이 있다.

추가하는 변수가 많을 수록 또는 더 많은 모델을 사용할수록 뭔가가 우연에 의해 '유의미한' 것으로 나타날 확률이 커진다.

지도학습에서는 이런 위험을 낮추기 위해, 홀드아웃 세트를 사용해서 이전에 보지 못했던 데이터를 통해 모델을 평가한다. 이런 홀드아웃 세트를 사용하지 않는 통계 및 머신러닝 방법은 지속적으로 통계적 잡음에 근거한 위험한 결론을 내리게 된다.

통계학에는 특정 상황에서 이러한 문제를 다루기 위한 몇 가지 방법이 있다. 예를 들어 여러 처리 그룹 간의 결과를 비교하는 경우, 여러 질문을 할 수 있다. 처리 A~C의 경우 다음과 같은 질문을 할 수 있다.

- A와 B가 서로 다른가?
- B와 C가 서로 다른가?
- A와 C가 서로 다른가?

또는 임상 실험의 경우, 여러 단계별로 치료 결과를 볼 수 있다. 각각의 경우에 여러 가지 질문을 하다 보니, 각 질문마다 유연에 속을 기회는 증가한다. 통계학의 수정 절차는 보통 단일 가설검정을 할 때 보다 통계적 유의성에 대한 기준을 더 엄격하게 설정함으로써 이를 보완한다. 이러한 수정 절차는 일반적으로 검정 횟수에 따라 '유의수준을 나누는' 방법이다. 이는 각 검정에 대해 더 작은 알파를, 즉 통계적 유의성에 대해 더 엄격한 잣대를 적용한다. 이러한 절차 중 하나인 본페로니 수정에서는 간단히 알파를 관측수 n으로 나눈다.

그러나 다중검정 문제는, 이렇게 잘 구조화된 경우 말고 데이터를 고문한다는 말이 나올 정도로 반복적으로 데이터를 '샅샅이 훑는' 현상과 관련이 있다. 달리 말하면, 충분히 복잡한 데이터가 주어졌을 때 흥미로운 것을 발견하지 못했다면 그저 오랫동안 열심히 들여다보지 않은 탓이다. 여느 때보다 더 많은 데이터가 사용 가능하다. 2002년에서 2010년 사이에 출판된 저널 논문 수가 거의 두 배로 증가했다. 이로 인해 다음과 같은 중복 문제를 포함하여 데이터에서 흥미로운 것을 발견할 수 있는 기회가 더욱 많아졌다.

- 여러 그룹 간에 쌍별 차이를 조사하는 것
- 여러 부분군에서의 결과를 알아보는 것(예를 들면 '전반적으로는 아무런 유의미한 결과를 찾을 수 없었다 하지만 30세 미만 미혼 여성들에서는 어떤 효과를 발견했다.')
- 여러 가지 통계 모형을 적용
- 모델에서 많은 변수들을 사용하는 것
- 수많은 서로 다른 질문들(즉, 서로 다른 가능한 결과들)을 묻는 것

NOTE 거짓 발견 비율

거짓 발견 비율(FDR)은 원래 주어진 여러 개의 가설검정을 가운데 하나가 유의미한 효과가 있다고 잘못 판단하는 비율을 나타내는 데 사용되었다. 게넘 연구의 동장과 함께 엄청난 수의 통계적 검정이 유전자 시퀀싱 프로젝트에서 중요한 역활을 하면서 특히 유용해졌다. 이때 이 용어는 검정 프로토콜에 해당하며, 하나의 거짓 '발견'이 한 가설검정의 결과(예를 들어 두 표본 사이의)일때, 사용하는 용어였다. 연구원들은 특정 수준에서 FDR을 제어하기 위해 검정 단계에서 변수를 설정했다. 한편 이 용어는 데이터 마이닝 쪽에서도 분류 문제를 다루는 데 사용된다. 이때의 거짓 발견은 한 레코드를 잘못 라벨링(분류) 한 것(특히 0을 1로 잘못 분류한 것)을 의미한다.

'중복도' 같은 같은 일반적인 문제를 포함하여 여러 가지 이유로, 더 많은 연구가 반드시 더 나은 연구를 의미하는 것은 아니다. 예를 들어 바이엘 제약 회사는 2011년 67개의 과학 연구를 재현하려 시도했으나, 그중 14개만 완전히 재현할 수 있다는 사실을 발견했다. 거의 2/3를 전혀 재현할 수 없었다.

어쨌든 정의가 분명하고 이미 잘 구조화된 통계 검정을 위한 수정 절차는, 데이터 과학자들이 일반적으로 사용하기에는 너무 특정한 경우를 위한 것이어서 문제에 맞게 변경하기가 어렵다.

중복에 대한 데이터 과학자의 결론은 다음과 같다.

- **예측 모델링의 경우, 교차타당성검사와 홀드아웃 표본 사용을 통해, 실제 우연히 발생한 것을 겉보기에 유효한 것처럼 보이도록 잘못된 모델을 만들 위험을 낮춘다.**
- **미리 분류되어 있는 홀드아웃 표본이 없는 다른 절차의 경우, 다음 사항에 의존해야 한다.**
    - **데이터를 더 여러번 사용하고 조작할수록 우연이 더 큰 역활을 할 수 있다는 것을 인식해야 한다.**
    - **재표본추출과 시뮬레이션 결과들을 사용하여 무작위 모델의 기준값을 만들어 관찰된 결과를 비교한다.**

주요개념

- **연구 조사나 데이터 마이닝 프로젝트에서 다중성(다중 비교, 많은 변수, 많은 모델 등) 은 일부가 우연히 유의미하다는 결론을 내릴 위험을 증가시킨다.**
- **여러 통계 비교(즉, 여러 유의성 검정)와 관련된 상황의 경우 통계적 수정 절차가 필요하다.**
- **데이터 마이닝에서, 라벨이 지정된 결과변수가 있는 (즉 분류 결과를 알고 있는) 홀드아웃 표본을 사용하면 잘못된 결과를 피할 수 있다.**

3.7 자유도

많은 통계 검정 관련 자료에서 '자유도'에 대한 설명을 볼 수 있다. 이 개념은 표본 데이터에서 계산된 통계량에 적용되며 변화가 가능한 값들의 개수를 나타낸다. 예를 들면 10개의 값으로 이뤄진 표본에서 평균과 9개 값을 알고 있다면, 마지막 10번째 값을 자연히 알 수 있다. 이 나머지 한 개의 값을 제외한 9개의 값만 변화가 가능하다.

