제6강: 모수추정 및 가설검정

금융 통계 및 시계열 분석

TRADE INFORMATIX

2014년 1월 24일

목차

문제

문제 1 : 기대 수익률 구하기

삼성전자 주가의 기대 수익률은?

문제 2 : 기대 수익률 판단

삼성전자 주가의 1달 평균 수익률이 0.1% 이다. 한달 평균 수익률로 계산한 기대 수익률은 값은 얼마나 정확한 값일까? 이를 근거로 주가가 오르고 있다고 판단할 수 있는가?

문제 3 : 기대 수익률 비교

삼성전자 주가의 한 달 평균 수익률은 4.0% 이고 현대차 주가의 한 달 평균 수익률은 3.9% 이다. 삼성전자 주가의 기대수익률은 현대차 주가의 기대수익률보다 높다고 할 수 있나?

문제 1의 접근 방법

문제 1 : 기대 수익률 구하기

삼성전자 주가의 기대 수익률은?

- ❏ 삼성전자 주가의 일간 수익률이 특정한 확률분포에서 생성된 확률변수라고 가정한다. 또 이 확률분포는 미래에도 변하지 않을 것이라고 가정하자.
- ❏ 이 확률분포의 평균값은 삼성전자 주가의 기대수익률이다.

확률분포의 모수 추정 (Estimation of distribution parameters)

- 가정
  - 관심을 가지는 변수의 집합이 어떤 확률분포에서 생성된 확률변수의 샘플집합이라는 가정
- 모수추정의 목표
  - 샘플의 특성으로부터 샘플이 생성된 확률분포의 모수(parameter)를 추정
- 모수추정 대상
  - 평균 (mean)
  - 분산 (variance)
  - 평균의 차이 (difference of two means)
  - 분산의 비율 (ratio of two variances)
- 모수추정의 응용
  - 주식의 기대 수익률 (expected return) : normal 분포의 평균
  - 주식 변동성 (volatility) : normal 분포의 분산

일점 모수 추정 방법론 (Methods of parameter estimation)

- (G)MM: (Generalized) Method of Moments
  - 샘플 모멘트가 실제 모멘트와 같다고 가정
  - 각 샘플이 같은 분포에서 나왔고 서로 독립적이면 (iid) 샘플 평균과 샘플 분산은 평균 및 분산의 불편추정치
- MLE: Maximum Likelihood Estimation
  - 우도 (likelihood) 가 최대가 되는 파라미터 값을 찾음
  - 분포가 정상분포이면 샘플 평균과 분산은 분포 평균 및 분포 분산의 불편추정치

일점 모수 추정과 베이지안 추정

- 일점 모수 추정 (point estimation)
  - 하나의 수치값으로 모수를 추정
  - 샘플 집합을 입력으로 하는 결정론적 함수 (deterministic function)
- 일전 모수 추정의 특징
  - 단일한 함수 형태로 추정식 (estimator) 를 구할 수 있으로 계산 편의성
  - 결과가 parametric
  - 추정치 자체로는 추정 신뢰도를 표현할 수 없으므로 신뢰구간 (interval of confidence), 검정 (testing) 등 추가적인 분석이 필요
  - 신뢰구간을 구하는 방법을 구간추정 (interval estimation) 이라고도 함
- 베이지안 추정 (Bayesian estimation)
  - 모수 추정 결과를 별도의 분포를 이용하여 표현
  - 결과가 non-parametric

Method of Moments 1 : 평균

- 평균 μ의 추정치 = 샘플 평균
- 증명 : 추정치 즉, 샘플평균의 기대값은 분포 평균

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\begin{aligned} E[\bar{x}] &= E\left[\frac{x_1 + \cdots + x_n}{n}\right] \\ &= (E[x_1] + \cdots + E[x_n])/n \\ &= (\mu + \cdots + \mu)/n \\ &= \mu \end{aligned}$$

Method of Moments 2 : 분산

❑ 분산 σ의 추정치 = 샘플 분산

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

$$\begin{aligned}
E[s^2] &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \frac{n}{n-1}E[(x_1-\bar{x})^2] \\
&= \frac{n}{n-1}\left(E[x_1^2] - 2E[x_1\bar{x}] + E[\bar{x}^2]\right) \\
&= \frac{n}{n-1}\left(E[x_1^2] - 2E\left[x_1\frac{1}{n}\sum_{i=1}^{n}x_i\right] + E\left[\frac{1}{n}\sum_{i=1}^{n}x_i\right]\right) \\
&= \frac{n}{n-1}\left(E[x_1^2] - \frac{1}{n}E[x_1^2] - \frac{(n-1)}{n}E[x_1x_2]\right) \\
&= \frac{n}{n-1}\left(\frac{n-1}{n}E[x_1^2] - \frac{(n-1)}{n}E[x_1x_2]\right) \\
&= E[x_1^2] - E[x_1]E[x_2] \\
&= E[x_1^2] - \mu^2 = \sigma^2
\end{aligned}$$

Likelihood

- 모수 추정치가 모수의 참값에 의존하는 확률분포를 가지는 것처럼,
- 반대로
  모수의 참값를 모르는 상태에서 특정한 샘플 혹은 모수 추정치가 나왔을 때,
  모수의 참값은 이 모수 추정치에 의존하는 Likelihood(우도) 분포를 가진다.
- Likelihood (우도)
  - 어떤 모수 추정치가 모수의 참값이라고 했을 때 샘플에서 그 모수 추정치가 나올수 있는 확률

$$L(\theta|x) = P(x|\theta)$$

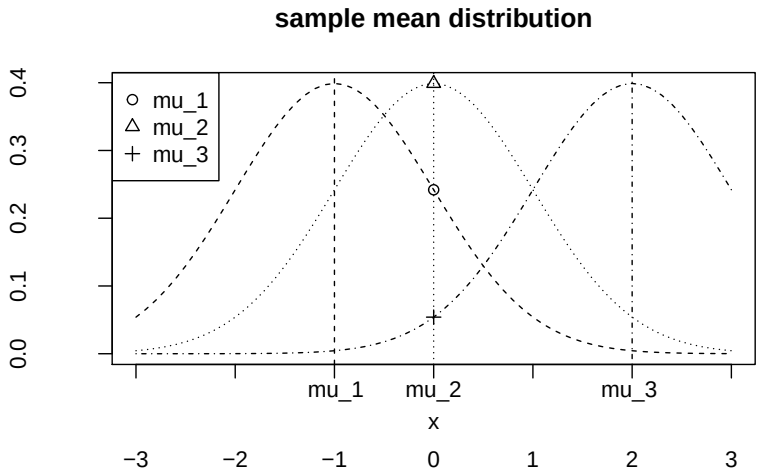
  - 확률분포와 likelihood 분포는 x와 μ 변수를 치환한 것에 불과. 따라서 분산을 알고 있는 normal 분포에서 평균을 추정하는 경우 ($\theta = \mu$)에는 두 함수는 x축 방향이 뒤바뀐 형태

