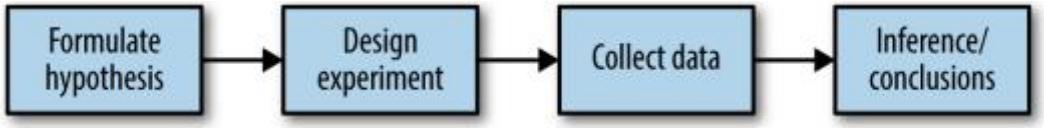
실험 설계(Design of experiment)

- experiment: 어떤 가설(hypothesis)을 확인(confirm)/기각(reject)하여 결론을 내기위해 설계
  예시) 어떠한 결정을 내릴 시 유의미한 차이가 발생하는지 등
  신약이 기존의 약보다 효력이 좋은지 / 판매가를 변경하면 수익이 오를지



classical statistical inference pipeline

* inference: 제한된 데이터셋에서 수행한 실험 결과를 더 큰 process나 모집단(population)에 적용하는 것

A/B Testing

- 방안, 절차 등 두가지 처리 방법(treatment)에 대해 어느 것이 나은지 확인하기 위해, 두 그룹으로 구성한 실험
- 일반적으로 기존 방식과 새로운 방식에 대해 비교하며, 기존 방식을 따르는 그룹을 대조군(control)으로 둠
  control group: 기존 방식대로 처리한 subject들의 그룹
  treatment group: 새로운/특정한 방식으로 처리한 subject들의 그룹
- treatment만 두 그룹간의 차이 발생의 원인이 되도록 제어하기, 위해 subject들은 랜덤하게 선정
- test statistic: 두개 그룹을 비교하기 위해 사용할 지표(평균 등)
  연구자 편향(bias)을 막기 위해, 실험을 실행하기 전 결정

적용 사례 예시) web context

- treatment: web page, price of a product, wording of a headline, ...
- subject: web visitor
- outcome: clicks, purchases, visit duration, number of pages visited, whether a particular page is visited

- 두가지 이상의 treatment를 비교하는 experimental design: multi-arm bandit

Hypothesis Tests (Significance test)

- 가설 검정 (유의성 검정)
- 어떤 현상이 우연히 일어난 것인지 / 분명한 차이가 있는 것인지 확인

- Null hypothesis: 귀무가설. 두 그룹은 본질적으로 같고, 값 차이는 우연히 일어났다고 가정
- Alternative hypothesis: 대립가설. 귀무가설과 반대되는 가정. 일반적으로 실험을 통해 입증하고 싶은 내용으로 설정
  Null = “no difference between the means of group A and group B”;
  alternative =“A is different from B” (could be bigger or smaller) //
  Null = “A ≤ B”; alternative = “A > B”
- 가설 검정을 통해, 귀무가설이 거짓 여부를 입증함으로써 대립가설의 참/거짓을 확인할 수 있음

- One-way(one-tail) test: 추정 값이 기준 값에서 왼쪽/오른쪽 중 한 방향으로만 벗어날 때 사용
  Null = “A = B”; alternative =“A > B” (“A > B”)
- Two-way(two-tail) test: 추정 값이 기준 값보다 크거나 작은 경우에 사용
  Null = “A = B”; alternative =“A ≠ B”

Resampling

- 통계량의 random variability를 평가하기 위해, observed data에서 값을 여러번 반복 추출하는 것

- bootstrap: sample에서 계산한 추정치(estimate)의 신뢰성(reliability)을 평가하기 위해 사용
- permutation: 두 개 이상 그룹을 포함하는 가설 검정에 사용

Permutation Test

1. 서로 다른 그룹(A, B, C, D, ...)의 데이터를 연결하여 single data set 구축
2. 새로 구축한 데이터셋을 섞고(shuffle), 이를 비복원추출하여 기존 A 그룹과 같은 사이즈인 새로운 데이터 그룹 형성
3. 마찬가지로 남은 데이터셋을 비복원 추출하여 기존 그룹 B, C, D, ...와 같은 크기의 새로운 데이터 그룹 형성
4. 기존 그룹들에서 계산했던 statistic/estimate(그룹 간 비율 차이 등)을 새로 형성한 그룹들에 대해서도 계산; 순열 반복 1회
5. 1~4 단계를 R회 반복하여, test statistic의 permutation distribution 생성
6. 최초 관찰했던 그룹간 차이(difference)와 permuted 그룹간 차이를 비교했을 때, 최초 관찰한 차이가 permutation distribution에서 드물게 나타나는 경우 해당 차이는 통계적으로 유의미하다고 판단함 (statistically significant). 즉 그룹간에 분명한 차이가 존재함

Permutation Test - example

문제 상황: 고품질의 서비스를 판매하는 회사에서 두가지 웹페이지(A, B) 중 어떤 웹페이지가 더 홍보를 잘하는지 알고 싶으나, 서비스의 가격이 높아 충분한 판매 데이터가 없음