용어 정리

- **표본크기 n : 해당 데이터에서 관측값의 개수(행 혹은 기록값의 개수와 같은 의미)**
- **d.f. (degrees of freedom) : 자유도**

자유도는 많은 통계 검정에서 입력으로 주어지는 값이다. 예를 들면 분산과 표준편차에 대한 계산에서 분모에 표시된 n-1를 자유도라고 부른다. 이것이 왜 이렇게 중요할까? 표본을 통해 모집단의 분산을 추정하고자 할 때 분모에 n을 사용하면 추정치가 살짝 아래쪽으로 편향될 것이다. 분모에 n-1을 사용하면 추정값에 편향이 발생하지 않는다.

전형적인 통계 수업이나 교제의 대부분은 다양한 표준 가설검정 방법(t검정, f검정 등)을 설명하는 데 더 많은 부분을 할애한다. 표본통계량이 전통적인 통계 공식에 맞게 표준화된 경우, 자유도는 표준화된 데이터가 그에 적합한 기준 분포(t분포, f분포 등)에 맞도록 하기 위한 표준화 계산의 일부이다.

과연 이것이 데이터 과학에서도 중요할까? 적어도 유의성 검정이란 측면에선 그렇지 않다.

첫째, 공식적인 통계 검정은 데이터 과학 분야에서 아주 드물게 사용된다. 다른 하나는 데이터 크기가 대게 충분히 크기 때문에, 분모가 n인지 n-1인지가 데이터 과학자들에게는 거의 차이가 없다.

그러나 관련성이 있는 영역이 하나있다. 회귀에서 요인변수를 사용할 때다(로지스틱 회귀 포함). 완전히 불필요한 예측변수들이 있는 경우 회귀 알고리즘을 사용하기 어렵다. 이것은 범주형 변수를 이진 지표(더미)로 요인화 할 때 가장 많이 일어난다. 요일을 한번 생각해보자. 일주일에 7일이 있지만 요일을 지정할 때 자유도는 6개이다. 예를 들면 월요일부터 토요일이 아닌 요일이라고 한다면 그날은 반드시 일요일이어야 한다. 따라서 월~토 지표를 포함하면서 동시에 일요일까지 포함한다면 **다중공선성**오차로 인해 회귀를 실패하게 된다.

주요개념

- **자유도 (d.f.)는 검정통계량을 표준화하는 계산의 일부이며, 이를 통해 기준 분포(t분포, f분포 등)와 비교할 수 있다.**
- **자유도 개념은 회귀를 할때 (다중공선성을 피하기 위해) 범주형 변수들을 n-1지표 혹은 더미 변수로 요인화하는 것의 이유가 된다.**

3.8 분산분석

A/B 검정 말고, 이제 여러 그룹, 예를 들어 A-B-C-D의 수치 데이터들을 서로 비교한다고 가정해보자. 여러 그룹 간의 통계적으로 유의미한 차이를 검정하는 통계적 절차를 분산분석(analysis of variance)줄여서 **ANOVA**라고 한다.

용어 정리

- **쌍별 비교 : 여러 그룹 중 두 그룹 간의 (예를 들면 평균에 대한) 가설검정**
- **총괄검정 : 여러 그룹 평균들의 전체 분산에 관한 단일 가설검정**
- **분산분해 : 구성 요소 분리, 예를 들면 전체 평균, 처리 평균, 잔차 오차로부터 개별 값들에 대한 기여를 뜻한다.**
- **F 통계량 : 그룹 평균 간의 차이가 랜덤 모델에서 예상되는 것보다 벗어나는 정도를 측정하는 표준화된 통계량**
- **SS : 어떤 평균으로부터의 편차들에 대한 제곱합**

다음은 4개의 웹 페이지의 점착성, 즉 방문자가 페이지에서 보낸 시간을 초 단위로 보여준다. 네 페이지는 무작위로 전환되며 각 웹 방문자는 무작위로 그중 한 곳에 접속된다. 각 페이지에는 총 5명의 방문자가 있으며 다음표의 각 열은 독립적인 데이터 집합이다. 페이지 1의 첫 번째 뷰어는 페이지 2의 첫 번째 뷰어와 아무 관련이 없다. 이와 같은 웹 테스트에서는 일부 방문자를 어떤 커다란 모집단에서 무작위로 선택하는 식의, 전통적인 랜덤표본추출 디자인을 원전히 구현할 수 없다. 우리가 선택하는 것이 아닌, 방문자가 오는 대로 바로 대상이 된다. 방문자는 시간대, 요일, 계절, 인터넷 환경, 사용하는 장치 등에 따라 체계적으로 다를 수 있다. 실험 결과를 검토할 때 이러한 요소들을 잠재적 편향의 요인으로 고려해야 한다.

자, 이제 한 가지 어려운 문제가 남았다. 단지 두 그룹을 비교하는 것이라면 문제는 단순할 것이다. 이 표로는 각 그룹 평균의 차이만 알 수 있다. 4개 평균에 대해서 다음과 같이 그룹 간에 6가지 비교가 가능하다.

- 1페이지와 2페이지 비교
- 1페이지와 3페이지 비교
- 1페이지와 4페이지 비교
- 2페이지와 3페이지 비교
- 2페이지와 4페이지 비교
- 3페이지와 4페이지 비교

  그림 3-8 네 그룹의 상자그림은 이들 사이의 상당한 차이가 있음을 보여준다.

우리가 이렇게 **한 쌍씩** 비교하는 횟수가 증가할수록 우연히 일어난 일에 속을 가능성이 커진다. 개별 페이지 간의 가능한 모든 비교에 대해 걱정하는 대신, '모든 페이지가 동일한 기본적인 점착성을 갖는가? 그리고 이들 사이의 차이는 우연에 의한 것이고 원래 4개의 페이지에 할당된 세션 시간 역시 무작위로 할당된 것인가? ' 라는 질문을 다루는 전체적인 **총괄검정**을 할 수 있다.