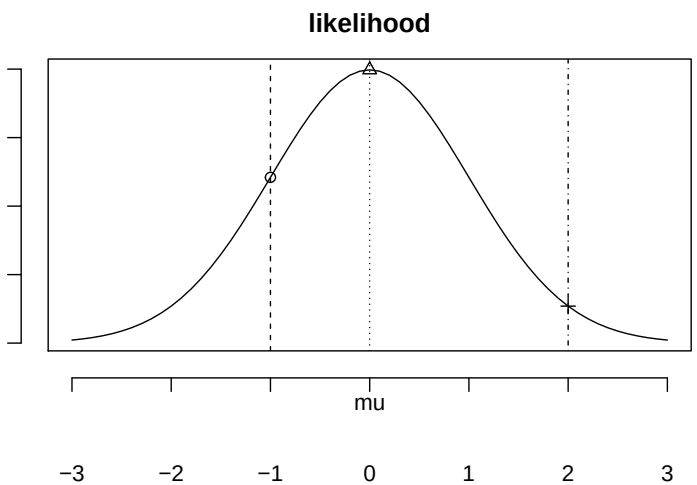
$$P(x;\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{x-\mu}{2\sigma^2}\right)$$

$$L(\mu;x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{\mu-x}{2\sigma^2}\right)$$

Likelihood 예시

```
> x <- seq(-3,3,0.1)
> d1 <- dnorm(x, -1, 1)
> d2 <- dnorm(x, 0, 1)
> d3 <- dnorm(x, 2, 1)
> layout(matrix(c(1,2), byrow=TRUE))
> par(mgp = c(0, 3, 0))
> plot(x, d1, type="l", lty=2,
+      xlim=c(-3,3),
+      main="sample mean distribution",
+      xlab="", ylab="")
> lines(x, d2, lty=3, xlim=c(-3,3))
> lines(x, d3, lty=4, xlim=c(-3,3))
> abline(v=-1, lty=2)
> abline(v=0, lty=3)
> abline(v=2, lty=4)
> points(0, dnorm(1), pch=1)
> points(0, dnorm(0), pch=2)
> points(0, dnorm(-2), pch=3)
> legend("topleft",
+        c("mu_1", "mu_2", "mu_3"), pch=1:3)
> mtext("x", side=1, line=1.6, adj=(0+3)/6.0)
> mtext("mu_1", side=1, line=0.4, adj=(-1+3)/6.0)
> mtext("mu_2", side=1, line=0.4, adj=(0+3)/6.0)
> mtext("mu_3", side=1, line=0.4, adj=(2+3)/6.0)
> par(mgp = c(0, 3, 0))
> plot(x, d2, type="l", line=1,
+      xlim=c(-3,3),
+      main="likelihood",
+      xlab="", ylab="")
> mtext("mu", side=1, line=1.4, adj=(0+3)/6.0)
> abline(v=-1, lty=2)
> abline(v=0, lty=3)
> abline(v=2, lty=4)
> points(-1, dnorm(1), pch=1)
> points(0, dnorm(0), pch=2)
> points(2, dnorm(-2), pch=3)
```





MLE에 의한 정상분포 모수 추정

- 정상분포 $\mathcal{N}(\mu, \sigma^2)$의 확률분포함수

$$P(x \mid \mu, \sigma^2) = f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- n개의 독립적인 샘플 값 $x_1, \cdots, x_n$이 나올 확률

$$\begin{aligned} P(x_1, \ldots, x_n, \mid \mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i, \mu, \sigma^2) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

MLE에 의한 정상분포 모수 추정 (계속)

- 우도함수 (Likelihood) 정의

$$\mathcal{L}(\mu, \sigma \mid x_1, \ldots, x_n) = f(x_1, \ldots, x_n, \mu, \sigma)$$

- 로그-우도함수 (Log-Likelihood) 정의

$$\log \mathcal{L}(\mu, \sigma \mid x_1, \ldots, x_n)$$

- 로그-우도함수 최대화

$$\frac{\partial}{\partial \mu} \log \mathcal{L}(\mu, \sigma) = 0, \quad \frac{\partial}{\partial \sigma} \log \mathcal{L}(\mu, \sigma) = 0$$

MLE에 의한 정상분포 모수 추정 : 평균

❑ 평균 추정

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \mu} \log \left(\left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left(- \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2} \right) \right) \\
&= \frac{\partial}{\partial \mu} \left(\log \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} - \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2} \right)
\end{aligned}
$$

$$
\hat{\mu} = \bar{x} = \sum_{i=1}^{n} x_i / n
$$

MLE에 의한 정상분포 모수 추정 : 분산

❑ 분산 추정

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \sigma} \log \left(\left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left(- \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2} \right) \right) \\
&= \frac{\partial}{\partial \sigma} \left(\frac{n}{2} \log \left(\frac{1}{2\pi\sigma^2} \right) - \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2} \right) \\
&= -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{\sigma^3}
\end{aligned}
$$

$$
\widehat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \widehat{\mu})^2 / n = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j
$$

$$
E\left[\widehat{\sigma}^2\right] = \frac{n-1}{n} \sigma^2
$$

문제 2의 접근 방법

문제 2 : 기대 수익률 판단

삼성전자 주가의 1달 평균 수익률이 0.1%이다. 한달 평균 수익률로 계산한 기대 수익률은 값은 얼마나 정확한 값일까? 이를 근거로 주가가 오르고 있다고 판단할 수 있는가?

- ❏ 주가가 오르고 있다는 것은 기대수익률이 양(positive)이라는 의미
- ❏ 한달 평균수익률은 기대수익률의 추정치이고 오차를 내포함
- ❏ 0.1%는 어느 정도의 정확도를 가진 숫자인가?
- ❏ 기대수익률이 양(positive)이라는 가설은 어느 정도의 근거를 가지는가?

