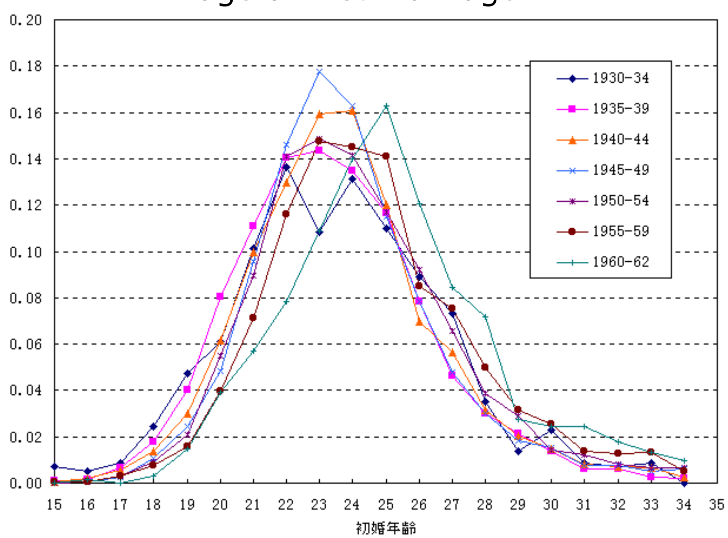
9. Statistics I

- Mean and variance
- Expected value
- Models of probability events

Statistic(s)

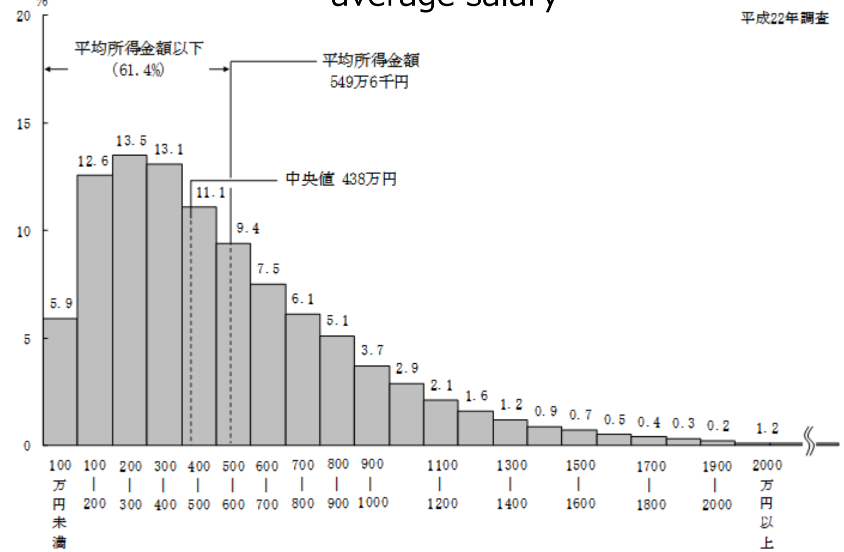
- Consider a set of distributed data (values)
  - E.g., age of first marriage and average salary of Japanese
- If we use only a single value to describe the data, we may choose
  - mean, median (the value separating the higher half of the data from the lower half), mode (the value that appears most often)
- If we can use one more value, we may want to represent dispersion of the data
  - variance = the width of dispersion of data

age of first marriage



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

average salary



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

Computation of statistics

- mean: `mean`

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

- median : `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
```

- variance : `var`
  - called *unbiased sample variance*

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

- standard deviation : `std`

$$\sigma = \sqrt{V}$$

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

Two different variances*

- Population variance
  - Defined for a set of *N* data: $$V = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad \cdots (*)$$
- Sample variance
  - Defined with *N* data that are samples chosen from a complete set of data
    - E.g., The case when we consider *height of Japanese* using randomly chosen *N* (say, =1000) persons
  - The definition in the last page gives an estimate of the true population variance of the complete set of data
    - If it is divided by *N* (not by *N*-1), then its expectation does not coincide with the true value (i.e., population variance of height of all Japanese)

Consider estimating the true variance (σ^2=1.0) of standard normal distribution using ten samples randomly drawn from it; this is repeated for 100,000 trials and the average of the 100,000 estimates are evaluated

When Eq (*) (divided by *N*) is used:

```
>> X = randn(10,100000);
>> m = mean(X);
>> Y = mean((X - ones(10,1)*m).^2);
>> mean(Y)
ans =  0.90047
```

When sample variance is used:

