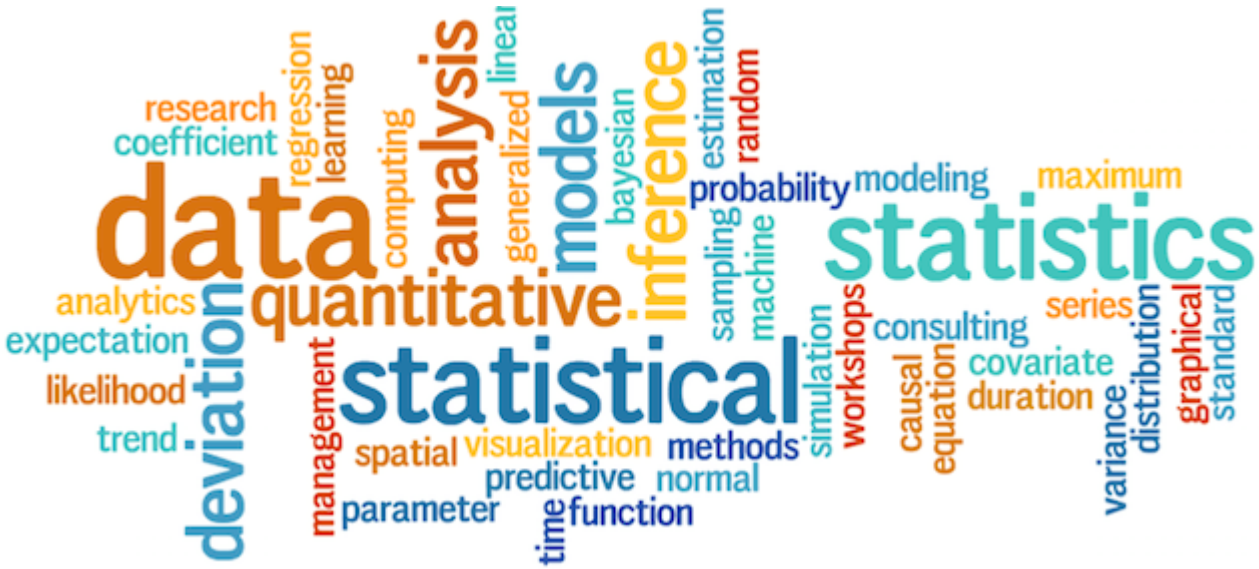


통계기반 데이터 분석 (Ch 3)

데이터사이언스 & A.I 전공
정화민 교수

확률 , 확률분포 (복습)

- 표준 정규분포 예

  예 : 어느 대학교 남학생들 키의 평균은 170cm, 표준편차는 6cm입니다. 이 대학교에서 남학생의 키가 182cm 이상일 확률은 다음과 같이 구합니다.(남학생의 키는 정규분포를 따로 있는 것으로 가정

$$P(X \geq 182) = 1 - P(X \leq 182) = 1 - \int_{-\infty}^{182} \frac{1}{6\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-170}{6}\right)^2} dt$$

  - 표준화 변환을 통한 표준정규분포로 계산

$$z = \frac{x-\mu}{\sigma} = \frac{182-170}{6} = \frac{12}{6} = 2$$

  이를 이용하여 표준정규분포에서 구하면 다음과 같음

  - $P(Z \geq 2) = 1 - P(Z \leq 2)$
  - 표준정규분포표에서 z값이 2가 되는 값, 즉 행에서 2.0을 찾고 열에서 0.00을 찾은 값은 0.977(유효숫자 소숫점 세째자리)

  다음페이지 표로부터 표준정규분포에서 2보다 작을 확률은 0.977이고, z가 2보다 클 확률은 1-0.977≈0.023임.

