Introductory Statistics
10: Distributions and Limit Theorems

Richard Veale

Graduate School of Medicine
Kyoto University

https://youtu.be/-U0heNrjL08

Lecture Video at above link

Summary

1) Some distributions:

Uniform

Binomial

Normal

2) “Expected Value”?

3) Limit Theorems:

Central Limit Theorem (CLT)

Law of Large Numbers (LLN)

Distributions

1) Some distributions:

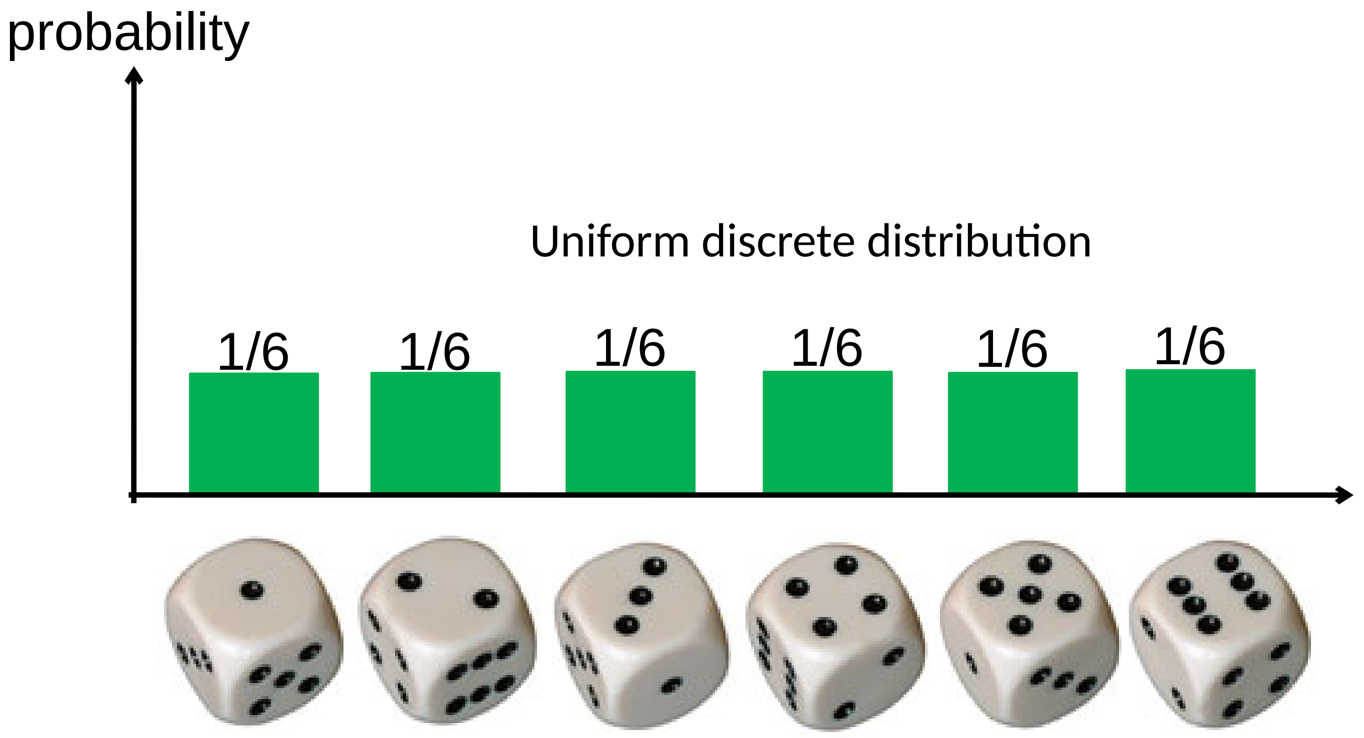
Uniform

Binomial

Normal