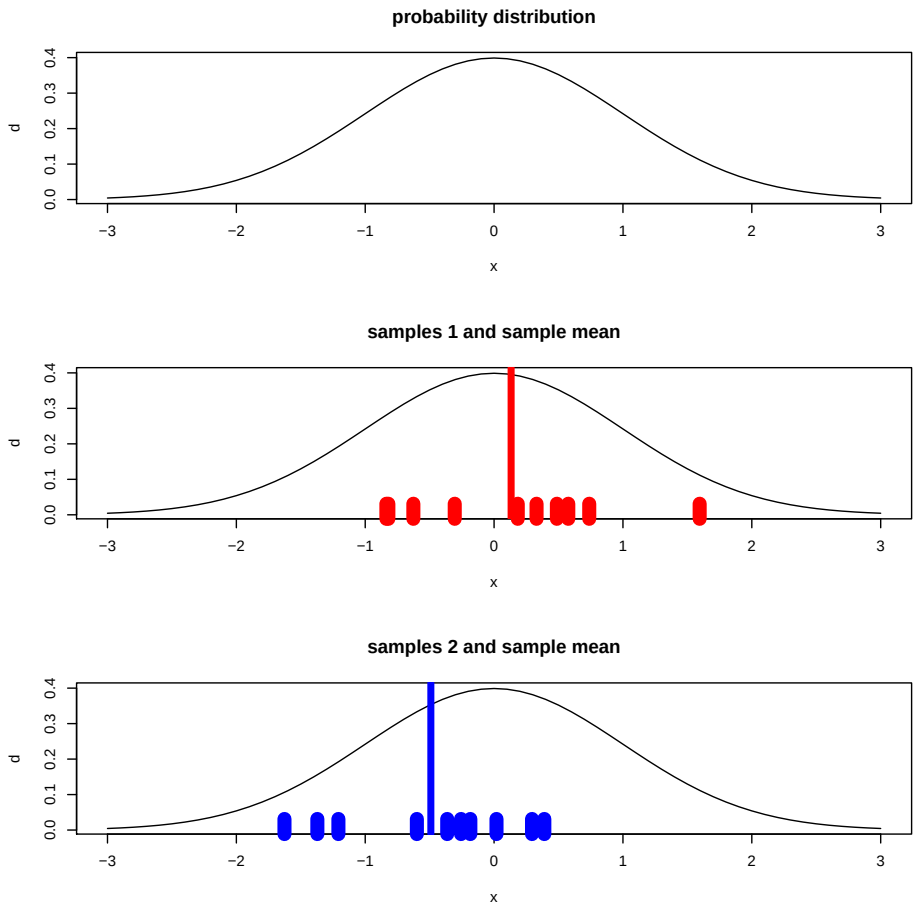
모수 추정치의 확률분포

- 모수 추정치는 샘플 집합 $\{x_i; i = 1, \cdots, n\}$의 함수로 계산됨
- 샘플 집합은 확률분포에 대한 일부 정보만을 가지므로 필연적으로 오차가 발생
- 샘플 집합이 확률변수의 결과이므로 그 함수값인 모수 추정치도 확률변수 (random variable)이고 그 나름의 확률분포를 가진다.
- 따라서 모수 추정치는 특정한 확률분포를 가지는 확률변수
- 모수 추정치의 확률분포는 다음 값에 의존한다.
  1. 원래 관심을 가진 확률변수의 모수
  2. 샘플의 갯수 n

확률분포의 모수 추정 예

❑ normal 분포의 평균 추정

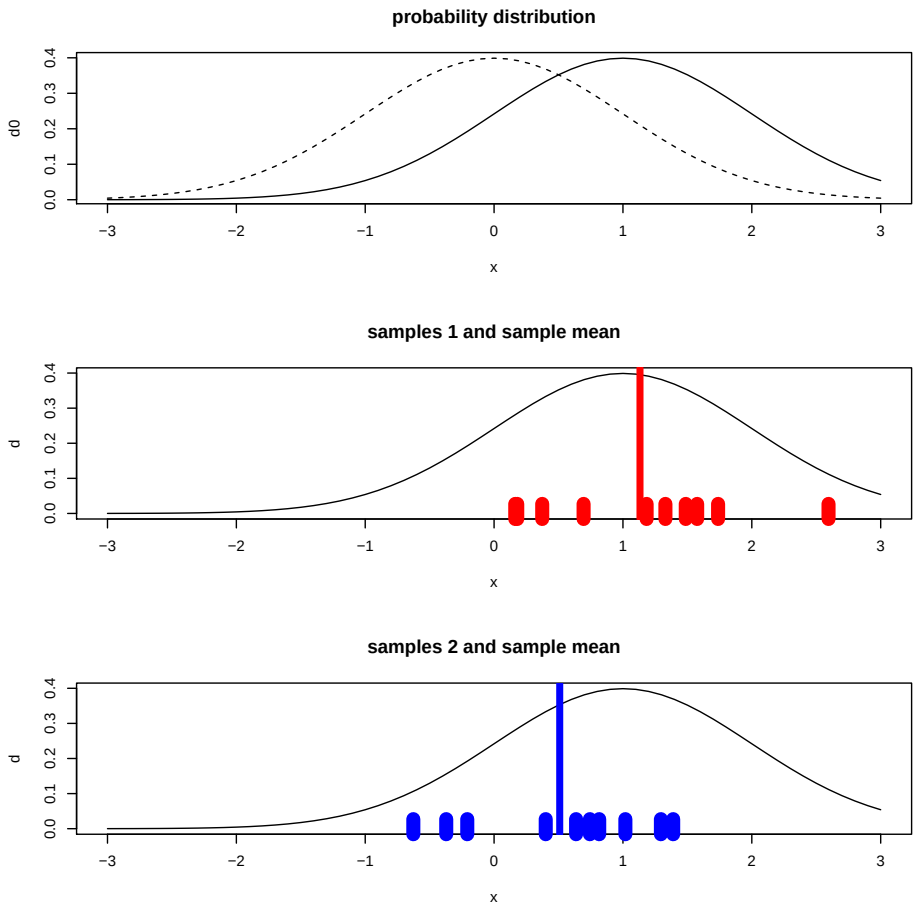
```
> x <- seq(-3,3,0.1)
> d <- dnorm(x, 0, 1)
> layout(matrix(c(1,2,3), byrow=TRUE))
> plot(x, d, type="l",
+      main="probability distribution")
> set.seed(1)
> y1 <- rnorm(10, 0, 1)
> mean(y1)

[1] 0.1322028

> set.seed(10)
> y2 <- rnorm(10, 0, 1)
> mean(y2)

[1] -0.4906568

> plot(x, d, type="l",
+      main="samples 1 and sample mean")
> rug(y1, 0.1, 1, 10, 'red')
> abline(v=mean(y1), lwd=5, col="red")
> plot(x, d, type="l",
+      main="samples 2 and sample mean")
> rug(y2, 0.1, 1, 10, 'blue')
> abline(v=mean(y2), lwd=5, col="blue")
```



확률분포의 모수 추정 예 (계속 1)

❑ 원래 확률분포의 모수가 달라지면?

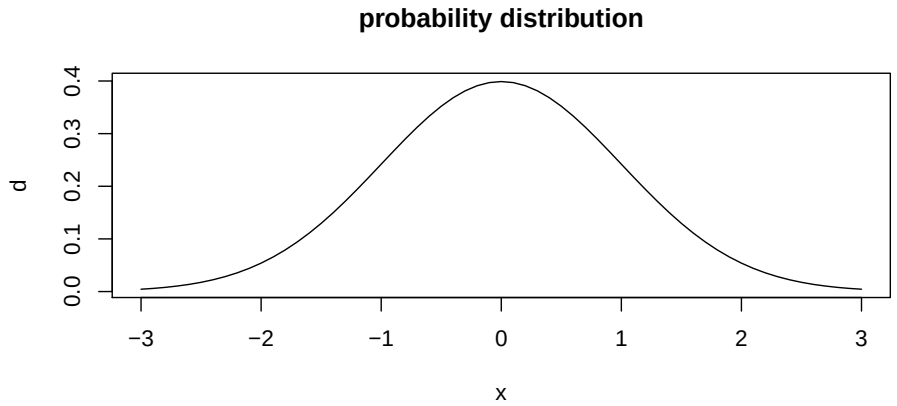
```
> x <- seq(-3,3,0.1)
> d0 <- dnorm(x, 0, 1)
> d <- dnorm(x, 1, 1)
> layout(matrix(c(1,2,3), byrow=TRUE))
> plot(x, d0, type="l", lty=2,
+      main="probability distribution")
> lines(x, d, type="l")
> set.seed(1)
> y1 <- rnorm(10, 1, 1)
> mean(y1)

[1] 1.132203

> set.seed(10)
> y2 <- rnorm(10, 1, 1)
> mean(y2)

[1] 0.5093432

> plot(x, d, type="l",
+      main="samples 1 and sample mean")
> rug(y1, 0.1, 1, 10, 'red')
> abline(v=mean(y1), lwd=5, col="red")
> plot(x, d, type="l",
+      main="samples 2 and sample mean")
> rug(y2, 0.1, 1, 10, 'blue')
> abline(v=mean(y2), lwd=5, col="blue")
```