ANOVA가 바로 이 검정에 사용되는 방법이다. 앞의 웹 페이지 점착성을 예로 들어, ANOVA의 토대가 되는 재표본추출 과정을 살펴보자.

1. **모든 데이터를 한 상자에 모은다.**
2. **5개의 값을 갖는 4개의 재표본을 섞어서 추출한다.**
3. **각 그룹의 평균을 기록한다.**
4. **네 그룹 평균 사이의 분산을 기록한다.**
5. **2~4 단계를 여러 번 반복한다. (예를 들면 1,000번)**

재표집된 분산이 관찰된 변화를 초과한 시간은 어느 정도일까? 이것이 바로 p값이다.

이런 형태의 순열검정은 3.3.1절에서 사용한 것보다 조금 더 복잡하다. 다행히도 **lmPerm** 패키지의 **aovp**함수는 이런 경우에 대한 순열검정 계산을 지원한다.

```
> summary(aovp(Time ~ Page, data=four_sessions))
[1] "Settings:  unique SS "
Component 1 :
            Df R Sum Sq R Mean Sq Iter Pr(Prob)
Page         3    831.4    277.13 4790  0.08142 .
Residuals   16   1618.4    101.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Pr(Prob) 값이 바로 p값으로 결과는 0.08142이다. Iter 열은 순열검정에서 수행한 반복 횟수를 나타낸다. 다른 열은 전통적인 ANOVA 테이블에 대응하며, 이에 대해 이어서 자세히 설명한다.

3.8.1 F통계량

두 그룹의 평균을 비교하기 위해 순열검정 대신 t 검정을 사용할 수 있는 것 처럼, **F 통계량**을 기반으로 한 ANOVA 통계 검정도 있다. F통계량은 잔차 오차로 인한 분산과 그룹 평균(처리 효과)의 분산에 대한 비율을 기초로 한다. 이 비율이 높을 수록 통계적으로 유의미 하다고 할 수 있다.

데이터가 정규분포를 따를 경우, 통계 이론에 따르면 해당 통계량은 특정 분포를 따르게 되어 있다.

이를 토대로 p값을 계산할 수 있다.

R의 aov 함수를 통해 **ANOVA 테이블**을 손쉽게 계산할 수 있다.

```
> summary(aov(Time ~ Page, data=four_sessions))
            Df Sum Sq Mean Sq
Page         3  831.4   277.1
Residuals   16 1618.4   101.2
            F value Pr(>F)
Page           2.74 0.0776 .
Residuals
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01
  ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Df는 자유도, Sum Sq는 제곱합, Mean Sq는 평균 제곱(평균 제곱 편차를 줄여서), F value는 F 통계량을 가리킨다. 총평균의 경우, 제곱합은 0에서부터 총평균까지의 거리를 구하고, 제곱한 다음 20(관측 수)을 곱한 값과 같다.

따라서 총평균에 대한 자유도는 정의에 따라, 1이 된다. 처리 방법에 대한 평균의 자유도는 3이다.

(3개의 평균과 함께 총평균이 정해지면 나머지 평균은 달라질 수 없다)

처리 평균에 대한 제곱합은 각 처리 평균과 총평균 사이의 편차를 제곱한 값들의 합이다. 잔차의 경우 자유도는 20(모든 관측치가 다를 수 있음)이며 SS는 개별 관측치와 처리 평균의 차에 대한 제곱합이다.

평균 제곱(MS)은 제곱합을 자유도로 나눈 값이다. F통계량은 MS(처리)/MS(오차)이다. F값은 이 비율에 따라 결정되며, 표준 F분포와 비교하여 그룹 평균 간의 차이가 랜덤 변이에서 예상되는 것보다 크 지 여부를 결정할 수 있다.

NOTE_ 분산분해

데이터에서 관측된 값들은 다른 구성 요소의 합으로 생각할 수 있다. 데이터 내의 관측값의 경우, 평균, 처리 효과, 잔차 오차로 분류할 수 있다. 이것을 '분산의 분해'라고 한다.

1. **총평균(웹 페이지 점착성 데이터의 경우 173.75)으로부터 시작한다.**

2. **음수인 처리 효과를 추가한다.**
3. **음수일 수 있는 잔차 오차를 더한다.**

이에 따르면, A-B-C-D 검정을 위한 표의 왼쪽 상단 값에 대한 분산분해는 다음과 같다.

1. **총평균에서 시작 : 173.75**
2. **처리 (그룹) 효과를 추가 : -1.75(172-173.75)**
3. **잔차 오차를 추가 : -8 (164-172)**
4. **최종값 : 164**

3.8.2 이원 분산분석

방금 설명한 A-B-C-D 검정은 변하는 요소(그룹)가 하나인 '일원'ANOVA이다. 그런데 '주말 대 평일'이라는 두 번째 요소를 고려한 각 조합(그룹 A 주말, 그룹 A 평일, 그룹 B 주말 등) 에 관한 데이터가 있다고 가정하자. 이럴 때 필요한 것이 '이원' ANOVA이다. 이는 '상호작용 효과'를 확인하는 식으로, 일원 ANOVA와 방식은 비슷하다. 총평균 효과와 처리 효과를 확인한 후, 각 그룹의 주말과 평일 데이터를 따로 분리한다. 그리고 그 부분집합들에 대한 평균과 처리 평균 사이의 차이를 찾아본다.

여러 요인과 그 효과를 모델링할 수 있는 회귀와 로지스틱 회귀 같은 완전한 통계 모델을 위한 첫걸음이 바로 이 ANOVA, 그리고 이원 ANOVA라고 할 수 있다.