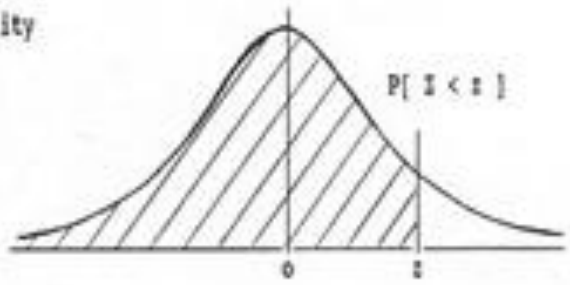
- 이제 다시 원래의 정규분포로 돌아가서 z 값으로 변환하여 2가 된 원래의 값을 구해보면 182. 이를 통해 182cm보다 클 확률은 0.023이 됨을 알 수 있음.

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e. $$P[\,Z < z\,] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

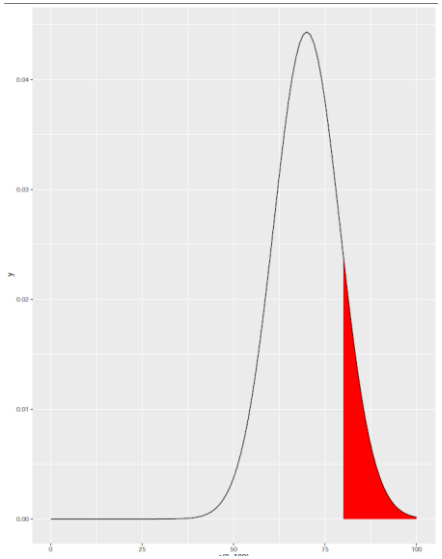
2주 (퀴즈 정답)

- 문제

  100명 학생의 수학평균은 70점 표준편차는 9
  나의 수학 점수 평균은 80점 일때 나의 수학점수 등수는 ?

```r
# 퀴즈정답 : 수학평균 70점, 표준편차 9 , 내 점수 90점 전체 학생 수 100명일때 나의 수학 점수 등수는 ?
pnorm(80, mean=70, sd=9)
1-0.8667397
100*0.1332603
# 나의 수학점수 등수는 13~14등 정도 됨
```

```r
install.packages("ggplot2")
library(ggplot2)
ggplot(NULL, aes(c(0, 100))) + geom_area(stat='function', fun=dnorm, args=list(mean=70, sd=9), xlim=c(80, 100), fill='red') +
  stat_function(fun=dnorm, args=list(mean=70, sd=9))
```