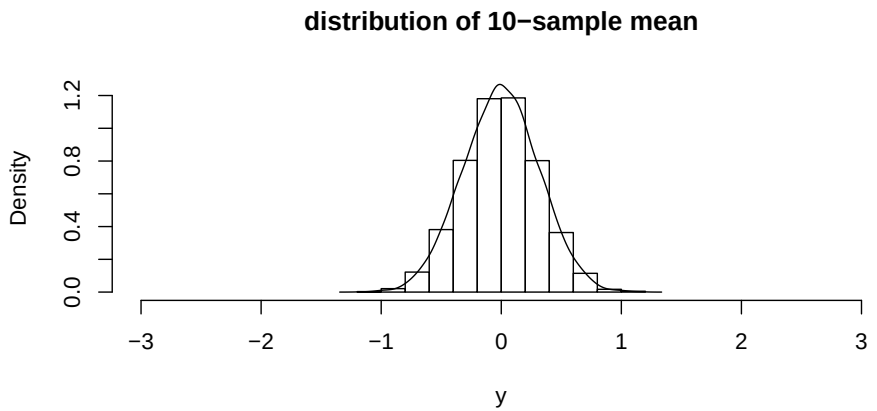


확률분포의 모수 추정 예 (계속 2)

❑ normal 분포의 평균 추정이 계속 반복된다면?

```
> x <- seq(-3,3,0.1)
> d <- dnorm(x, 0, 1)
> layout(matrix(c(1,2), byrow=TRUE))
> plot(x, d, type="l", , xlim=c(-3,3),
+      main="probability distribution")
> set.seed(1)
> y <- double(10000)
> for (i in 1:10000) {
+   y[i] <- mean(rnorm(10, 0, 1))
+ }
> hy <- hist(y, freq=FALSE, plot=FALSE)
> dy <- density(y, bw = "sj")
> ym <- max(c(hy$density, dy$y))
> hist(y, xlim=c(-3,3), ylim=c(0, ym),
+   freq=FALSE,
+   main="distribution of 10-sample mean")
> lines(dy)
```

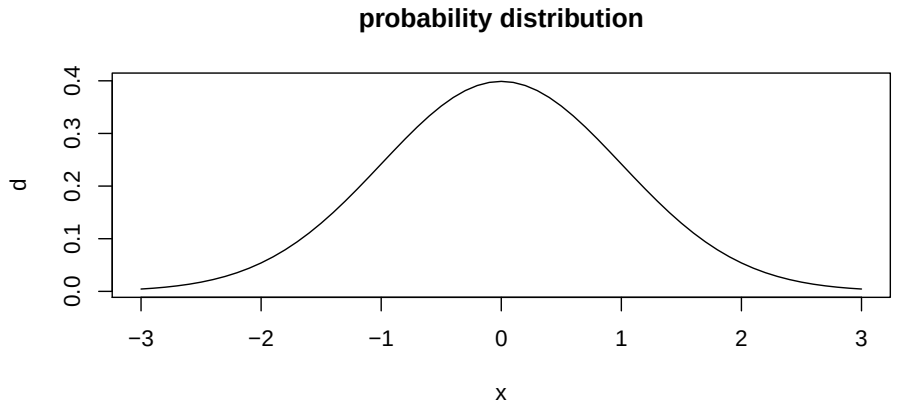


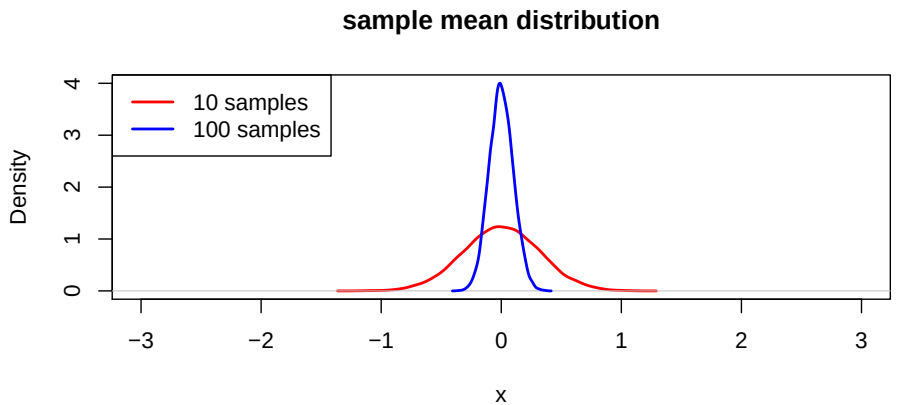


확률분포의 모수 추정 예 (계속 3)

❑ 샘플의 갯수가 달라진다면?

```
> x <- seq(-3,3,0.1)
> d <- dnorm(x, 0, 1)
> layout(matrix(c(1,2), byrow=TRUE))
> plot(x, d, type="l", , xlim=c(-3,3),
+      main="probability distribution")
> set.seed(1)
> y1 <- double(10000)
> y2 <- double(10000)
> for (i in 1:10000) {
+   y1[i] <- mean(rnorm(10, 0, 1))
+   y2[i] <- mean(rnorm(100, 0, 1))
+ }
> dy1 <- density(y1, bw = "sj")
> dy2 <- density(y2, bw = "sj")
> ym <- max(c(dy1$y, dy2$y))
> plot(dy1,
+   xlim=c(-3,3), ylim=c(0, ym),
+   col="red", lwd=2, xlab="x",
+   main="sample mean distribution")
> lines(dy2,
+   xlim=c(-3,3), ylim=c(0, ym),
+   col="blue", lwd=2)
> legend("topleft",
+        col=c("red", "blue"),
+        lwd=c(2,2),
+        c("10 samples",
+          "100 samples"))
```





신뢰구간

- 구간추정 (interval estimation)
  - 모수 추정치는 확률분포를 가지는 확률변수이므로 정확한 하나의 값을 추정하는 방법 (point estimation) 보다 특정한 신뢰도를 가지는 구간을 추정하는 방법 (interval estimation) 을 주로 사용한다.
  - 추정의 신뢰도는 주로 검정통계량의 확률값을 사용한다.
- 구간추정은 검정과 반대의 과정
  - 추정된 값이 미리 정한 극단적인 값을 가지는 경우를 가정하고 이 때의 올바른 모수값이 있을 수 있는 범위를 구함
- 신뢰구간과 신뢰수준
  - 모수 추정치 $\hat{\theta}$가 $\theta_1 \leq \hat{\theta} \leq \theta_2$ 인 구간에 있을 때의 검정통계량의 확률이 α이면 $1-\alpha$를 신뢰수준 (confidence level), 그 구간을 신뢰구간 (confidence inteval) 이라고 한다.
  - 보통 신뢰수준 $1-\alpha$가 먼저 정해지고 α에 따른 critical value를 이용한 신뢰구간을 계산

가설검정

- 통계적 가설 (statistical hypothesis)
  - 모집단의 확률 분포, 모수 등에 대한 가정
- 가설검정 (hypothesis testing)
  - 모집단의 확률 분포, 모수 등에 대한 가정에 대한 논리적 판단
- 귀무가설 (null hypothesis) H_0
  - 채택 (accept)/기각 (reject) 하려는 특정한 가설
- 대립가설 (alternative hypothesis) H_a
  - 귀무가설과 반대되는 가설. 귀무가설이 채택되면 대립가설은 기각되고 반대로 귀무가설이 기각되면 대립가설은 채택된다.