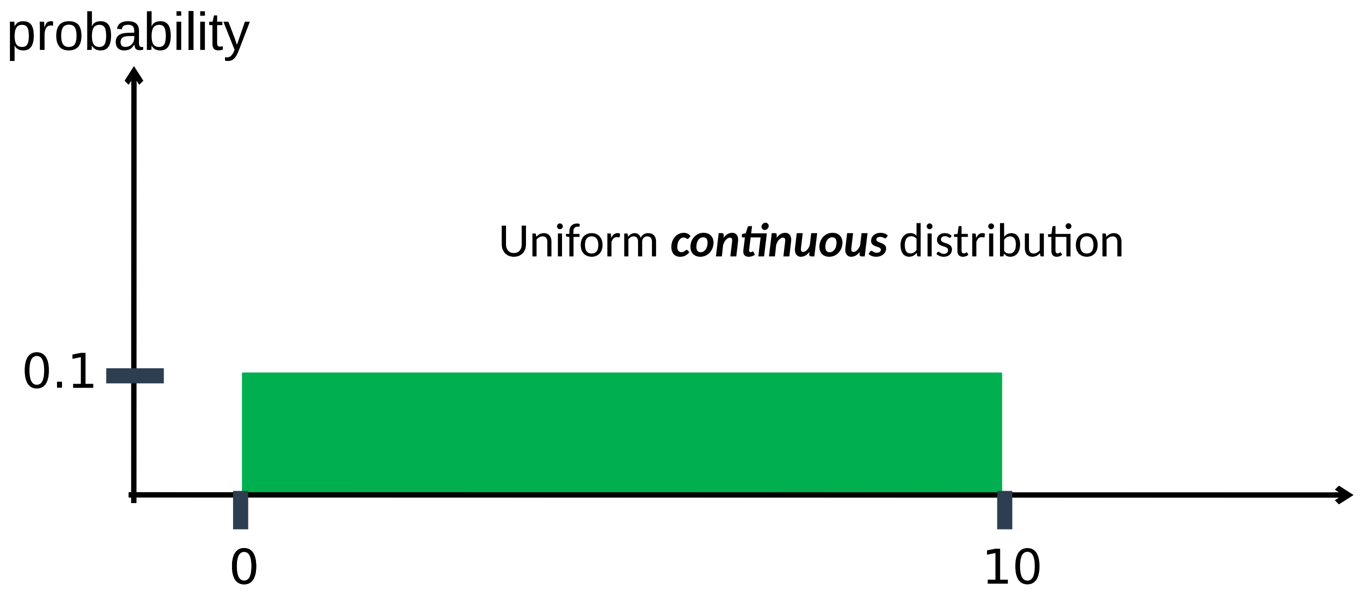
Uniform Distribution

Uniform (Equal) probability of all outcomes



Uniform Distribution

Uniform (Equal) probability of all outcomes



Equal probability of getting:
1.2333, 9.4738, 6.66666666,
0.2766666, 0.5, ...

“Expected Value”

Expected value:

$$E(X)=\mu=\sum_{x} x\cdot P(X=x)$$

Mean of random variable X

The expected value of a random variable is the value of the average which we would expect to find in a very large sample.