변수 설정: proxy variable 사용. 웹페이지를 오래 보는 사람은 서비스를 구매할 확률이 높다고 판단하여, 웹페이지 잔류 시간(session time)을 변수로 사용하기로 결정

실험 설계: 웹페이지 A, B의 average session time을 비교하여 통계적으로 유의미한 차이가 있는지 검정

- 대리변수(proxy variable): 획득이나 사용이 어려운 어떤 변수를 대신하여 사용하는 변수

In [ ]:

```
```

- random permutation test(randomization test): random shuffling & dividing
- exhaustive permutation test(exact test)
- bootstrap permutation test

exhaustive permutation test

- 데이터가 나누어질 수 있는 모든 경우의 수 고려 (sample size가 작을 때만 실용적)

bootstrap permutation test

- 복원추출로 진행

Statistical Significance and p-Values

- statistical significance: 우연히 일어날 수 있는 것보다 극단적인 결과를 내는지 측정하는 것
  chance variation 범위를 넘어서는 경우 통계적으로 유의하다고 판단함
- p-value: chance model에서 관찰된 결과보다 극단적인(extreme) 값이 나올 수 있는 빈도(frequency)

Outcome	Price A	Price B	전체
Purchase	200	182	382
Not Purchase	23,539	22,406	45,945
전체	23,739	22,588	46,327

Outcome	Price A	Price B	전체
Purchase	196	186	382
Not Purchase	23,543	22,402	45,945
전체	23,739	22,588	46,327

In [ ]:

Alpha(α)

- null hypothesis 기각(reject) 기준이 되는 임계값(threshold)
- 일반적으로 alpha level은 5%(0.05), 1%(0.01) 사용

Type I error, Type II error

- Type I error: 참인 가설을 기각하는 오류(false negative)
- Type II error: 거짓인 가설을 채택하는 오류(false positive)

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

- 데이터의 타입(count data / measured data), sample 개수 등에 따라 적절한 타입의 significance test를 사용함

자유도(Degrees of Freedom)

- 표본 데이터에서 계산된 statistic에 적용되는 개념으로, 자유롭게 변경할 수 있는 값의 '개수'를 의미하며, 여러 확률 분포의 모양에 영향을 미침
  예시) 10개의 값이 있는 표본집단의 mean을 알고 있다면, 자유도는 9

t-Test

- 단일 표본 평균의 분포를 근사하기 위해 개발된 t-분포에서 이름을 따옴
- 데이터 타입이 numeric(not binary)인 A/B test에서 많이 사용됨

In [ ]:

ANOVA

- numeric data를 가진 multiple group 간 통계적으로 유의미한 차이가 있는지 검정하는 절차

	Page 1	Page 2	Page 3	Page 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Average	172	185	176	162
Grand average				173.75

- 4개의 웹페이지에, 각각 5명의 방문자가 잔류했던 시간(second)
- 각 웹페이지에 대한 기록은 서로 완전히 독립적이라고 가정
- 각 웹페이지에 대해 평균 잔류 시간이 유의미하게 차이가 있는지 검정

In [ ]:

- 그룹 간 비교는 6가지 경우가 가능함
- 1 vs 2 / 1 vs 3 / 1 vs 4 / 2 vs 3 / 2 vs 4 / 3 vs 5

- 모든 그룹이 서로 차이가 없다라는 귀무가설은, 실행하는 검사 수가 많을수록 귀무가설의 진위와 상관없이 기각될 확률이 낮아짐
- 위 6가지에 대해 각각 검정을 시행하는 대신, '모든 웹페이지는 차이가 없고, 관찰된 웹페이지 간 차이는 우연히 발생할 수 있는 경우'인가를 검정할 수 있는 실험을 할 수 있음 -> ANOVA

1. 모든 데이터를 섞어 하나의 데이터셋으로 만듬
2. 새로운 데이터셋을 추출하여, 기존 그룹의 개수/데이터셋 크기와 동일한 새로운 그룹 형성
3. 새로운 그룹들의 평균 기록

4. 그룹 평균들의 분산 기록
5. 2~4 단계를 R회 가량 반복
6. resampled variance와 observed variance를 이용하여 p-value 계산

In [ ]:

```
```

F-Statistic

- 두 그룹의 평균을 비교하는 permutation test 대신 t-test를 사용했던 것처럼,
  ANOVA는 F-statistic에 기반함
- F-statistic: residual error로 인한 분산에 대한 그룹 평균 간 분산의 비율을 기반으로 하며,
  이 비율이 높을수록 통계적으로 유의하다고 판단함

In [ ]:

```
```

Two-Way ANOVA

- one-way ANOVA: A/B/D/D test 등, 한가지 요소만 차이나는 경우
- two-way ANOVA: 두가지 요소가 있는 경우
  예시) group A/B,...와 주말/평일

- 'interaction effect'을 식별하여, one-way ANOVA와 유사한 방식으로 처리함
  먼저 grand average의 effect과 treatment effect를 확인한 후, 각 그룹(treatment) 내에서 주말/평일 observation을 분리하여 treatment average와 해당 하위 분류의 average의 차이를 확인