기본적인 가설검정

- 주로 평균과 분산에 대한 가설을 테스트한다.
  - 예 1 : 평균이 특정한 값이다. ($H_0 : \mu = \mu_0$)
  - 예 2 : 두 평균이 같은 값이다. ($H_0 : \mu_1 - \mu_2 = 0$)
  - 예 3 : 분산이 특정한 값이다. ($H_0 : \sigma = \sigma_0$)
  - 예 4 : 두 분산이 같은 값이다. ($H_0 : \frac{\sigma_1}{\sigma_1} = 1$)
- 기본적으로 정상분포 (normal distribution) 을 대상
  - 평균 검정의 경우에는 CLT(Central Limit Theorem) 에 의해 샘플의 갯수가 큰 경우 (보통 $n > 25$) 에는 정상분포와 같은 방법을 사용할 수 있다.

고급 가설검정

검정 이름	R 명령
runs test	`runs`
binomial test	`binom.test`
chi-square goodness of fit test	`chisq.test`
Kolmogorov-Smirnov goodness of fit test	`ks.test`
Shapiro-Wilk normality test	`shapiro.test`
correlation test	`cor.test`
Wilcox-Mann-Whitney rank sum test	`wilcox.test`
Kruskal-Walls test	`kruskal.test`
Friedman test	`freidman.test`

가설의 채택 기준

- 귀무가설의 채택 기준
  - 귀무가설이 존재하면 그 귀무가설하에서의 특정한 검정통계량의 확률분포가 결정됨
  - 귀무가설이 맞다는 가정하에 결정된 검정통계량의 확률분포하에서, 샘플에서 계산한 검정통계량과 같은 값 혹은 그보다 더 희귀한 값이 나올 확률이 미리 정한 기준치보다 낮으면 기각. 아니면 채택
- one-tailed / two-tailed 검정 방식
  - one-tailed test : 같은 부호이면서 더 희귀한 검정통계량 값이 나오는 경우 기각
  - two-tailed test : 부호와 상관없이 크기가 더 희귀한 검정통계량 값이 나오는 경우 기각

검정통계량 (test statistic)

- 모수 추정치와 마찬가지로 샘플집합 $\{x_i; i = 1, \cdots, n\}$의 함수로 계산되는 수치
- 검정통계량 (test statistic) 계산의 목적
  - 가설 검정시 가설이 맞는지 틀린지를 확인
  - 모수 추정시 추정의 정확도 분석
- 모수 추정치와 마찬가지로 검정통계량은 특정한 확률분포를 가지는 확률변수이고 다음 다음 값에 의존한다.
  - 원래 관심을 가진 확률변수의 모수
  - 샘플의 갯수 n
- 단순한 모수추정시에는 모수 추정치를 정규화한 함수(값)를 사용
  - 평균 추정시 : 샘플 평균 $\sim$ t-statistics
  - 분산 추정시 : 샘플 분산 $\sim$ chi-squared statistics
- 복잡한 모수추정이나 가설검정시에는 모수 추정치와 다른 함수 사용
  - 회귀분석의 계수추정시 : F-statistics 사용

예 : 분산값을 알고 있는 normal 분포 $N(\mu, \sigma)$의 평균

- 분산값을 알고 있는 normal 분포 $N(\mu, \sigma)$의 N개의 샘플의 평균값은 다음과 같은 normal 분포를 따름

$$\sum_{i=1} Nx_i \sim N\left(\mu, \frac{\sigma}{\sqrt{N}}\right) \tag{3}$$

- 만일 귀무가설이 $H_0 : \mu = \mu_0$라고 하면 이때의 검정통계량(샘플 평균값)의 확률분포는

$$N\left(\mu_0, \frac{\sigma}{\sqrt{N}}\right) \tag{4}$$

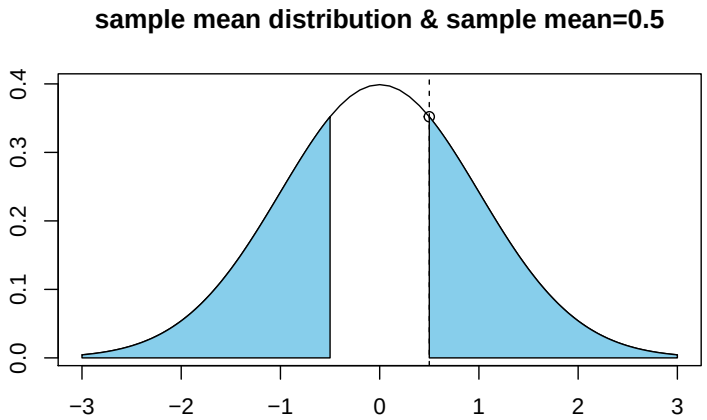
- 이때 실제로 나온 샘플에 대한 평균값이 $\bar{\mu}$라고 하면 $\bar{\mu}$ 혹은 그보다 더 희귀한 값 즉, $\bar{\mu}$보다 크거나 $-\bar{\mu}$보다 작은 값이 나올 확률은

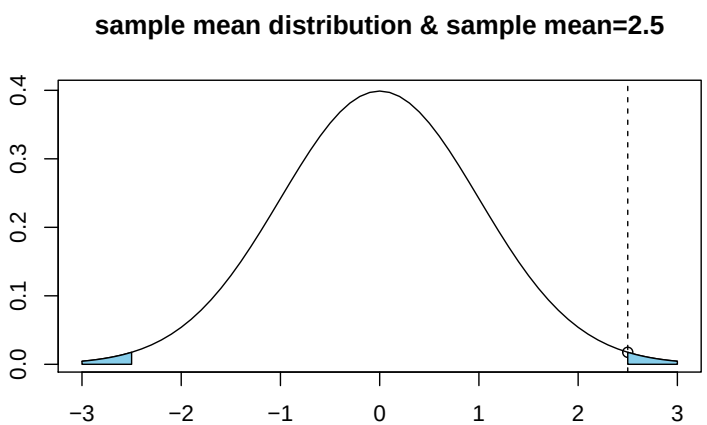
$$1 - c.d.f\left(\mu_0, \frac{\sigma}{\sqrt{N}}\right) + c.d.f\left(-\mu_0, \frac{\sigma}{\sqrt{N}}\right) \tag{5}$$