주요개념

- **ANOVA는 여러 그룹의 실험 결과를 분석하기 위한 통계적 절차이다.**
- **A/B 검정과 비슷한 절차를 확장하여 그룹 간 전체적인 편차가 우연히 발생할 수 있는 범위 내에 있는지를 평가하기 위해 사용된다.**
- **ANOVA의 결과 중 유용한 점 중 하나는 그룹 처리, 상호작용 효과, 오차와 관련된 분산의 구성 요소들을 구분하는 데 있다.**

3.9 카이제곱검정

웹 테스트 시, 종종 단순한 A/B 검정을 넘어 동시에 여러 가지 처리를 한 번에 테스트할 필요가 있다. **카이제곱검정**은 횟수 관련 데이터에 주로 사용되며 예상되는 분포에 얼마나 잘 맞는지를 검정한다. 통계적 관행에서 **카이제곱통계량**은 일반적으로 변수 간 독립성에 대한 귀무가설이 타당한지를 평가하기 위해 r x c 분할표를 함께 사용한다.

카이제곱검정은 원래 칼 피어슨에 의해 1900년에 처음 개발됐다.

'카이' 라는 용어는 피어슨이 논문에서 사용한 그리스 문자 χ로부터 유래한다.

용어 정리

- ***카이제곱통계량: 기댓값으로부터 어떤 관찰값까지의 거리를 나타내는 측정치***
- ***기댓값 : 어떤 가정(보통 귀무가설)으로부터 데이터가 발생할 때,그에 대해 기대하는 정도***
- ***d.f. degees of freedom : 자유도***

NOTE_ rxc는 각각 '행과 열'을 의미한다. 즉 2 x 3 은 2행 3열을 의미한다.

3.9.1 카이제곱검정 : 재표본추출 방법

A,B,C 세 가지 헤드라인을 비교한다고 가정하자. 이때 각각 1,000명의 방문자에 관한 결과는 다음과 같다.

결과적으로 이들 셋의 효과가 확실히 다른 것처럼 보인다. 실제 수는 적지만, A는 B에 비해 거의 두 배의 클릭을 유도했다. 재표본추출을 통해, 재표본추출을 통해, 클릭율이 우연히 발생할 수 있는 것보다 유의미한 정도로 큰 것인지를 검정할 수 있다. 이 검정을 하려면 클릭의 '기대' 분포가 필요하며, 이 경우 각 헤드라인 모두가 동일한 클릭율을 갖는다는 가정이 귀무가설에 속한다. 전체 클릭율은 34/3,000이다. 이 가정하에 분할 표는 다음과 같다,

피어슨 잔차는 다음과 같이 정의된다.

$$R = \frac{\text{관측값} - \text{기댓값}}{\sqrt{\text{기댓값}}}$$

R은 실제 횟수와 기대한 횟수 사이의 차이를 나타낸다.

카이제곱통계량은 바로 이 피어슨 잔차들의 제곱합이다.

$$\chi^2 = \sum_{i}^{r} \sum_{j}^{c} R^2$$

여기에서 r과 c는 각각 행과 열의 수를 의미한다. 따라서 이 경우 카이제곱통계량은 1.666이 된다. 과연 이 값이 귀무가설로부터 얻을 수 있는 값보다 크다고 할 수 있을까?

재표본추출 알고리즘으로 이를 검정할 수 있다.

1. 34개의 1(클릭한 경우)과 2,966개의 0(클릭하지 않은 경우) 이 들어 있는 상자를 만들자.
2. 상자의 내용물을 잘 섞은 다음, 1,000개의 표본을 세 번씩 가져와서 각각의 클릭 수를 계산한다.
3. 이렇게 얻은 횟수와 기대한 횟수의 차이를 제곱해서 합산한다.
4. 2~3단계를 1,000번 반복한다.
5. 재표본추출을 통해 얻은 편차의 제곱합이 얼마나 자주 관측값을 초과하는가? 이것이 바로 p값이다.

chisq.test 함수를 통해 이 값을 계산할 수 있다. 그 결과는 다음과 같다.

```
> ## Chi square test
> clicks <- matrix(click_rate$Rate, nrow=3, ncol=2, byrow=TRUE)
> dimnames(clicks) <- list(unique(click_rate$Headline),
unique(click_rate$Click))
> chisq.test(clicks, simulate.p.value=TRUE)

    Pearson's Chi-squared test
    with simulated p-value (based
    on 2000 replicates)

data:  clicks
X-squared = 1.6659, df = NA,
p-value = 0.4878
```

검정 결과는 관찰된 결과가 귀무가설(랜덤)로부터 얼마든지 얻을 수 있는 결과임을 보여준다.

3.9.2 카이제곱검정 : 통계적 이론

점근적 통계 이론은 카이제곱통계량의 분포가 **카이제곱분포**로 근사화될 수 있음을 보여준다. 적절한 표준 카이제곱분포는 **자유도**에 의해 결정된다. 분할표에서 자유도는 다음과 같이 행(r)과 열(c)의 수와 관련이 있다.

자유도 = $(r-1) \times (c-1)$

카이제곱분포는 일반적으로 한쪽으로 기울어져 있고 오른쪽으로 긴 꼬리가 있다. 자유도가 1, 2, 5, 10인 경우의 분포는 다음과 같으니 참고하자.관찰된 통계량이 카이제곱분포의 바깥쪽에 위치할수록 p 값은 낮아진다.

chisp.test함수를 이용해 카이제곱분포에 대한 p값을 계산할 수 있다.

```
> chisq.test(clicks, simulate.p.value=FALSE)

    Pearson's Chi-squared test

data:  clicks
X-squared = 1.6659, df = 2,
p-value = 0.4348
```

p값을 보면, 재표본추출해서 얻은 p값 보다 약간 작다. 이는 카이제곱분포가 실제 통계 분포가 아닌 근사적이기 때문이다.