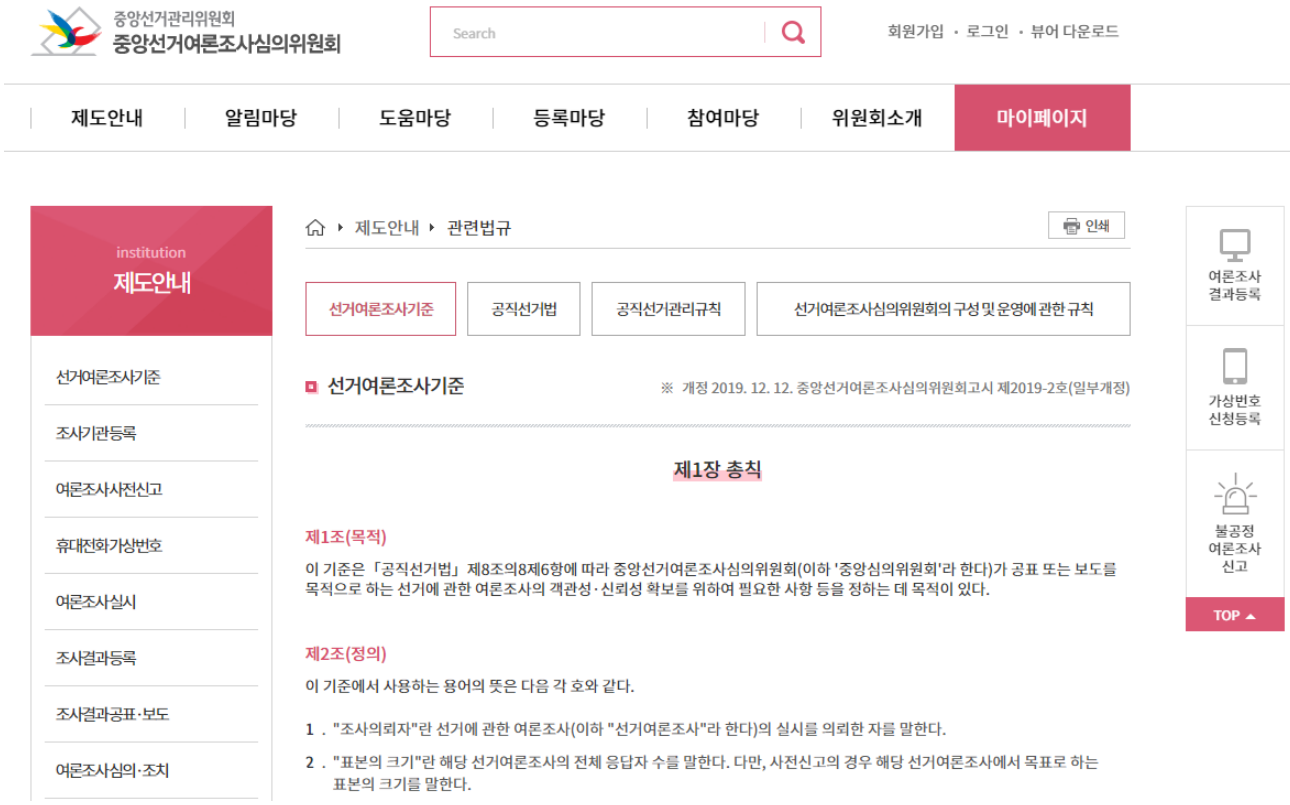


표본분포

- 표본분포 : 표본들로부터 모집단의 특성을 유추

  예 : 중앙선거여론조사 공정심의위원회 (여론조사 심의)

  - 선거기간 : 특정 후보자와 정당을 얼마나 많은 유권자가 지지하는지
  - 국정지지도, 잠재적 대선후보군에 대한 지지도 등



Source: www.nesdc.go.kr/

표본분포

- 모수(parameter): 모집단의 특성을 나타냄.
  -예 : 선거 시 A 후보에 대한 전체 유권자의 지지율 *P*

- 통계량(statistic): 표본으로 부터 관찰되는 표본의 특성
  -예 : 평균, 표준편차, 중앙값 등

- 표본분포(sampling distribution): 표본조사를 실시하고 조사를 위해 전체 표본을 한번 추출해서 표본의 특성을 구하고 모집단에 대하여 추측을 함
  - 비 복원추출 : 한번 추출한 표본은 다시 추출하지 않는 것
  - 복원추출: 한번 추출한 표본을 표본추출틀에 다시 넣는 방법
  - 예: 여론조사방법에서 표본추출틀은 표본 추출을 위한 모집단의 목록으로 이 조사에서는 무선전화번호를 사용하였음을 밝히고 있음 , 표본추출방법인 RDD는 표본추출틀 내에서 무작위(Random)로 전화번호 숫자(Digit)를 만들어 전화 연결 (Dialing) 하는 것을 말함

```
# 1~10까지의  6개 비복원 추출
sample(1:10,6)
```

```
## [1]  3  2  6  9  1 10
```

```
# 1~10까지의 복원 추출로 표본 뽑기
sample(1:10,6,replace=T)
```

```
## [1] 7 8 4 2 4 7
```

추정(estimation)

· 국회의원 선거철이 되면 다음과 같은 신문기사를 가끔 볼 수 있다
"~정치인 지지율 조사에서 A후보는 40%, B후보는 25%의 지지율을 얻었다. 이번 조사는 여론조사 전문기관인 00가 00구 성인남녀 천명을 대상으로 전화면접조사로 실시되었고, 신뢰수준 95%에서 표본오차는 ±3.1%포인트이다."
· 여기서 '신뢰수준 95%에서 표본오차 3.1%포인트'란 말의 의미는 무엇일까? 그 의미는 다음과 같다. 위와 같이 동일한 형태의 여론조사를 100번 실시했을 경우에 95번은 A후보가 40%에서 ±3.1% 인 36.9% ~ 43.1%, B후보는 25%에서 ±3.1% 인 21.9% ~ 28.1% 사이의 지지율을 얻을 것으로 기대된다는 의미이다.
즉, 신뢰수준이란 표본에 의한 조사 결과의 확실성 정도를 표현하는 것이며, 보통은 95%의 신뢰도를 사용하여 그러한 판단을 내린다.