- 이 값이 미리 정해놓은 어떤 값보다 작으면 $H_0 : \mu = \mu_0$라고 보기에는 너무 희귀한 값이 나온 셈이므로 기각

예 : 분산값을 알고 있는 normal 분포 $N(\mu, \sigma)$의 평균 (계속)

```
> x <- seq(-3,3,0.1)
> d <- dnorm(x, 0, 1)
> layout(matrix(c(1,2), byrow=TRUE))
> plot(x, d, type="l", xlim=c(-3,3),
+  main="sample mean distribution & sample mean=0.5",
+  xlab="", ylab="")
> 2 * pnorm(-0.5,0,1)

[1] 0.6170751

> abline(v=0.5, lty=2)
> points(0.5, dnorm(0.5), pch=1)
> cord.x <- c(0.5,seq(0.5,3,0.01),3)
> cord.y <- c(0, dnorm(seq(0.5,3,0.01)), 0)
> polygon(cord.x,cord.y,col='skyblue')
> cord.x <- c(-0.5,seq(-0.5,-3,-0.01),-3)
> cord.y <- c(0, dnorm(seq(-0.5,-3,-0.01)), 0)
> polygon(cord.x,cord.y,col='skyblue')
> plot(x, d, type="l", xlim=c(-3,3),
+  main="sample mean distribution & sample mean=2.5",
+  xlab="", ylab="")
> 2 * pnorm(-2.5,0,1)

[1] 0.01241933

> abline(v=2.5, lty=2)
> points(2.5, dnorm(2.5), pch=1)
> cord.x <- c(2.5,seq(2.5,3,0.01),3)
> cord.y <- c(0, dnorm(seq(2.5,3,0.01)), 0)
> polygon(cord.x,cord.y,col='skyblue')
> cord.x <- c(-2.5,seq(-2.5,-3,-0.01),-3)
> cord.y <- c(0, dnorm(seq(-2.5,-3,-0.01)), 0)
> polygon(cord.x,cord.y,col='skyblue')
```





p-value, 유의수준

- p-value
  - 귀무가설이 맞다는 가정하에 결정된 검정통계량의 확률분포하에서,
    샘플에서 계산한 검정통계량과 같은 값 혹은 그보다 더 희귀한 값이 나올 확률
  - 아주 작으면 기각, 충분히 크면 채택
- significance level (유의수준) α
  - 귀무가설 채택의 기준이되는 p-value의 값
  - 보통 0.05(5%) 혹은 0.01(1%) α 값을 사용
  - 1종 오류를 범할 확률
- critical-value
  - p-value가 미리정한 기준치 즉, 유의수준 α보다 커지는 검정통계량의 값

검정오류

- Type-1 Error (1종 오류)
  - 귀무가설이 맞음에도 불구하고 너무 희귀한 검정통계량이 나오는 바람에 맞는 귀무가설을 기각하는 오류
  - 1종 오류를 범할 확률은 유의수준 α와 동일
- Type-2 Error (2종 오류)
  - 귀무가설이 틀림에도 불구하고 귀무가설하의 확률분포에서 있을 법한 검정통계량이 나오는 바람에 틀린 귀무가설을 채택하는 오류
  - 1종 오류를 범할 확률을 Power of Test(검정력)라 부른다.

R에서 지원하는 기본적인 가설검정

검정 이름	목적	R 명령
one-sample z-test	분산값이 알려진 정상분포의 평균에 대한 가설 검정	(없음)
one-sample t-test	분산값을 모르는 정상분포의 평균에 대한 가설 검정	`t.test`
paired t-test	분산값을 모르는 정상분포의 차이의 평균에	`t.test`
two-sample t-test for equal-variances	분산값이 같은 두 정상분포의 평균의 차이에 대한 가설 검정	`t.test`
two-sample t-test for unequal-variances	분산값이 다른 두 정상분포의 평균의 차이에 대한 가설 검정	`t.test`
chi-squared test	정상분포의 분산에 대한 가설 검정	(없음)
Two-sample F test	두 정상분포의 분산의 비율에 대한 가설 검정	`var.test`

샘플 평균의 분포

- 분산 σ가 알려진 정상분포 $N(\mu, \sigma)$의 경우
  - 샘플 평균 $\hat{\mu}$을 정규화 (standardization) 한 z-statistics는 표준 정상분포 (standard normal distribution)을 따른다.

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{6}$$

$$\text{z-statistics} = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{N}} \sim N(0,1) \tag{7}$$

- 분산 σ을 모르는 정상분포 $N(\mu, \sigma)$ 혹은
  정상분포가 아닌 일반적인 분포이지만 샘플 개수가 많은 경우 $(N > 25)$
  - 샘플 평균 s을 정규화 (standardization) 한 t-statistics는 DOF = $N-1$인 student-t 분포 t_{N-1}를 따른다.

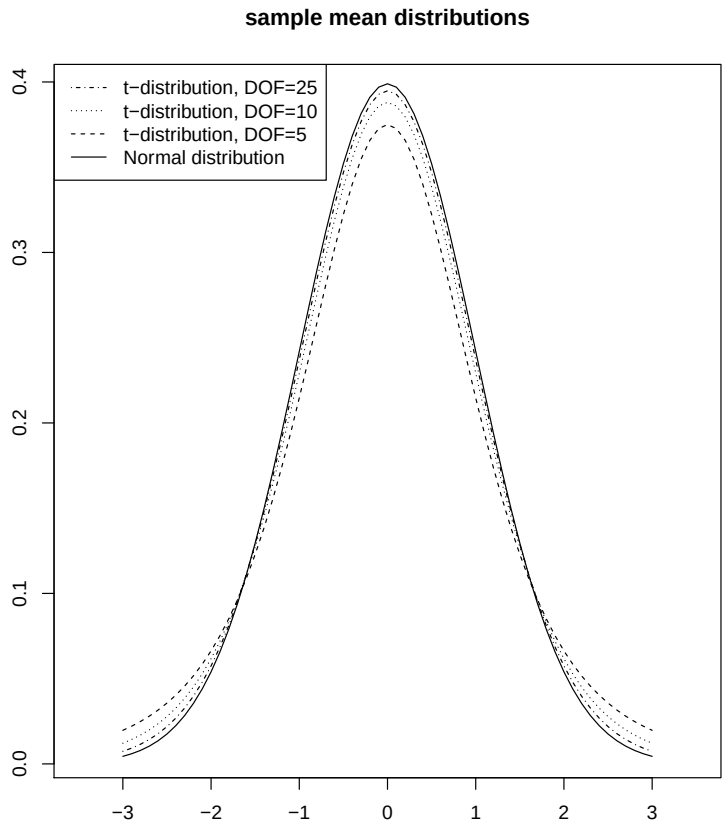
$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 \tag{8}$$

$$\text{t-statistics} = \frac{\hat{\mu} - \mu}{s^2/\sqrt{N}} \sim t_{N-1} \tag{9}$$

샘플 평균 분포의 특성

- 샘플 평균은 확률분포 평균의 unbiased 추정치
- 샘플의 갯수 N 이 증가하면 샘플 평균의 분산은 $\sqrt{N}$ 에 비례하여 감소

```
> x <- seq(-3,3,0.1)
> d1 <- dnorm(x, 0, 1)
> d2 <- dt(x, 5-1)
> d3 <- dt(x, 10-1)
> d4 <- dt(x, 25-1)
> plot(x, d4, type="l",
+  xlim=c(-3.5,3.5), lty=4,
+  main="sample mean distributions",
+  xlab="", ylab="")
> lines(x, d3, lty=3)
> lines(x, d2, lty=2)
> lines(x, d1, lty=1)
> legend("topleft",
+        c("t-distribution, DOF=25",
+          "t-distribution, DOF=10",
+          "t-distribution, DOF=5",
+          "Normal distribution"),
+        lty=4:1)
```



평균의 구간 추정 (신뢰수준 $1-\alpha$)

- 분산 σ가 알려진 정상분포 $N(\mu, \sigma)$의 경우

$$\hat{\mu} - z_{1-0.5\alpha}\sigma/\sqrt{N} < \mu < \hat{\mu} + z_{1-0.5\alpha}\sigma/\sqrt{N} \tag{10}$$