3.9.3 피셔의 정확검정

이렇게 카이제곱분포는 재표본 검정의 좋은 근사치를 제공한다. 사건 발생 횟수가 매우 낮을 때(한 자리 숫자이거나, 특히 5개 이하인 경우)는 예외이지만, 이런 예외적인 경우에도 재표본추출 방법을 통해 더 정확한 p값을 얻을 수 있다. 실제로 대부분의 통계 SW는 발생 할 수 있는 **모든** 조합(순열)을 실제로 열거하고, 빈도를 집계하고, 관찰된 결과가 얼마나 극단적으로 발생할 수 있는지를 정확하게 결정하는 절차를 제공한다. 이를 위대한 통계학자 피셔의 이름을 붙여, 이를 **피셔의 정확검정**이라고 한다. 피셔의 정확검정을 위한 R코드는 다음과 같이 아주 간단하다.

```
> fisher.test(clicks)

    Fisher's Exact Test for Count
    Data

data:  clicks
p-value = 0.4824
alternative hypothesis: two.sided
```

이렇게 얻은 p값은 재표본추출 방법을 사용하여 얻은 p값 0.4853과 아주 가깝다 .

일부 값이 매우 낮고 다른 값이 상대적으로 매우 높은 경우(예를 들면 전환율 계산에서 분모) 모든 가능한 순열을 계산하기는 어렵기 때문에, 완전한 정확검정 대신, 순열검정을 수행해야 할 수 있다. 앞서 사용한 R 함수에는 이 근사 방법을 사용할지 여부(simulate.p.value를 TRUE OR FALSE로 지정), 반복할 횟수(B), 그리고 정확한 결과에 대한 계산을 얼마나 수행할지를 제한하는 계산 제약 조건(workspace) 등 몇가지 인수들을 제공한다.

cf) 과학 사기를 밝히다.

미국 1991년에 데이터 조작사건이 있었다. 이 사건의 중요한 요소중 하나는 실험실에서 관찰한 데이터의 숫자 분포에 관한 통계적 증거에 있었다. 수사관은 **균등한 확률분포**를 따를 것으로 기대되는 **중간** 자리의 숫자들에 초점을 맞췄다. 즉, 중간 숫자들은 동일한 확률로 무작위로 나타나야 했다(첫 자리는 주로 하나의 값이었고, 마지막 자리는 반올림에 의해 영향을 받으므로 제외). 다음 표에는 실제 데이터의 내부 숫자의 도수가 나와 있다.

다음 그림은 이렇게 수집한 315개의 숫자 분포를 보여준다. 분명히 무작위가 아닌 것처럼 보인다.

수사관들은 기댓값(31.5, 정확히 균일한 분포에서 각 숫자가 뽑힐 때)과의 차이를 계산하고, 카이제곱검정을 사용해(재표본추출 절차를 사용해도 똑같았을 것이다.) 실제 분포가 정상적인 랜덤 변이의 범위를 훨씬 넘는다는 것을 보였다.

3.9.4 데이터 과학과의 관련성

카이제곱검정이나 피셔의 정확검정은 데이터 과학과의 직접적인 연관성을 찾기가 참 어렵다.

A-B 나 A-B-C나 상관없이 대부분 실험에서의 목표는 단순히 통계적 유의성을 조사하는 것이 아닌 최적의 처리 방법을 찾는 것이다. 이를 위해서는 멀티암드 밴딧 방법이 더 정확한 해결책이라고 할 수 있다.

데이터 과학에서 카이제곱검정, 특히 피셔의 정확검정을 활용하는 대표적인 예로, 웹 실험에 적합한 표본크기를 판별하는 일을 들 수 있다. 이러한 실험은 종종 클릭률이 매우 낮기에 수천 번의 실험에도 불구하고 집계 비율이 너무 낮아 실험을 통해 확실한 결론을 내리기 어렵다. 이러한 경우 피셔의 정확검정, 카이제곱검정, 그리고 기타 검정은 검정력이나 표본크기를 계산하는 데 유용할 수 있다.

카이제곱검정은 출판을 하기 위해 통계적으로 유의미한 p값을 찾는 논문 연구에서 널리 사용된다. 데이터 과학 응용 분야에서는 카이제곱검정이나 이와 유사한 재표본추출 시뮬레이션을 필터로서 더 많이 사용한다. 즉 어떤 효과나 특징에 대해 기본적인 유의성 검정을 넘어 더 심층적인 분석이 필요할지 여부를 결정한다. 예를 들면 공간 통계학에서 공간 데이터가 어떤 특징영분포를 따르는지 여부(랜덤일 경우보다 특정 영역에 범죄가 집중되고 있는가?) 를 결정하는 데 사용된다. 또한 머신러닝에서는 자동으로 특징을 선택하기 위해 사용한다. 특징에 따라 클래스의 분포가 어떠한지 조사하고, 특정 클래스의 분포가 랜덤 변이에 비해 비정상적으로 크거나 작은 특징을 알아내는 등에 사용한다.

주요 개념

- **통계학에서 흔한 절차는 관측된 데이터가 독립성 가정(예를 들면 특정 항목을 구매하려는 성향이 성별과 무관함)을 따르는지 검증하는 것이다.**
- **카이제곱분포는 카이제곱통계량을 비교할 기준 분포(독립성 가정 포함)이다.**

3.10 멀티암드 밴딧 알고리즘

멀티암드 밴딧 (MAB) 알고리즘은 실험설계에 대한 전통적인 통계적 접근 방식보다 명시적인 최적화와 좀 더 빠른 의사 결정을 가능하게 하며,여러 테스트, 특히 웹 테스트를 위해 이를 사용한다.

용어정리

- **멀티암드 밴딧 : 고객이 선택할 수 있는 손잡이가 여러 개인 가상의 슬롯머신을 말하며, 각 손잡이는 각기 다른 수익을 가져다준다. 다중 처리 실험에 대한 비유라고 생각할 수 있다.**
- **손잡이 : 실험에서 어떤 하나의 처리를 말한다. (예를 들면 '웹 테스트에서 헤드라인 A')**
- **상금(수익) : 슬롯머신으로 딴 상금에 대한 실험적 비유(예를 들면 '고객들의 링크 클릭 수')**

전통적인 A/B 검정은 특정하게 설계된 실험을 통해 수집된 데이터를 이용하여 '처리 A나 처리 B 둘 중 어느 쪽이 더 좋은가?' 와 같이 정해진 질문에 대한 답을 준다. 일단 답을 얻고 나면 실험은 멈추고 결과에 따라 행동한다.

이러한 접근에 몇 가지 어려움을 느낄 수도 있다.

