推測統計(検定)

Code

inferential statistics

推測統計

少ない data から大きな集団の特徴を掴む

母集団と**標本**と**サンプリング**

↓

推定・検定

母集団と標本

- **母集団** → 情報を得たい対象全体
- **標本** → 母集団の一部

標本抽出のサンプリング

母集団	**標本**
母平均： μ	標本平均： $\overline{X}$
母分散： σ^2	標本分散： s^2
母標準偏差： σ	標本標準偏差： s

仮説検定

ある仮説が偶然起こった事なのか **統計学的** に判断する方法

推定と同様に **標本分布** を元に考える

↓

仮説

検定の考え方

> 1. 帰無仮説 ： H_0

- 本当は **対立仮説** を言いたいが, 逆の仮説を立てる

> 2. 対立仮説 ： H_1

- **本来自分が言いたい仮説**

> 3. 有意水準の決定(危険率, 棄却率)

- **よくある事と珍しい事の%**
  - 珍しい事が起これば **帰無仮説** が棄却 **対立仮説** を採択
  - よくある事が起これば **対立仮説** が棄却 **帰無仮説** を採択
  - 帰無仮説 H_0 : **正規分布の 95% の範囲に入るはず**
  - 対立仮説 H_1 : **正規分布の 5% の範囲に入るはず**

上記の仮説 (検定) の仕方は　**片側検定**

↓

正規分布のグラフの片側の面積

違う仮説の立て方て　**両側検定**　もある

↓

正規分布のグラフの両側の面積

$$標準化 \quad = \quad \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

仮定

　近所の酒屋でお酒の量り売りをしているが, 最近お酒の量が減った気がする。１００g のお酒を 30日間買って量を測ったところ, 平均値が 98g, 標準偏差が 5g だった。

Hide

```
xb <- 98
s <- 5
m <- 100
n <- 30

z <- (xb - m)/(s/sqrt(n))
z
```

```
[1] -2.19089
```

- $\overline{X}$: 98 ｜ μ : 100 ｜ σ : 5 ｜ $\sqrt{n}$: 30

$$標準化 \quad = \quad \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- **pnorm( )** : 正規分布の面積を求める
    - マイナス記号の場合は片側検定の面積

Hide

```
pnorm(-1.64, mean = 0, sd =1)
```

```
[1] 0.05050258
```

- **qnorm( )** : 上記の%となる点を見つけてくれる

Hide

```
qnorm(0.05, mean = 0, sd = 1)
```

```
[1] -1.644854
```

推定と検定のまとめ

　統計量が異なるとそれぞれ異なる　**標本分布**　が必要

母集団のどのような統計量を求めたいのか?sample数の大きさ etc ….

↓

上記のような基準を常に意識して区間推定をする

	sample数小	**sample数大**
母平均	t分布	正規分布 (t分布)
母分散 (母標準偏差)	x^2 分布 (カイ二乗分布)	x^2 分布 (カイ二乗分布)
母比率	F分布	正規分布

t分布を使用した検定

- 同時に区間推定もしてくれる
    - **95%の信頼区間** : 95.68908 ~ 99.91539

Hide

```
x <- rnorm(30, mean = 98, sd =5)

t.test(x, conf.level = 0.95)
```

```
	One Sample t-test

data:  x
t = 94.659, df = 29, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 95.68908 99.91539
sample estimates:
mean of x 
 97.80224 
```

t分布検定

- 母平均が 100 の時に片側検定（下側）

Hide

```
t.test(x, mu = 100, alternative = "less")
```

```
	One Sample t-test

data:  x
t = -2.1271, df = 29, p-value = 0.02102
alternative hypothesis: true mean is less than 100
95 percent confidence interval:
     -Inf 99.55779
sample estimates:
mean of x 
 97.80224 
```