  - 위 식에서 $z_{1-0.5\alpha}$는 정상분포의 누적확률분포가 $1-0.5\alpha$가 되는 값

- 분산 σ을 모르는 정상분포 $N(\mu, \sigma)$ 혹은
  정상분포가 아닌 일반적인 분포이지만 샘플 개수가 많은 경우 $(N > 25)$

$$\hat{\mu} - t_{N-1,1-0.5\alpha}s/\sqrt{N} < \mu < \hat{\mu} + t_{N-1,1-0.5\alpha}s/\sqrt{N} \tag{11}$$

  - 위 식에서 $t_{N-1,1-0.5\alpha}$는 자유도 $N-1$인 student-t분포의 누적확률분포가 $1-0.5\alpha$가 되는 값

평균 가설검정 (유의수준 α)

- 귀무가설 $H_0 : \mu = \mu_0$
- 분산 σ가 알려진 정상분포 $N(\mu, \sigma)$의 경우
  - z-statistics에 대한 p-value가 0.5α보다 작으면 기각
- 분산 σ을 모르는 정상분포 $N(\mu, \sigma)$ 혹은
  정상분포가 아닌 일반적인 분포이지만 샘플 개수가 많은 경우 $(N > 25)$
  - t-statistics에 대한 p-value가 0.5α보다 작으면 기각

R에서의 평균 구간 추정과 가설 검정

❑ `t.test(x, alternative, mu, conf.level)` 이용

- ▶ `x` : 샘플 벡터
- ▶ `alternative` : "two-sided", "less", "greater"
- ▶ `mu` : 귀무가설의 평균값
- ▶ `conf.level` : 신뢰구간 $(1-\alpha)$, 디폴트 0.95

```
> set.seed(1)
> x1 <- rnorm(10,0,1)
> t.test(x1)

One Sample t-test

data:  x1
t = 0.5356, df = 9, p-value = 0.6052
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.4261948  0.6906003
sample estimates:
mean of x
0.1322028

> x2 <- x1 + 0.7
> t.test(x2)

One Sample t-test

data:  x2
t = 3.3714, df = 9, p-value = 0.008239
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.2738052 1.3906003
sample estimates:
mean of x
0.8322028
```

문제 3의 접근 방법

문제 3 : 기대 수익률 비교

삼성전자 주가의 한 달 평균 수익률은 4.0% 이고 현대차 주가의 한 달 평균 수익률은 3.9% 이다. 삼성전자 주가의 기대수익률은 현대차 주가의 기대수익률보다 높다고 할 수 있나?

- ❏ 삼성전자 주가의 일간 수익률과 현차대 주가의 일간 수익류은 두 개의 서로 다른 확률분포에서 생성된 확률변수라고 가정. 또 이 확률분포들은 미래에도 변하지 않을 것이라고 가정.
- ❏ 이 확률분포의 평균값은 삼성전자 주가의 기대수익률이다.
- ❏ 이 두 개의 확률분포에서 뽑은 두 개의 샘플 집합에서 계산한 평균차이는 어떤 확률분포를 가지는가?

독립적인 두 샘플의 평균의 차이의 분포

- 두 개의 정상분포 $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$에서 독립적으로 뽑은 N_1, N_2 개의 샘플 대상
- 샘플 평균의 차이를 정규화한 t-statistics는 student-t 분포
- 두 분포의 분산이 같은 경우 ($\sigma_1 = \sigma_2$)

$$\text{DOF} = N_1 + N_2 - 2 \tag{12}$$

$$\text{t-statistics} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{s^2}{N_1} + \frac{s^2}{N_2}}} \sim t_{\text{DOF}} \tag{13}$$

$$s^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2} \tag{14}$$

- 두 분포의 분산이 다른 경우 ($\sigma_1 \neq \sigma_2$)

$$\text{DOF} = \frac{(s_1^2/N_1 + s_2^2/N_2)^2}{(s_1^2/N_1)^2/N_1 + (s_2^2/N_2)^2/N_2} \tag{15}$$

$$\text{t-statistics} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_1^2}{N_2}}} \sim t_{\text{DOF}} \tag{16}$$

R에서 독립적인 two-sample 평균 가설검증

❑ `t.test(x, y, var.equal)` 이용

► `var.equal` : TRUE면 두 분포의 분산이 같은 경우. 디폴트는 FALSE

```
> set.seed(1)
> x1 <- rnorm(10,0,1)
> x2 <- rnorm(10,0,1)
> t.test(x1, x2, var.equal=TRUE)

Two Sample t-test

data:  x1 and x2
t = -0.2786, df = 18, p-value = 0.7837
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9963182  0.7630338
sample estimates:
mean of x mean of y
0.1322028 0.2488450

> x1 <- rnorm(10,0,1)
> x2 <- rnorm(10,0,2)
> t.test(x1, x2)

Welch Two Sample t-test

data:  x1 and x2
t = -0.6315, df = 14.602, p-value = 0.5374
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6442119  0.8939446
sample estimates:
 mean of x  mean of y
-0.1336732  0.2414604
```

paired 샘플의 차이의 평균의 분포

- 두 개의 정상분포 $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$에서 동시에 뽑은 N개의 샘플 대상
- 샘플값의 차이의 평균을 정규화한 t-statistics는 $N-1$자유도의 student-t 분포

$$\text{t-statistics} = \frac{\hat{\mu}_d}{s_d/\sqrt{N_1}} \sim t_{\text{DOF}} \tag{17}$$

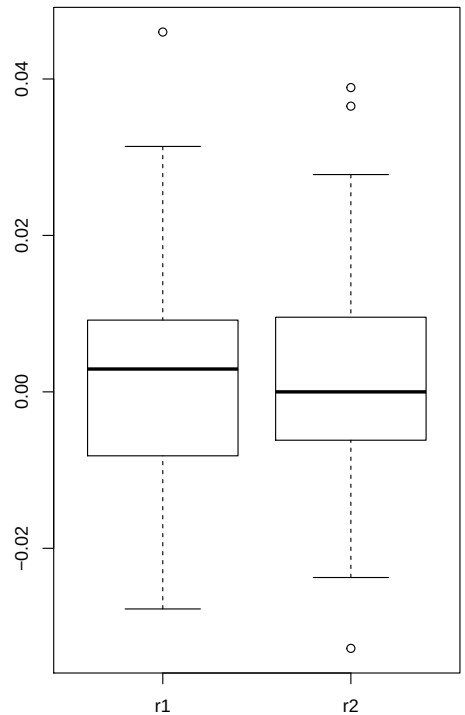
- $\hat{\mu}_d$, s_d^2는 각각 샘플차이의 평균과 분산 추정치

R에서 paired two-sample 평균 가설검증

❑ `t.test(x, y, paired)` 이용

► `paired` : `TRUE`면 paired two-sample. 디폴트는 `FALSE`

```
> require(TTR)
> require("rquantbook")
> api <- "krx_stock_daily_price"
> d1 <- "2013-08-01"
> d2 <- "2013-09-17"
> df1 <- get_quantbook_data(api, date_start=d1, date_end=d2,
+   ticker="005930")

rmongodb package (mongo-r-driver) loaded
Use 'help("mongo")' to get started.

> df2 <- get_quantbook_data(api, date_start=d1, date_end=d2,
+   ticker="005380")
> r1 <- na.omit(ROC(df1$close))
> r2 <- na.omit(ROC(df2$close))
> boxplot(cbind(r1, r2))
> t.test(r1, r2, paired=TRUE)

Paired t-test

data:  r1 and r2
t = -0.2471, df = 31, p-value = 0.8065
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.006647324  0.005210639
sample estimates:
mean of the differences
          -0.0007183428
```