첫째, 결론을 내리기 어려울 수 있다. '입증되지 않은 효과', 즉 실험 결과를 통해 효과가 있다는 것을 유추할 수는 있지만, 효과가 있더라도 그것을 입증할 만한(전통적인 통계 표준을 만족시킬 만한) 크기의 표본이 없을 수 있다. 어떤 결론을 내릴 수 있을까?

둘째, 우리는 실험이 끝나기 전에 이미 얻은 결과들을 이용하기 시작할 수도 있다.

셋째, 마음을 바꿔서 실험이 끝난 후에 추가적으로 들어오는 데이터를 기반으로 다른 것을 시도하고 싶을 수 있다. 실험과 가설검정에 대한 전통적인 방법들은 1920년대에 시작된 것으로, 다소 유연하지 않다. 컴퓨터 성능과 SW의 출현으로 더 강력하고 유연한 접근 방식이 가능해졌다. 게다가 데이터 과학, 그리고 비즈니스 전반에서는 통계적 유의성보다는 제반 비용과 결과를 최적화하는 데 더 관심이 있다.

웹 테스트에서 널리 사용되는 밴딧 알고리즘을 사용하면 한 번에 여러 가지 처리를 테스트하고 기존의 통계 설계보다 빠르게 결론을 얻을 수 있다. 이 알고리즘은 도박에서 사용되는 슬롯머신을 지칭하는 속어에서 이름을 가져왔다. 이 알고리즘은 도박에서 사용되는 슬롯머신을 지칭하는 속어에서 이름을 가져왔다. 슬롯머신을 속칭 팔 하나인 강도라고 부르기 때문이다(도박꾼들이 지속적으로 조금씩 돈을 잃게 되는 구조이므로). 둘 이상의 손잡이가 달려 있고 각 손잡이는 다른 속도로 돈을 지불하는 솔롯머신을 상상해보자. 그것이 이 알고리즘의 정식 이름(팔 여러 개 달린 강도)이라고 할 수 있다.

우리의 목표는 가능한 많은 돈을얻는 것이고, 더 구체적으로 말하면 많은 상금이 나오는 손잡이를 나중에 확인하는 것이 아닌 빨리 확인하는 것이다. 어려운 점은 손잡이를 잡아당길 때 얼마를 지불할지 모른다는 것이다. 손잡이를 당겼을 때의 결과만 알 수 있다. 어떤 '손잡이' 든지 간에 '상금'이 모두 같은 금액이라고 가정하자. 다른 점은 승리할 확률이다. 처음에 각 손잡이 마다 50번 시도한 후 다음과 같은 결과를 얻었다고 가정하자.

- 손잡이 A : 50번 중 10번 승리
- 손잡이 B : 50번 중 2번 승리
- 손잡이 C: 50번 중 4번 승리

그러면 단순히 다음과 같은 극단적인 결론을 내릴 수 있을 것이다. '손잡이 A가 최고인 것으로 보인다. 다른 손잡이는 시도하지 말고 A만 당기자' 이것은 초기 시험에서 얻은 정보를 최대한 활용하는 방법이다. A가 정말로 우월하다면, 우리는 그 이익을 초기에 얻게 된다. 하지만 사실은 B나 C가 더 좋다면 우리는 이 사실을 발견할 기회를 놓치게 된다. 또 다른 극단적인 접근법은 '모두가 무작위 인것으로 보인다. 모두 똑같이 잡아당기자'이다. 이것은 A외의 다른 것들의 확률을 알 수 있는 최대한의 기회를 제공한다. 그러나 그 과정에서 우리는 어쩔 수 없이 수익이 낮을 것으로 예상되는 행위를 자주 시도해야 한다. 이것을 얼마나 지속해야 할까? 밴딧 알고리즘은 하이브리드 접근 방식을 취한다. A의 우위를 활용하기 위해 A를 더 자주 잡아당기는 것으로 시작하긴 하지만 그렇다고 B와 C를 포기하지는 않는다. A에서 계속해서 성과를 거둔다면, B와 C를 당길 기회를 A에게 더 줘서 A를 더 자주 잡아당긴다. 반면 C가 더 좋아지고 A가 더 나빠지기 시작하면 A로 가던 기회를 C에게 돌린다. 그중 하나가 A보다 우수하고 이것이 초기 실험에서 우연히 감춰졌던 결과라면, 이제는 더 많은 테스트를 통해 이 사실이 밝혀질 수 있는 기회가 생기게 된다.

이제 이것을 웹 테스트에 적용하는 방법을 생각해보자. 여러 개의 슬롯머신 손잡이 대신에 웹 사이트에서 여러 가지 제안, 헤드라인, 색상 등을 테스트 할 수 있다. 고객은 클릭(상품 판매자 입장에서는 '승리') 하거나 클릭하지 않을 것이다. 처음에는 여러 제안이 무작위로 균등하게 표시된다. 그러다가 한 제안이 다른 제안 보다 좋은 결과를 내기 시작하면 더 자주 표시('잡아당기기 ') 될 수 있게 한다.

그러나 잡아당기는 비율을 수정하는 알고리즘을 위한 파라미터는 무엇이 되어야 할까? 잡아당기는 비율을 언제 어떻게 수정해야 할까?

엡실론-그리디 알고리즘이라는 A/B검정을 위한 간단한 알고리즘은 다음과 같다.

1. 0부터 1사이의 난수를 생성한다.
2. 이 숫자가 0과 엡실론(0과 1 사이의 값으로, 일반적으로 아주 작다) 사이에 존재하면, 50/50의 확률로 동전 뒤집기를 시행한다.
   a. 그 결과 동전이 앞면이면 제안 A를 표시한다.
   b. 동전이 뒷면이면 제안 B를 표시한다.
3. 숫자가 엡실론보다 크면, 지금까지 가장 좋은 결과를 보인 제안을 표시한다.

엡실론은 이 알고리즘을 제어하는 단일 파라미터이다. 엡실론이 1이면 우리는 결국 간단한 표준 A/B 검정(매 실험마다 A와 B를 무작위로 할당)을 하게 되는 셈이다. 엡실론이 0이라면 완전한 **탐욕 알고리즘**이 되어버린다. 더 이상의 실험 없이, 피실험자(웹 방문자)들을 항상 지금까지 알려진 가장 좋은 제안에 할당한다.

