

9. Statistics 統計学

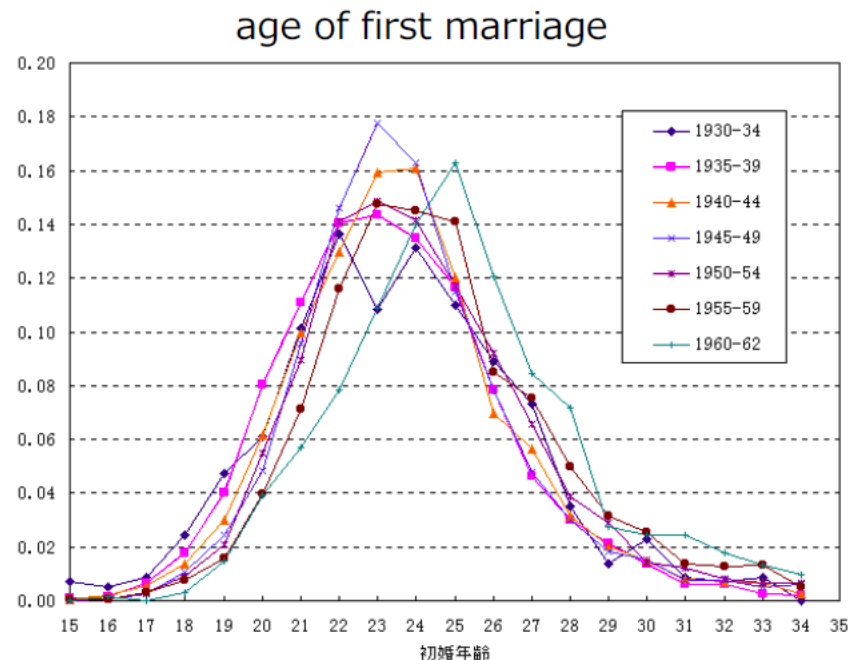
- Mean and variance 平均と分散
- Expected value 期待値
- Models of probability events 確率事象モデル