Source: https://kostat.go.kr/

추정(estimation)

- 추측통계학
  - 표본으로부터 특성을 관찰하여 모집단의 특성을 유추하는 통계학의 한 분야
  - 추측통계학의 두 가지 연구 방법
    1) 추정 : 모집단으로부터 추출된 표본으로부터 특성(통계량)을 파악하여 이를 바탕으로 모수를 유추하는 방법
    2) 가설검정 : 모수에 대한 가설을 수립하고 이로부터 어떤 가설을 선택할 것인지를 통계적으로 결정하는 방법
  - 추정의 종류
    1) 점추정 : 표본의 특성을 나타내는 계산식(통계량) 중 모수를 유추하는 데 있어 최적의 계산식을 통해 구한 하나의 추정값을 구하는 방법
    2) 구간추정 : 하나의 점(값)이 아닌 모수의 참값이 포함될 것으로 기대하는 구간을 추정하는 방법

추정(estimation)

- **신뢰구간**
  - 구간추정을 위해 표본으로부터 구한 하한과 상한을 각각 $\widehat{\theta_L}, \widehat{\theta_U}$ 이라 할 때, $0 < \alpha < 1$ 인 α에 대해

$$P\left(\widehat{\theta_L} < \alpha < \widehat{\theta_U}\right) = 1 - \alpha$$

를 만족하는 구간 $(\widehat{\theta_L}, \widehat{\theta_U})$을 "모수 θ에 대한 $100(1-\alpha)\%$ 신뢰구간"이라 부르고, $(1-\alpha)$를 신뢰수준이라 함.

  - 구간추정

표본평균의 분포를 표준정규분포로 변환한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같이 하한으로 $\widehat{\theta_L} = -z_{\alpha/2}$, 상한으로 $\widehat{\theta_L} = z_{\alpha/2}$를 갖는 영역입니다.

$$P\left(\widehat{\theta_L} < Z < \widehat{\theta_U}\right) = P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha$$

다음 페이지 그림 참조

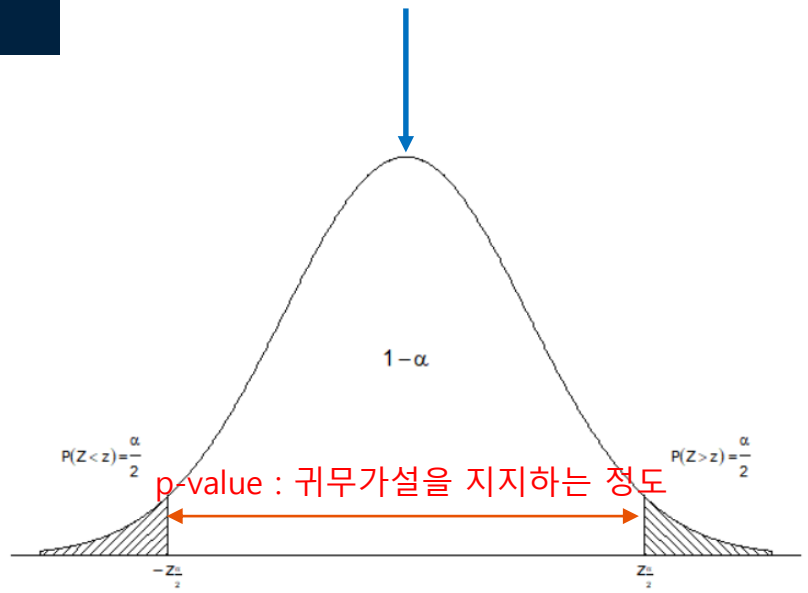
추정(estimation)