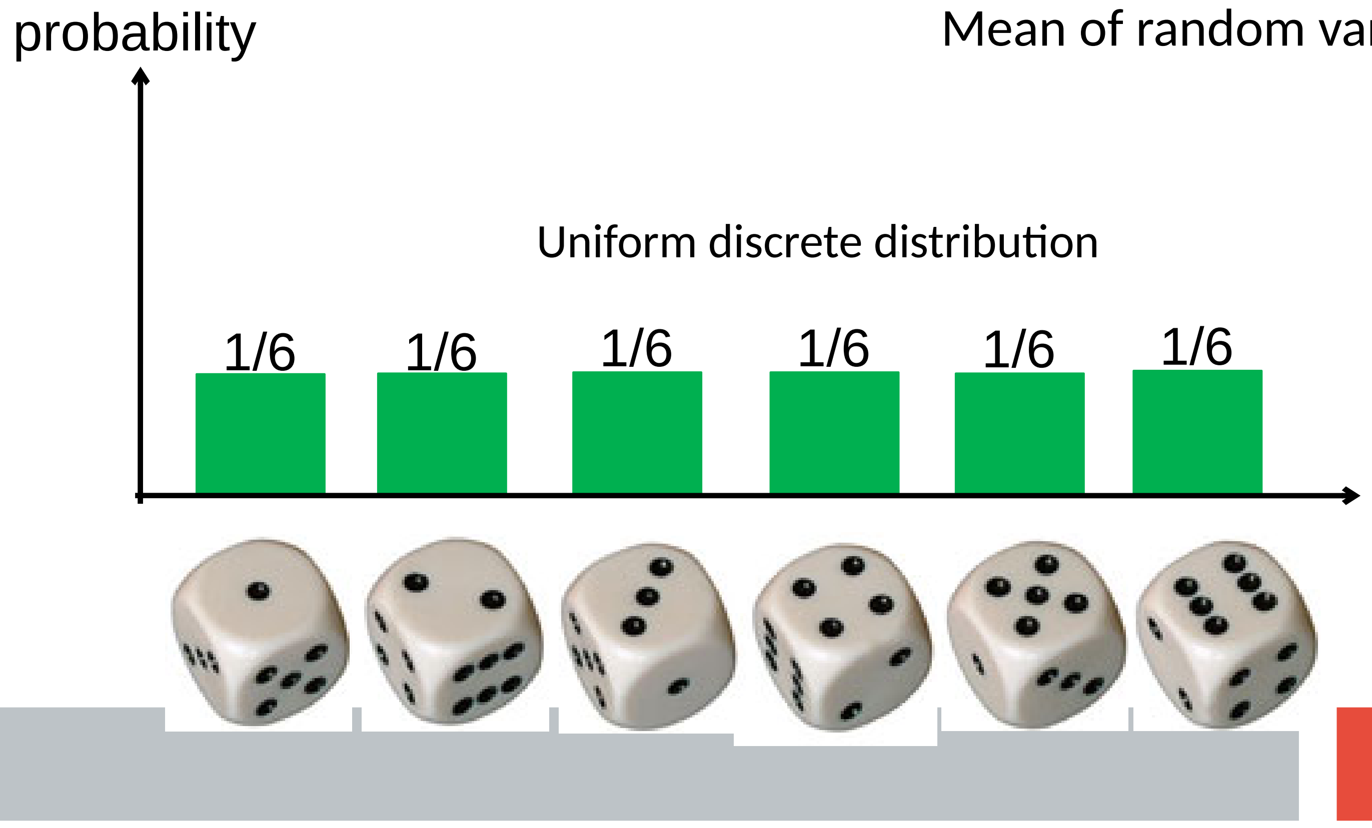
Expected value of uniform discrete

$$x \in \{1,2,3,4,5,6\}; P(X=x)=\frac{1}{6}$$

Expected value:

$$E(X)=1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+3\cdot\frac{1}{6}+4\cdot\frac{1}{6}+5\cdot\frac{1}{6}+6\cdot\frac{1}{6}=\frac{21}{6}=3.5=\mu$$

Mean of random variable X



Expected value of uniform discrete

$$x \in \{1,2,3,4,5,6\}; P(X=x) = \frac{1}{6}$$

Expected value:

$$E(X) = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = \frac{21}{6} = 3.5 = \mu$$

How about: die with “A”, “B”, “C”, “D”, “E”, “F” on the faces?

It is still a uniform discrete distribution.