Statistics 統計学

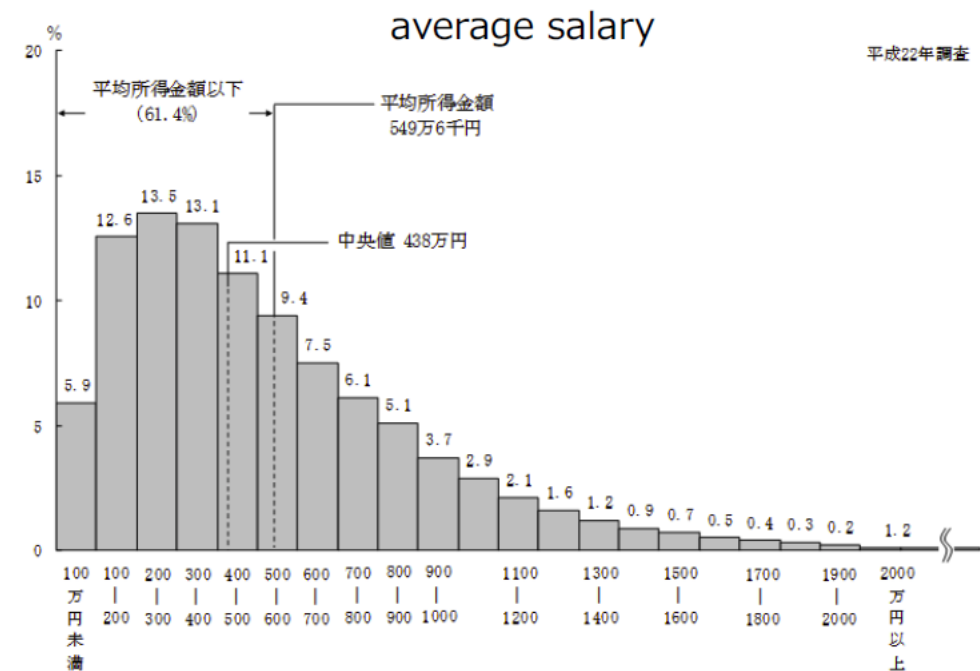
- ある分散データ(distributed data)について考える

例) 初婚年齢, 日本人の所得

- このような分布データを1つの数字で表す場合には
平均, 中央値 (小さい順に並べたときの真ん中の値), 最頻値 (もっともデータ数の多い値)
- もう一つ数字を使えるならデータのばらつきを表したい
分散 = データのばらつきの幅を表す



<http://www.mhlw.go.jp/shingi/0112/s1211-3a.html>



<http://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa10/2-2.html>

統計計算

平均:mean

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

中央值:median

```
>> X = randn(10000,1);  
>> mean(X)  
ans = 0.0034172  
>> var(X)  
ans = 1.0268  
  
>> X = rand(10000,1);  
>> mean(X)  
ans = 0.50384  
>> var(X)  
ans = 0.083720
```

分散:var

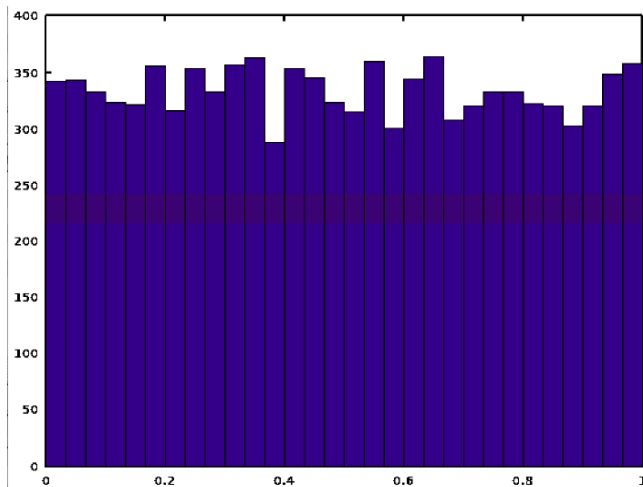
不偏標本分散

$$V = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2$$

標準偏差:std $\sigma = \sqrt{V}$

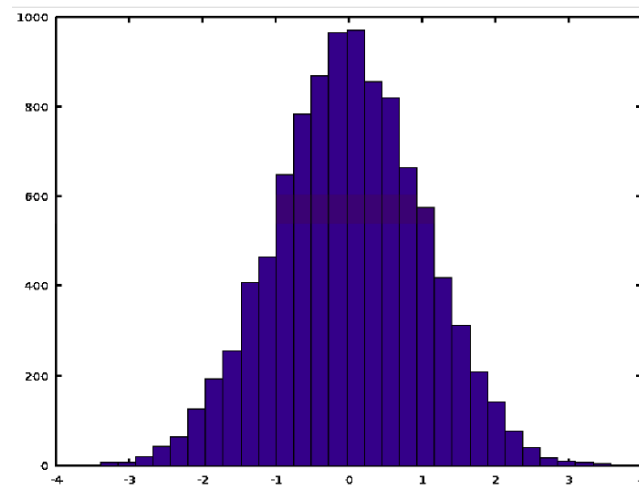
```
>> X = randn(10000,1);  
>> std(X)  
ans = 0.99576  
>> sqrt(var(X))  
ans = 0.99576  
>> median(X)  
ans = -0.0051996
```

```
>> hist(rand(10000,1),30)
```



rand 0 から 1 までの
均一な分布を持つよう
なランダム数値を発生
させる

```
>> hist(randn(10000,1),30)
```



randn 平均0 分散 1
の分布を持つような
ランダム数値を発生
させる

```
>> X = randn(10000,1);  
>> mean(X)  
ans = 0.0034172  
>> var(X)  
ans = 1.0268  
  
>> X = rand(10000,1);  
>> mean(X)  
ans = 0.50384  
>> var(X)  
ans = 0.083720
```