- 구간추정 예 : R 코딩 값 참조

```r
# 표본평균의 분포에서의 신뢰구간, 표본평균의 분포에서의 신뢰구간(1-a)
# 코딩예: 표준정규분포에서 평균을 중심으로 95% 면적 오류확률 a=0.05 그래프
# 95% 신뢰구간 (이후 가설검증의 기준이 됨  95% 이내 귀무가설 채택)
par(mar=c(0,1,0,1))
x <- seq(-3, 3, by=0.01)
y <- dnorm(x)
plot(x, y, axes=F, type="l", ylim=c(-0.1, 0.5), xlab="", ylab="")
abline(h=0)
ll <- qnorm(0.025)
ul <- qnorm(0.975)
polygon(c(-3, x[x<ll], ll), c(0, y[x<ll], 0), density=20)
polygon(c(ul, x[x>ul], 3), c(0, y[x>ul], 0), density=20, angle=135)

text(0, 0.2, expression(1-alpha))
text(-2.5, 0.1, expression(plain(P)(Z<z) == over(alpha, 2)), cex=0.7)
text(2.5, 0.1, expression(plain(P)(Z>z) == over(alpha, 2)), cex=0.7)
text(-1.96, -0.02, expression(-z[over(alpha, 2)]), cex=0.8)
text(1.96, -0.02, expression(z[over(alpha, 2)]), cex=0.8)
```

귀무가설이 참일 때 데이터의 분포



가설(hypothesis)

- **가설검정의 과정**

  - 모집단 특성의 상태에 대한 주장인 가설에 대해 표본으로부터 얻은 정보를 바탕으로 이를 채택할지 기각할지를 판단함으로써 모집단의 상태에 대해 결정하는 과정으로, 다음의 4단계를 거쳐 이루어 짐.

    - 1단계 : 가설 수립
    - 2단계 : 표본으로부터 검정을 위한 통계량 계산
    - 3단계 : 가설 선택의 기준 수립
    - 4단계 : 판정

- **가설종류**

  - 영가설(귀무가설, H_0)

    - 주로 기존에 알려진 것과 차이가 없음을 나타냄.

  - 대안가설(대립가설, H_1)

    - 주로 기존에 알려진 것과 차이가 있음을 나타냄.
    - 연구자가 밝히고자 하는 가설로 “연구가설”이라고도 함.

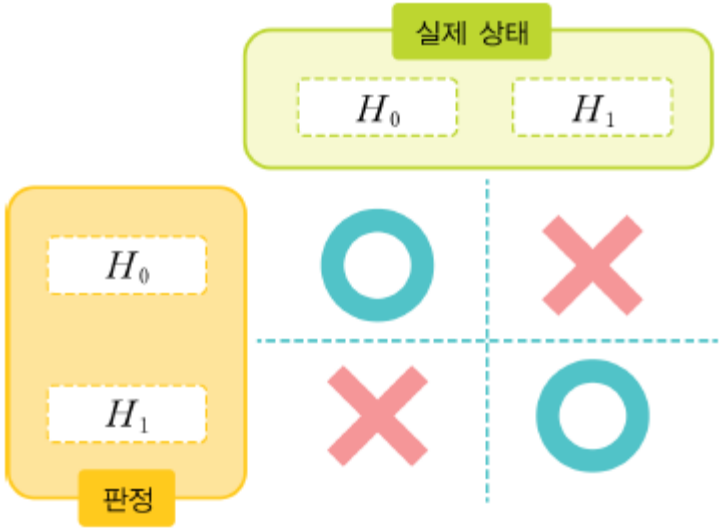
가설(hypothesis)

- **가설검정시 판정의 오류**

- 실제 영가설이 참일 때 가설검정을 통해 대안가설을 선택하거나, 실제 영가설이 거짓일 때 가설검정을 통해 영가설을 선택하는 경우

-제 1종 오류 : 영가설이 참인데 대안가설을 선택하는 오류
(예: 실제 유죄이지만 무죄 판결을 받아 사회로 돌아감)