But, what is the expected value?
→ You can’t do it

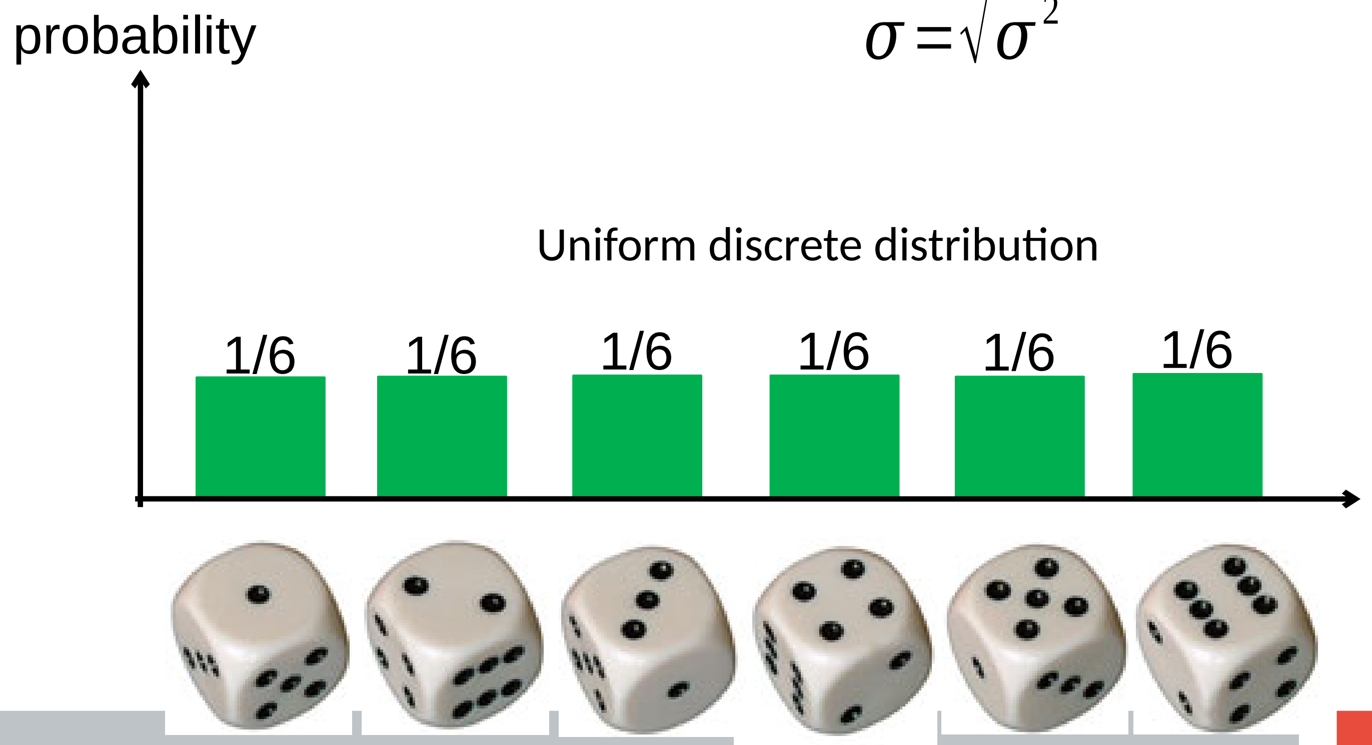
Variance of uniform discrete

Variance of X?

$$E\left[(X-\mu)^2\right]=\sigma^2=\sum_x (x-\mu)^2 \cdot P(X=x)$$

σ^2 : variance of random variable X

$$\sigma=\sqrt{\sigma^2}$$



Variance of uniform discrete

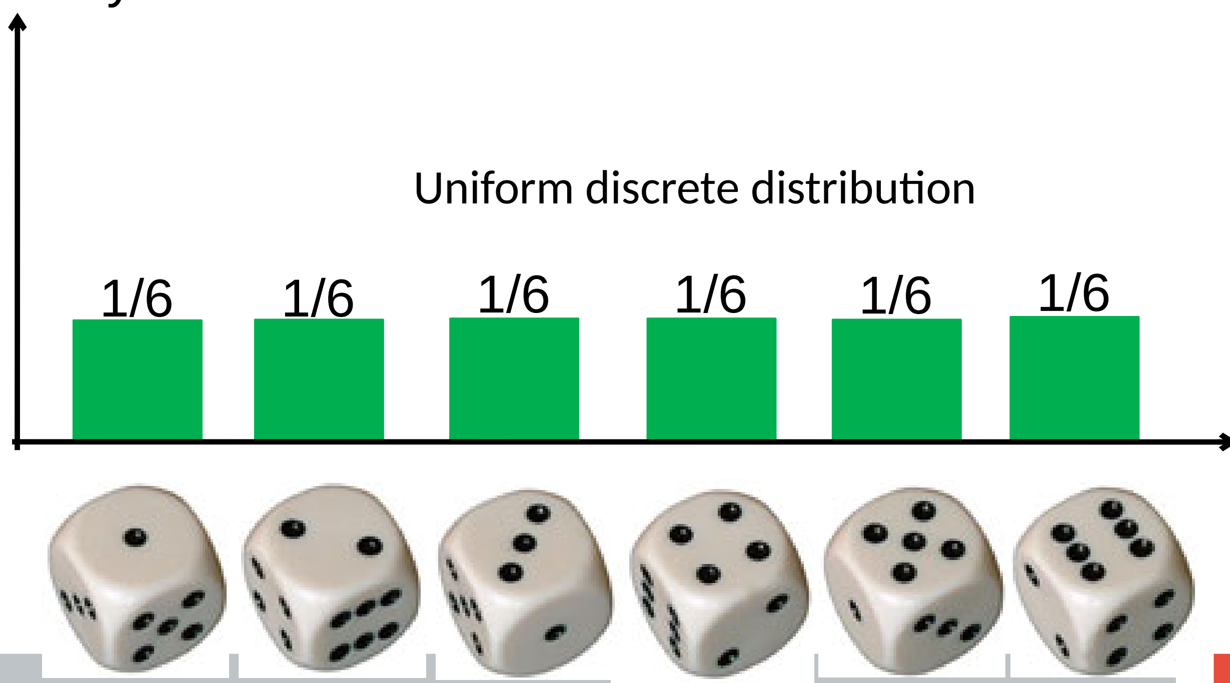
Variance of X?