両側検定

Hide

```
t.test(x, mu = 100)
```

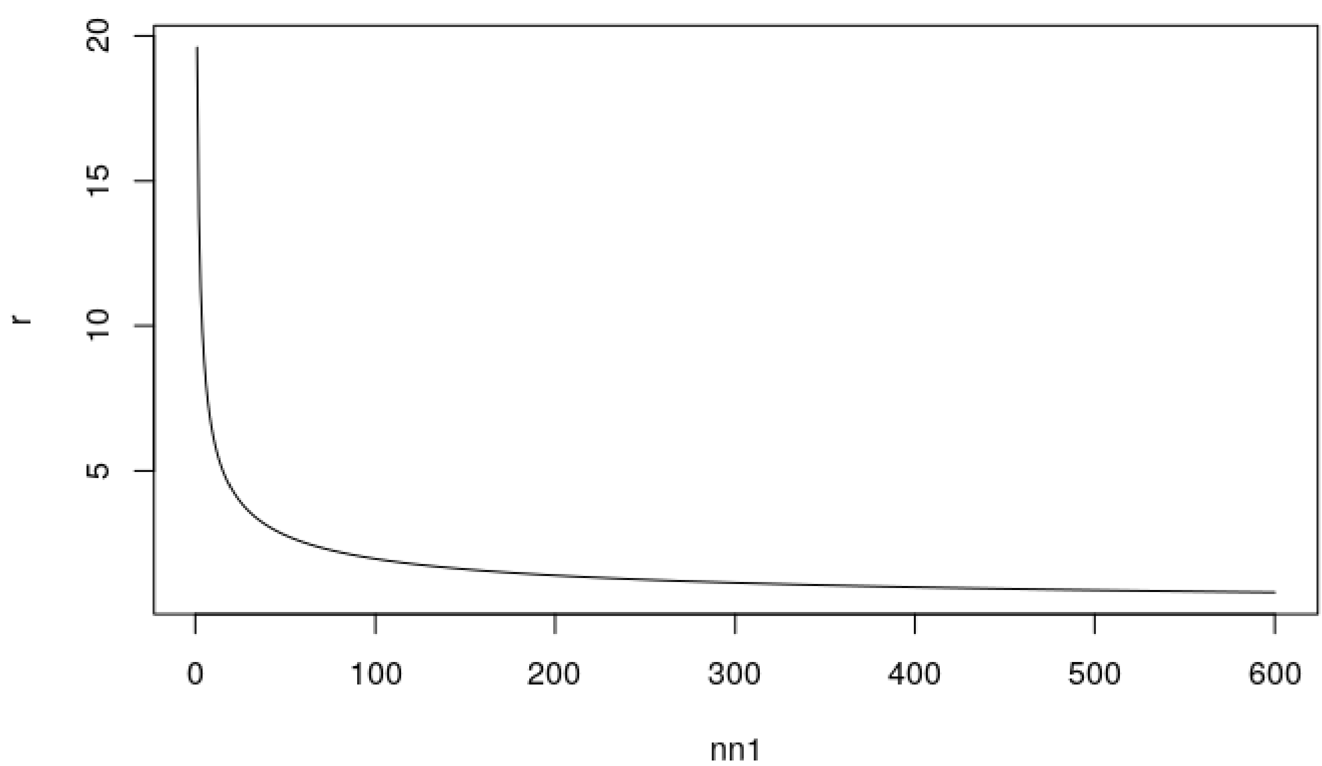
```
	One Sample t-test

data:  x
t = -2.1271, df = 29, p-value = 0.04204
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 95.68908 99.91539
sample estimates:
mean of x 
 97.80224 
```

sample数の変化と区間の関係

- **r** = 推定区間
- **nn1** = sample数

Hide

```
nn1 <- c(1:600)
r <- 1.96*(10/sqrt(nn1))
plot(nn1, r, type = "l")
```

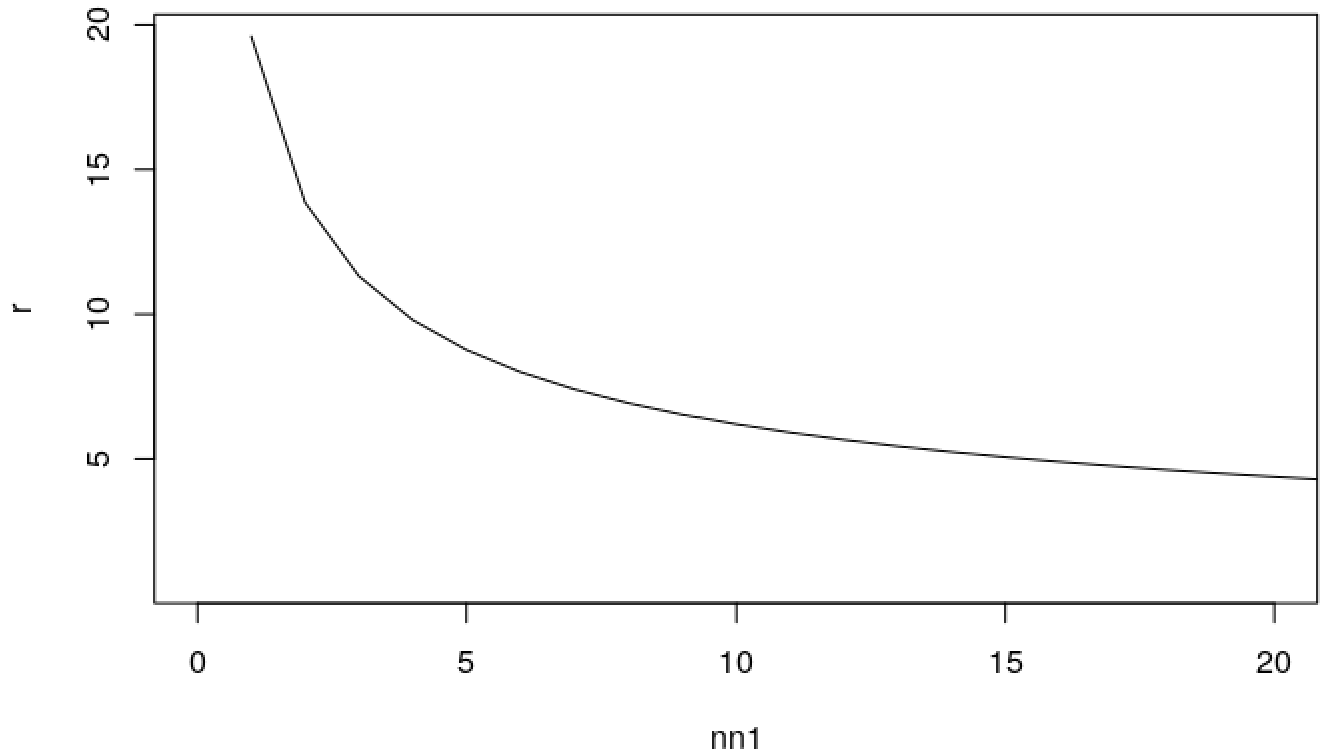


sample数が増えるに従って推定区間が

plot 拡大

Hide

```
plot(nn1, r, type = "l", xlim = c(0, 20))
```



ある程度母集団の標準偏差が推定される場合は

sample数を求める為に区間を基準に決めることが出来る