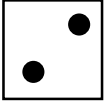
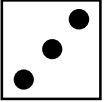
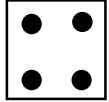
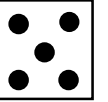
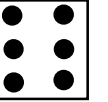
```
>> X = randn(10000,1);  
>> std(X)  
ans = 0.99576  
>> sqrt(var(X))  
ans = 0.99576  
>> median(X)  
ans = -0.0051996
```

期待値(Expected value)

- 確率変数 X の期待値 $E[X]$

$$E[X] = \sum_{i=1}^{\infty} x_i P(X = x_i)$$

例: 6面サイコロを振って、出た目の数×1000円貰えるゲーム
このゲームの期待値は？

確率変数 X	x_1  1000円	x_2  2000円	x_3  3000円	x_4  4000円	x_5  5000円	x_6  6000円
確率 $P(X=x_i)$	1/6	1/6	1/6	1/6	1/6	1/6

$$E(X) = 1000 \times 1/6 + 2000 \times 1/6 + 3000 \times 1/6 + 4000 \times 1/6 + 5000 \times 1/6 + 6000 \times 1/6 = 3500$$

期待値(Expected value)

- 前のスライドのサイコロゲームをモンテカルロ法で計算

```
>> X=rand(10000,1);
```

```
>> Y=floor(X*6)+1;
```

```
>> mean(Y*1000)  
ans =3445.2
```

X: ランダム(0から1の間) な数列

Y: サイコロの出た目

Xの値を6倍して

$0 < 6 * X < 1$ $Y = 1$

$1 \leq 6 * X < 2$ $Y = 2$

⋮

$5 \leq 6 * X < 6$ $Y = 6$

サイコロの出た目Y×1000したあとで
10000回分の平均値を取る

2つの異なる分散

- 母分散(population variance) $V = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$
- 標本分散(sample variance)

- 完全なデータ列から抽出した標本を基にした分散

例) 日本人の身長データを取るときに, ランダムにN人のデータを選ぶ

3 ページ目の分散は以下のように定義されている

不偏分散: $V = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2$ **N-1で割っていることに注意**

これは母分散の真の値を正しく見積もることができる

Nで割った値は真の値（母分散）と一致しないことが統計的に知られている

2つの異なる分散

- 10個の標本から正規分布の分散を計算する.
- それを10,000回行い平均値を出す.

randn関数をつかうので真の分散の値は1であるはずだが・・・

Nで割った場合

$$V = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

```
>> X=randn(10,10000);  
>> m=mean(X);  
>> Y=mean((X-ones(10,1)*m).^2);  
>> mean(Y)  
ans =0.90047
```

1からずれている

標本分散を使う場合

$$V = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2$$

```
>> X=randn(10,10000);  
>> m=mean((var(X))  
ans =1.0005
```