$$E\left[(X-\mu)^2\right]=\sigma^2=\sum_x (x-\mu)^2\cdot P(X=x)$$

σ^2 : variance of random variable X

$$\sigma=1.708$$

probability

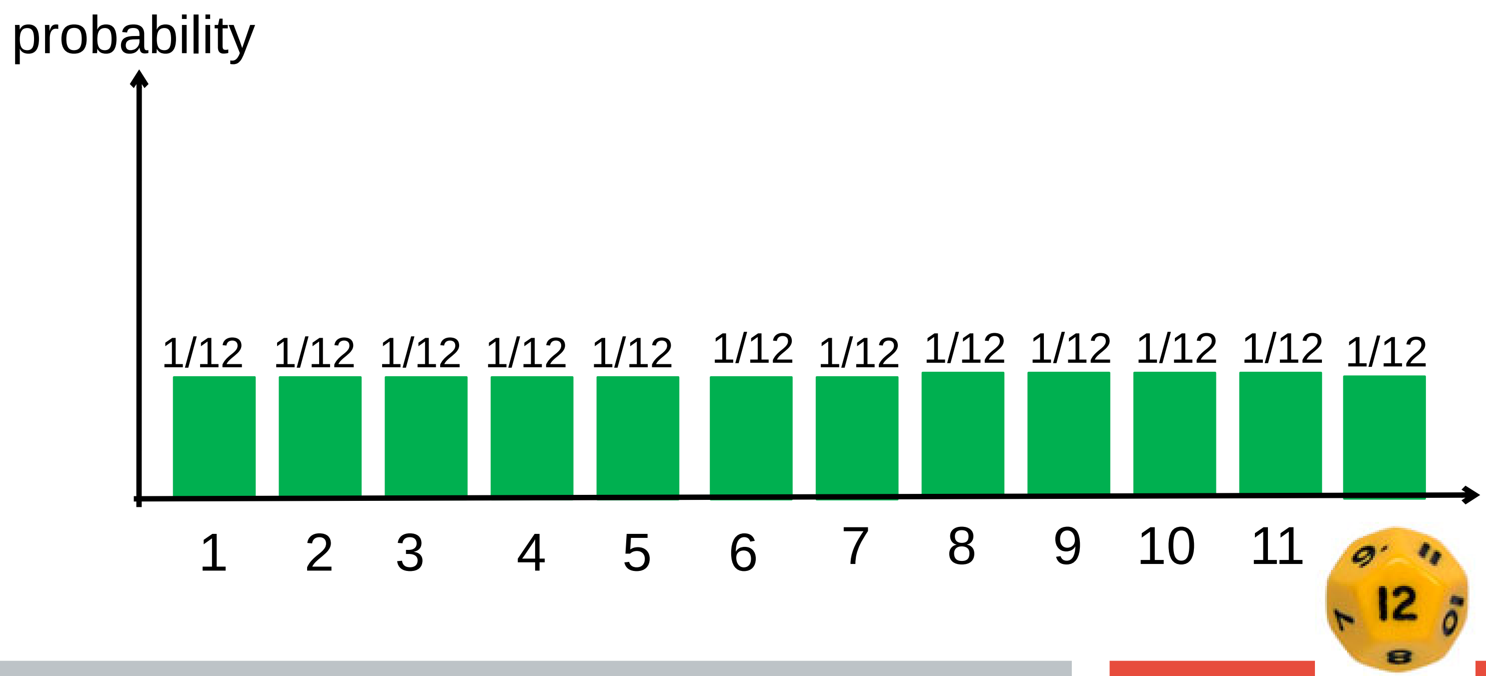


Discrete Uniform Distribution

Twelve-sided die?