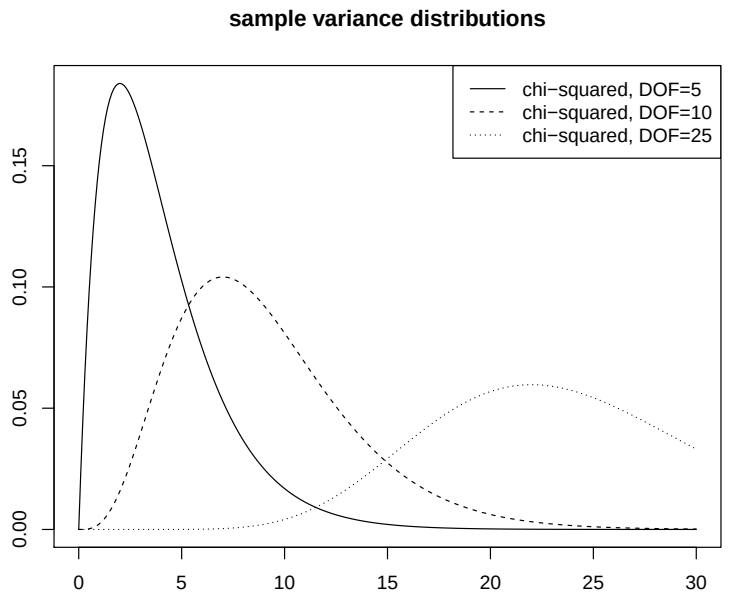
샘플 분산의 분포

❑ 정상분포 $N(\mu, \sigma)$의 경우

► 샘플 분산 s은 다음과 같이 scaled된 자유도 $N-1$ chi-squared 분포를 따른다.

$$s \sim \frac{\sigma^2}{N-1}\chi^2_{N-1} \tag{18}$$

```
> x <- seq(0,30,0.1)
> d1 <- dchisq(x, 5-1)
> d2 <- dchisq(x, 10-1)
> d3 <- dchisq(x, 25-1)
> plot(x, d1, type="l", lty=1,
+  main="sample variance distributions",
+  xlab="", ylab="")
> lines(x, d2, lty=2)
> lines(x, d3, lty=3)
> legend("topright",
+        c("chi-squared, DOF=5",
+          "chi-squared, DOF=10",
+          "chi-squared, DOF=25"),
+        lty=1:3)
```



분산의 구간 추정 (신뢰수준 $1-\alpha$)

❑ 정상분포 $N(\mu, \sigma)$의 경우

$$\frac{(N-1)s}{\chi^2_{N-1,1-0.5\alpha}} < \sigma^2 < \frac{(N-1)s}{\chi^2_{N-1,0.5\alpha}} \tag{19}$$

- ▶ 위 식에서 $\chi^2_{N-1,0.5\alpha}$는 $N-1$ 자유도 chi-squared 분포의 누적확률분포가 0.5α가 되는 값

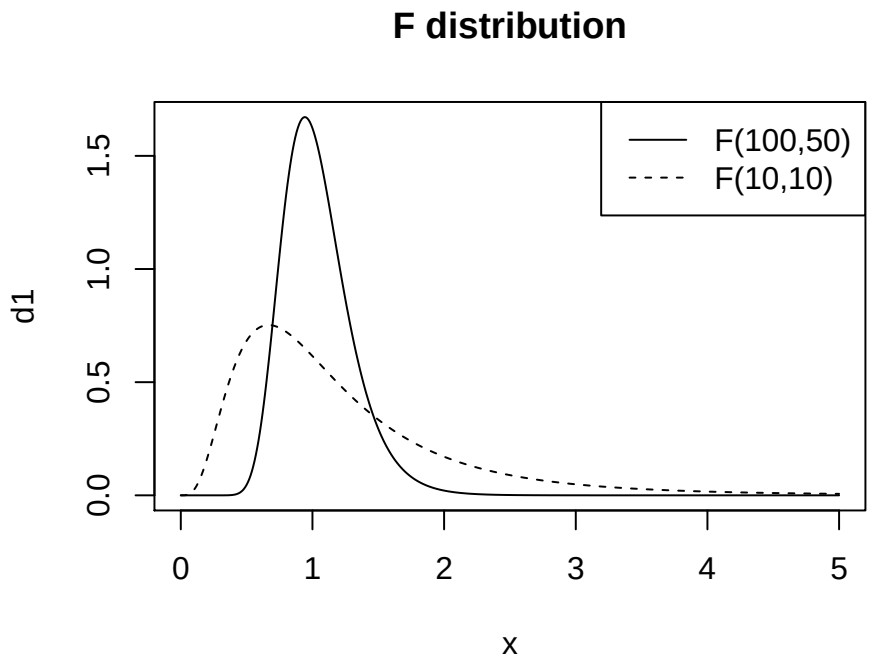
```
> x <- rnorm(10, 0, 1)
> var(x)

[1] 0.3549372

> 9*var(x)/qchisq(1-0.5*0.05,9)

[1] 0.1679269

> 9*var(x)/qchisq(0.5*0.05,9)

[1] 1.182953
```

독립적인 두 샘플의 분산의 비율의 분포

- 두 개의 정상분포 $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$에서 독립적으로 뽑은 N_1, N_2 개의 샘플 대상
- 샘플 분산의 비율를 정규화한 F-statistics는 $(N_1 - 1, N_2 - 1)$ 자유도의 F 분포

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{N_1-1, N_2-1} \tag{20}$$

```
> x <- seq(0,5,0.01)
> d1 <- df(x, 100, 50)
> d2 <- df(x, 10, 10)
> plot(x, d1, type="l", xlim=c(0,5),
+      main="F distribution")
> lines(x, d2, xlim=c(0,5), lty=2)
> legend("topright", lty=1:2,
+        c("F(100,50)",
+          "F(10,10)"))
```



분산의 비율의 구간 추정 및 검정

❑ 구간 추정

$$\frac{1}{F_{\text{DOF},1-0.5\alpha}}\frac{s_1^2}{s_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_{\text{DOF},0.5\alpha}}\frac{s_1^2}{s_2^2} \tag{21}$$

- ▶ DOF=$(N_1 - 1, N_2 - 1)$

❑ 검정 : $H_0 : \sigma_1 = \sigma_2$

$$\frac{s_1^2}{s_2^2} \sim F_{\text{DOF}} \tag{22}$$

❑ `var.test(x, y, conf.level)` 이용

- ▶ `x`, `y` : 샘플 벡터

```
> set.seed(1)
> x1 <- rnorm(50, 0, 1)
> x2 <- rnorm(100, 0, 1)
> var.test(x1, x2)

F test to compare two variances

data:  x1 and x2
F = 0.7822, num df = 49, denom df = 99, p-value =
0.343
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4897019 1.3014987
sample estimates:
ratio of variances
         0.7822426
```