여기서 조금 더 복잡한 알고리즘은 '톰슨의 샘플링'을 사용하는 방법이다. 여기서는 각 단계마다 '표본을 추출'(손잡이를 당김)하여 최고의 손잡이를 선탠할 확률을 최대화한다. 당연히 어느 것이 좋은 손잡이인지 모른다(이것이 늘 문제다). 그러나 연속적인 추출을 통해 얻는 수익을 관찰하면 더 많은 정보를 얻을 수 있다. 톰슨 샘플링은 베이지언 방식을 사용한다. 즉 **베타 분포**(베이지언 문제에서 사전 정보를 지정하는 일반적인 메커니즘)를 사용하여 수익의 일부 사전 분포를 가정한다. 각 추출 정보가 누적되면서 정보가 업데이트 되기 때문에, 다음번에 최고 손잡이를 선택할 확률을 효과적으로 최적화할 수 있다.

밴딧 알고리즘은 3가지 이상의 처리를 효율적으로 다루고 '최고'를 위한 최적의 선택을 하도록 돕는다. 전통적인 통계검정의 경우, 3가지 이상의 처리를 위한 의사 결정은 전통적인 A/B 검정의 의사 결정보다 훨씬 복잡하며, 이 경우 밴딧 알고리즘의 정점이 훨씬 더 커진다.

주요개념

- 전통적 A/B검정은 임의표집 과정을 기본으로 하기 때문에, 수익이 낮은 것을 너무 많이 시도할 수 있다.
- 이와 대조적으로 MAB는 실험 도중에 얻은 정보를 통합하고 수익이 낮은 것의 빈도를 줄이는 쪽으로 표본 추출 과정을 변경한다.
- 또한 두 가지 이상의 처리를 효과적으로 다룰 수 있다.
- 추출 확률을 수익이 낮은 처리에서 수익이 높으리라 추정되는 쪽으로 이동시키기 위한 다양한 알고리즘이 존재한다.

3.11 검정력과 표본크기

웹 테스트를 수행할 경우 실행 시간 (즉, 처리당 얼마나 많은 노출이 필요할까?)은 어떻게 결정할까? 웹 테스트에 대한 수많은 관련 자료들을 인터넷에서 쉽게 찾을 수 있다. 하지만 모든 경우에 ㅏㄱ ㅁ자는 일반적인 방법은 없고, 다만 원하는 달성 목표에 따라 조절해야 한다.

용어정리

- **효과크기 : '클릭률의 20% 향상'과 같이 통계 검정을 통해 판단할 수 있는 효과의 최소 크기**
- **검정력 : 주어진 표본크기로 주어진 효과크기를 알아낼 확률**
- **유의수준 : 검증 시 사용할 통계 유의수준**

표본크기에 대한 고려는 '가설검정이 실제로 처리 A와 B의 차이를 밝혀낼 수 있을까?'라는 질문과 바로 연결된다. 가설검정의 결과라고 할 수 있는 p값은 A와 B 사이에 실제 차이가 있는지에 따라 달라진다. 물론 실험에서 누가 어떤 그룹에 속하느냐는 선택의 운에 따라 결과가 달라질 수도 있다. 그렇다 하더라도 실제 차이가 크면 클수록, 그것을 밝혀낼 가능성도 따라서 커질 것이고, 그 차이가 작을수록 더 많은 데이터가 필요하다는 생각에는 모두 동의할 수 있다. 야구에서 3할 5푼 타자와 2할 타자를 구분하기 위해서는 많은 타석이 필요하지는 않다. 하지만 3할 타자와 2할 8푼 타자를 구분하기 위해서는 많은 타석 정보가 필요할 것이다.

검정력이란 바로 특정 표본 조건(크기와 변이)에서 특정한 **효과크기**를 알아낼 수 있는 확률을 의미한다. 예를 들어 25타서에서 3할 3푼 타자와 2할 타자를 구분할 수 있는 확률이 0.75라고 (가정해서) 말 할 수 있다. 여기서 효과크기란 바로 1할 3푼(0.130)를 의미한다. 그리고 '알아낸다'는 것은 가설검정을 통해 차이가 없을 것이라는 영가설을 기각하고, 실제 효과가 있다고 결론을 내리는 것을 의미한다. 다시말해 두 타자를 대상으로한 25타석(n-25) 실험은 0.130의 효과크기에 대해 0.75 혹은 75%의 (가설상의) 검정력을 가진다고 볼 수 있다.

몇 가지 '움직이는 부분'이 있다. 가설검정에서 표본 변이, 효과크기, 표본크기, 유의수준 등을 특정하는 데 필요한 수많은 통계적 가설과 수식에 말려들기 십상이다. 실제로 검정력을 계산하기 위한 특별한 목적의 통계 SW가 있다. 대부분의 데이터 과학자들은 (예를 들어 논문 출판에 필요한) 검정력을 구하기 위해 형식적인 절차를 모두 지킬 필요는 거의 없다. 하지만 A/B 검정을 위해 데이터를 수집하고 처리하는 데 비용이 발생하는 경우, 가끔 사용해야 할 수도 있다. 이럴 경우, 데이터 수집을 위해 대충 얼마의 비용이 발생할지 안다면 데이터를 수집하고도 결론을 내리지 못하는 상황을 피할 수 있을 것이다. 여기 꽤 직관적인 방법 하나르 소개한다.

1. 최대한 (사전 정보를 이용해서) 결과 데이터가 비슷하게 나올 수 있는 가상의 데이터를 생각해보자. 예를 들면 2할 타자를 위해 20개의 1과 80개의 0이 들어 있는 상자를 생각 한다든지, 아니면 웹 페이지 방문 시간을 관측한 자료가 담겨 있는 상자를 생각할 수 있다.
2. 첫 표본에서 원하는 효과크기를 더해서 두 번째 표본을 만든다. 예를 들면 33개의 1과 67개의 0을 가진 두 번째 상자, 혹은 각 초기 방문시간에 25초를 더한 두 번째 상자를 만들 수 있다.
3. 각 상자에서 크기 n인 부트스트랩 표본을 추출한다.
4. 두 부트스트랩 표본에 대해 순열 가설검정(혹은 수식 기반의 가설검정)을 진행한다. 그리고 여기에 통계적으로 유의미한 차이가 있는지 기록한다.
5. 3~4단계를 여러 번 반복한 후, 얼마나 자주 유의미한 차이가 발견되는지 알아본다. 이 확률이 바로 검정력 추정이다.