```
>> X = randn(10,100000);
>> mean(var(X))
ans =  1.0005
```

Expected value (of a random variable)

- Expected value of a (discrete) random variable X is defined to be

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

- Consider a game in which you roll a six sided die and you win (the number shown on the face of the die) × 1,000 JPY; how much money can you get paid for this game?
  - The expected value of the income gives an answer

$$E[X] = 1000 \times \frac{1}{6} + 2000 \times \frac{1}{6} + 3000 \times \frac{1}{6} + 4000 \times \frac{1}{6} + 5000 \times \frac{1}{6} + 6000 \times \frac{1}{6} = 3500$$

  - You can evaluate it approximately using Monte Carlo simulation

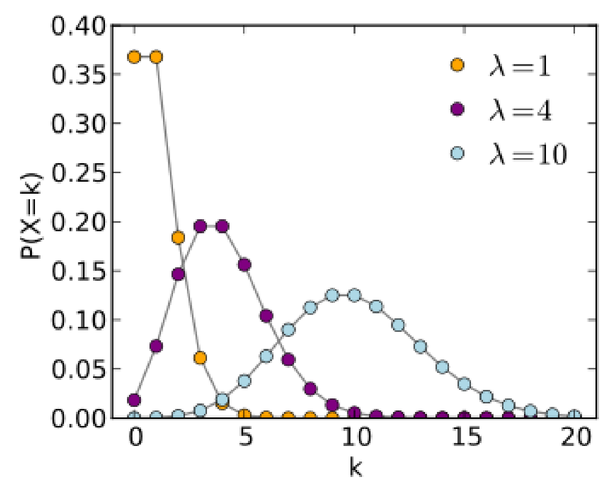
$$E[X] \approx \frac{1}{N} \sum_{n=1}^{N} X_n$$

```
>> X=rand(10000,1);
>> Y=floor(X*6)+1;
>> mean(Y*1000)
ans =  3445.2
```

Model of probability events: Poisson distribution

- Consider events that will happen λ times in a fixed interval of time in an average sense
  - E.g., E-mails received in thirty minutes
- Probability that k events occurs in this time interval is given by

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



https://commons.wikimedia.org/wiki/File:Poisson_pmf.svg

  - Expected value of X: $E[X] = \lambda$

- This is called Poisson distribution
  - Random numbers distributed with a Poisson distribution are generated by `randp(l,m,n)`, where $l=\lambda$ and $m \times n$ is the size of matrix

```
>> randp(4,1,10)
ans =
   7    3    4    4    6    4    5    4    3    3
>> hist(randp(4,1,10000))
```

Model of probability events: binomial distribution

- Consider tossing a coin *n* times; let *X* be the counts (out of *n*) for which we see the head side
  - We assume the outcome of each tossing is independent of earlier ones
- Let *p* be the probability of the head; the probability of *X*=*k* is given by

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

$$\left(\binom{n}{k}=\frac{n\times(n-1)\times\cdots\times(n-k+1)}{k\times(k-1)\times\cdots\times 1} \longrightarrow \texttt{nchoosek(n,k)}\right)$$

  - Expected value of *X*: $E[X]=np$
- This is called binomial distribution and denoted by *B*(*n*,*p*)

X's distributed with *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
```

Average of 10,000 *X*'s:

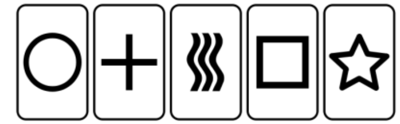
```
>> Y=sum(rand(10,10000)<0.4);
>> mean(Y)
ans =   4.0098
```

$$\updownarrow$$

$$E[X]=np$$

Example use of binomial distribution

- Consider predicting a card randomly chosen from the five cards on the right when they are face down; when you do this prediction *ten* times, *six* of them are correct



- Can you declare that you are a psychic?
- Let's calculate the probability that six out of ten are correct
  - Suppose you are *not* a psychic; then it will be completely random whether or not you can make a correct prediction at each trial; its probability is a constant p=1/5=0.2
  - The number X of correct predictions will distribute with $B(10,p)$
  - Thus, $p(X=k)$ for $k=1,2,3,\cdots$ is calculated as follows:

```
>> for k=0:10, nchoosek(10,k)*0.2^k*(1-0.2)^(10-k), end
ans =  0.10737   k=0
ans =  0.26844   k=1
ans =  0.30199   k=2
ans =  0.20133
ans =  0.088080
ans =  0.026424
ans =  0.0055050   k=6
ans =   7.8643e-004
ans =   7.3728e-005
ans =   4.0960e-006
ans =   1.0240e-007
```

Assuming you are *not* a psychic, the probability of correctly predicting cards six and more times is only about 0.6%, which is a very rare event; thus it is very likely that you are a psychic!

Exercise 9.1

- A, B, C and D are the last 4 digits from your student number (see Excercise 4.1)
- In an area of a country, it is known that earthquakes occur 0.7*(A+1) times in B+1 days in an average sense since the dawn of the history
- However, there were 10+C+D earthquakes in the last four weeks
- Calculate the probability of 10+C+D and more earthquakes occurring in four consecutive weeks