確率のモデル (1) : ポアソン分布

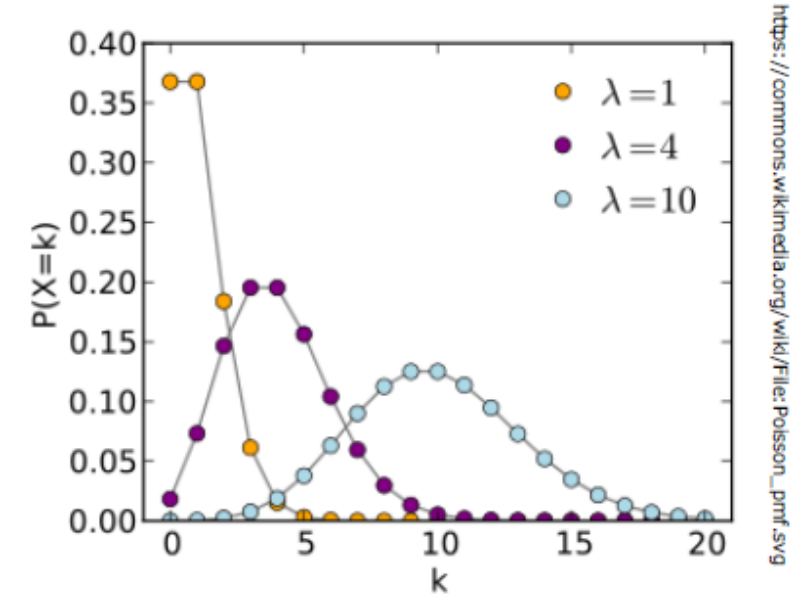
- ある固定された期間内に平均的に λ 回発生する事象を考える

例) 30分の間にe-mailを受け取る回数

- この期間にk回の事象が起きる確率は

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{で与えられる}$$

- Xの期待値は $E[X] = \lambda$
- これはポアソン分布と呼ばれる
- これはポアソン分布を持つようなランダム数を発生させる場合
randp(l,m,n)を使う $l=\lambda$ 期待値 λ のポアソン分布を持つ数値を $m \times n$ 行列に代入



```
>> randp(4,1,10)
```

```
ans=
```

```
7    3    4    4    6    4    5    4    3    3
```

```
>> hist(randp(4,1,10000))
```

確率のモデル (2) : 2 項分布

- コインを n 回投げる; n 回の試行のうち表が出る回数を X
(毎回のコイントスはそれ以前の結果とは独立)
- p は表が出る確率とすると $X=k$ となる確率は次のように与えられる

$$P[X = k] = \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{for } k = 0, 1, 2, \dots, n$$

$$\binom{n}{k} = \frac{n \times (n - 1) \dots \times (n - k + 1)}{k \times (k - 1) \dots \times 1} \quad \text{nchoosek}(n,k)$$

- X の期待値は $E[X] = np$
- これは 2 項分布と呼ばれる. 表記の仕方は $B(n,p)$

確率のモデル (2) : 2 項分布

例) 表になる確率 $p=40\%$ のコインで10回投げる

X : 2項分布 $B(10, 0.4)$

```
>> X=rand(1,10)<0.4;
ans=0  0  0  1  0  1  1  1  1  0
>> sum(X)
ans=5
```

ランダムな数値が0.4未満であれば1
0.4以上の場合は0
=> 1:表 0:裏

これを10000回行ったときの平均は

```
>> Y=sum(rand(10,10000)<0.4);
>> mean(Y)
ans=4.0098
```

$$E[X]=np=10*0.4=4.0$$

2項分布の使用例

- 裏側になった5種類のカードがある。ランダムに1枚選んだカードがどの模様のカードか予想する。

10回行って6回正解した場合、超能力者といえるか？



- 超能力者でないなら；毎回正しい答えを選ぶことができるかは完全にランダムである。つまり $p=1/5$ の確率で正解を選ぶ。
- 正解を選ぶ回数は2項分布 $B(10, p)$
- 10回中 k 回正解する確率 $p(X=k)$ $k=1, 2, 3, \dots$ を計算するには

```
>> for k=0:10, nchoosek(10,k)*0.2^k*(1-0.2)^(10-k), end
```

```
ans=0.10737          k=1
```

```
ans=0.26844          k=2
```

```
ans=0.30199
```

```
ans=0.088080
```

```
ans=0.026424
```

```
ans=0.0055050
```

```
ans=7.8643e-004
```

$k=6$

6回正解できる確率は0.6%程度

のかなり低い確率=>超能力者のようだ

Exercise 9.1

- ある国のある地域では、昔から1日の平均にして0.7回の地震が起きることが知られている
- しかし、直近の過去4週間で29回の地震があった
- 4週間で29回以上地震が起きる確率を計算せよ

Hint: ある期間の平均回数を表すモデルはポアソン分布 $\Rightarrow$ randpを使う