제 2종 오류 : 영가설이 거짓인데 영가설을 선택하는 오류
(예: 실제 무죄이나 유죄 판결을 받아 양형에 따른 수감 생활0



민감도, 특이도, 정확도

* 용어정의 :

- 민감도(Sensitivity) : 특정질환을 가지고 있는 사람들 중 검사결과가 양성으로 나오는 비율.
- 특이도(Specificity) : 특정질환을 가지고 있지 않은 사람들 중 검사 결과가 음성으로 나오는 비율.
- 정확도(Accuracy) : 전체를 기준으로 맞은 걸 맞다고 아닌걸 아니라고 맞춘 비율

Table 1. 민감도(Sensitivity) 및 특이도(Specificity) 계산

항목		확진검사	
		질병유	질병무
검사	양성	TP (True Positive)	FP (False Positive)
	음성	FN (False Negative)	TN (True Negative)
		TP+FN	FP+TN
$\frac{TP+TN}{TP+FP+FN+TN} \times 100$ 정확도(Accuracy) 공식		$\frac{TP}{TP+FN} \times 100$ 민감도 공식	$\frac{TN}{FP+TN} \times 100$ 특이도 공식

Table 3. Baseline의 민감도(Sensitivity) 및 특이도(Specificity)

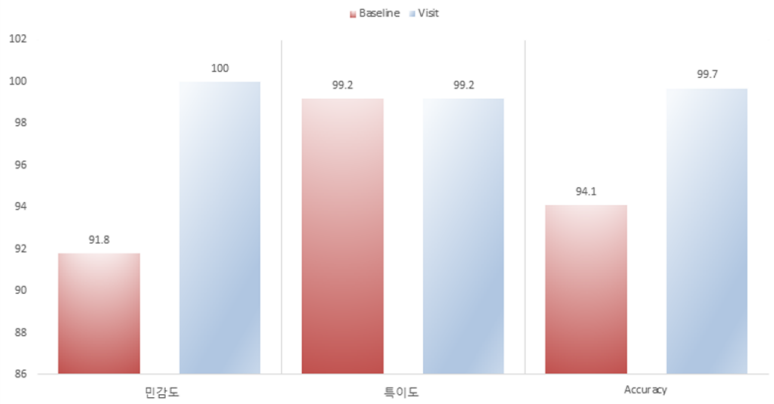
항목			DLP최종진단		전체
			성병(질병유)	정상	
Baseline (기존 스왑 검사)	양성	빈도	516	2	518
		전체 중 %	62.5%	0.2%	62.7%
	음성	빈도	46	262	308
		전체 중 %	5.6%	31.7%	37.3%
전체		빈도	562	264	826
		전체 중 %	68.0%	32.0%	100.0%
민감도 / 특이도			91.8% 민감도	99.2% 특이도	-

Table 6. Visit(신형스왑 사용검체)의 민감도(Sensitivity) 및 특이도(Specificity)

항목			DLP최종진단		전체
			성병(질병유)	정상	
Visit(신규 스왑검사)	양성	빈도	562	2	564
		전체 중 %	68.0%	0.2%	68.3%
	음성	빈도	0	262	262
		전체 중 %	0.0%	31.7%	31.7%
전체		빈도	562	264	826
		전체 중 %	68.0%	32.0%	100.0%
민감도 / 특이도			100% 민감도	99.2% 특이도	-

Table 9. 민감도(Sensitivity) , 특이도(Specificity), 정확도 (Accuracy) 비교

항목	Baseline(구형스왑 사용)	Visit(신형스왑 사용)
민감도	91.8%	100%
특이도	99.2%	99.2%
정확도	94.1%	99.7%



[그림 4] Baseline(기존스왑 사용)과 Visit(신형스왑 사용)의 정확도 비교

정확도 사례

Source https://www.youtube.com/watch?v=Pcf1RDfYyCY