3.11.1 표본크기

검정력 계산의 주된 용도는 표본크기가 어느 정도 필요한가를 추정하는 것이다. 예를 들면 기존 광고와 새로운 광고를 비교하기 위해 클릭률(노출당 클릭이 발생하는 백분률)을 조사한다고 가정하자. 이 조사를 위해 얼마나 많은 클릭 수를 수집해야 할까? 50% 정도의 큰 차이에만 관심이 있다면 상대적으로 적은 수의 표본으로도 목표를 이룰 수 있을 것이다. 하지만 그것보다 훨씬 작은 차이에도 관심이 있다면 훨씬 큰 표본이 필요하다. 이런 식으로 새 광고가 기존 광고에 비해 얼만큼(예를 들어 10%) 더 효과적이어야 하는지, 어느 정도가 아니면 기존 광고로 계속 갈지에 대한 기준을 설정하는 것이 표준적인 접근법이다. 이러한 목표, 즉 '효과크기'가 표본크기를 좌우한다.

예를 들어 현재 클릭률이 약 1.1% 수준인데, 여기서 10% 증가한 1.21%를 원한다고 가정하자. 이댸 우리는 두 상자, 1.1%의 1이 들어 있는 상자 A(110개의 1과 9,890의 0)와 1.21%의 1이 들어 있는 상자 B(121개의 1과 9,879개의 0)가 있다고 생각할 수 있다. 먼저 각 상자에서 300개씩 뽑는다고 하자(이는 마치 각 광고에 대한 300의 '첫인상'이라고 할 수 있다). 결과가 다음과 같다고 가정하자.

- 상자 A : 3개의 1
- 상자 B : 5개의 1

어떤 가설검정을 해도 이 차이 (5개와 3개)가 유의미하지 않다고 나올 것이라고 쉽게 눈치챘을 것이다. 이 표본크기(300개)와 효과크기 (10% 차이)의 조합은 가설검정을 통해 이 차이를 보이기에는 너무 작다.

따라서 이번에는 표본크기를 증가시켜 2,000개의 첫 인상을 알아보자. 그리고 더 큰 효과크기를 생각해보자.

다시 클릭률은 여전히 1.1% 수준이라고 가정한다. 대신 50% 증가한 1.65를 원한다고 생각해보자. 아까와 마찬가지로 두 상자, 1.1% 1이 들어 있는 상자 A(110개의 1과 9,890개의 0)와 1.65% 1이 들어 있는 상자 B(165개의 1과 9,835개의 0)가 있다고 생각할 수 있다. 이제 각 상자에서 2,000개를 뽑는다. 이렇게 뽑은 결과가 다음과 같다고 하자.

- 상자 A: 19개의 1
- 상자 B: 34개의 1

하지만 유의성 검정을 해도 이 차이(34-19)여전히 '유의미허자 엃다'라고 결론 날 것이다 (앞선 5-3의 차이보다는 유의미한 결과에 훨씬 더 가깝기는 하지만). 검정력을 계산하기 위해서는 이러한 과정을 여러 번 반복해야 한다. 아니면 검정력 계산을 지원하는 SW를 사용할 수도 있다. 하지만 앞 예제를 통해 알 수 있듯이 50% 정도의 효과를 알기 위해선 수천 개 이상의 광고 첫인상 정보가 필요할 것이다. 요약하면, 검정력 혹은 필요한 표본크기의 계산과 관련한 다음 4가지 중요한 요소들이 있다.

- **표본크기**
- **탐지하고자 하는 효과크기**
- **가설검정을 위한 유의수준**
- **검정력**

이 중 3가지를 정하면 나머지 하나를 알 수 있다. 가장 일반적으로, 표본크기를 알고 싶을 경우가 많다. 이때 나머지 3가지 요솔ㄹ 정해야 한다. 아래 R코드는 같은 크기의 두 표본을 고려한 검정을 위해 사용된다. 이를 위해 pwr 패키지를 사용한다.

```
pwr.2p.test(h=...,sig.level=...,power=...)
```

여기서 h는 효과크기(비율)이고 n은 표본크기, sig.level은 검정을 수행할 유의수준(알파), power는 검정력(효과크기를 알아낼 확률)이다.

주요개념

- **통계 검정을 수행하기 앞서, 어느 정도의 표본크기가 필요한지 미리 생각할 필요가 있다.**
- **알아내고자 하는 효과의 최소 크기를 지정해야 한다.**
- **또한 효과크기를 알아내기 위해 요구되는 확률(검정력)을 지정해야 한다.**
- **마지막으로, 수행할 가설검정에 필요한 유의수준을 정해야 한다.**

3.12 마치며

실험설계 원칙(대상을 서로 다른 처리군으로 랜덤하게 할당하는 원칙)을 통해 실험이 얼마나 잘 진행되었는지 타당한 결론을 도출할 수 있다. '아무런 변화도 가하지 않은' 대조 처리군을 포함하는 것은 필수적이다. 전통 적인 통계 추론 주제들, 즉 가설검정, p값, t 검정 등 같은 주제들은 기존의 통계 강의나 참고자료에서 가장 많은 시간과 공간을 차지한다. 데이터 과학의 입장에서 보면 이런 기존의 틀은 거의 필요하지 않다. **하지만 랜덤 변이가 사람을 속이는 데 중요한 역활을 한다는 것을 이해하는 것은 중요하다.** 데이터 과학자들은 직관적인 재표본추출 과정 (순열과 부트스트랩) 을 통해 데이터 분석에서 우연에 의한 변이가 어느 정도까지 영향을 미치는지 측정할 수 있게 됐다.