Expected value: E(X) = 6.5

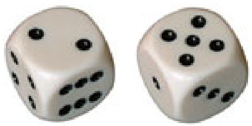
$$E\left[(X-\mu)^2\right] = \frac{143}{12} = \sigma^2 \qquad \sigma \approx 3.452$$



Sum of two random variables

Instead of a 12-sided die, how about the sum of 2 dice?

$Y = X_1 + X_2$ $\qquad P(X_1=1)=P(X_2=1)=1/6$

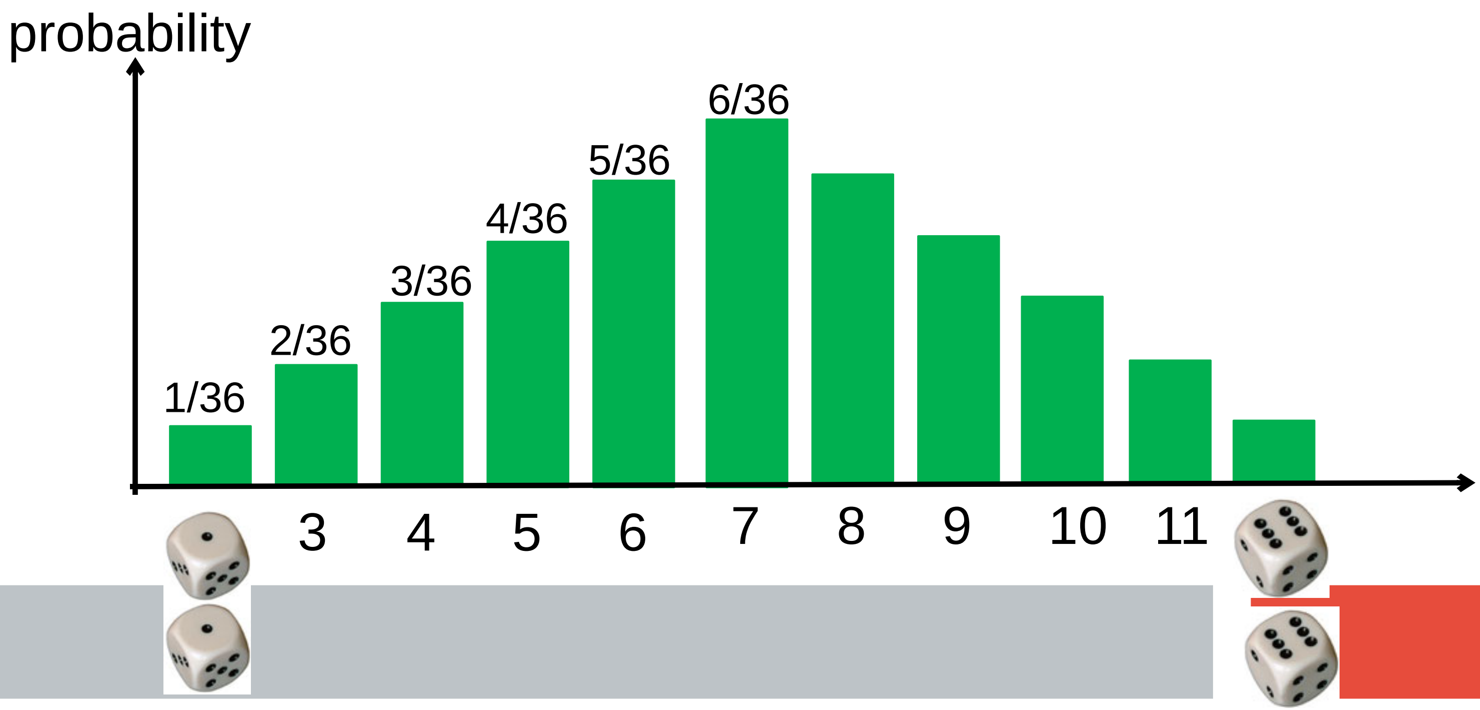


$P(Y=2) = P(X_1=1) \cdot P(X_2=1) = 1/36$

$P(Y=12) = P(X_1=6) \cdot P(X_2=6) = 1/36$

$P(Y=3) = P(X_1=1) \cdot P(X_2=2) + P(X_1=2) \cdot P(X_2=1) = 2/36$

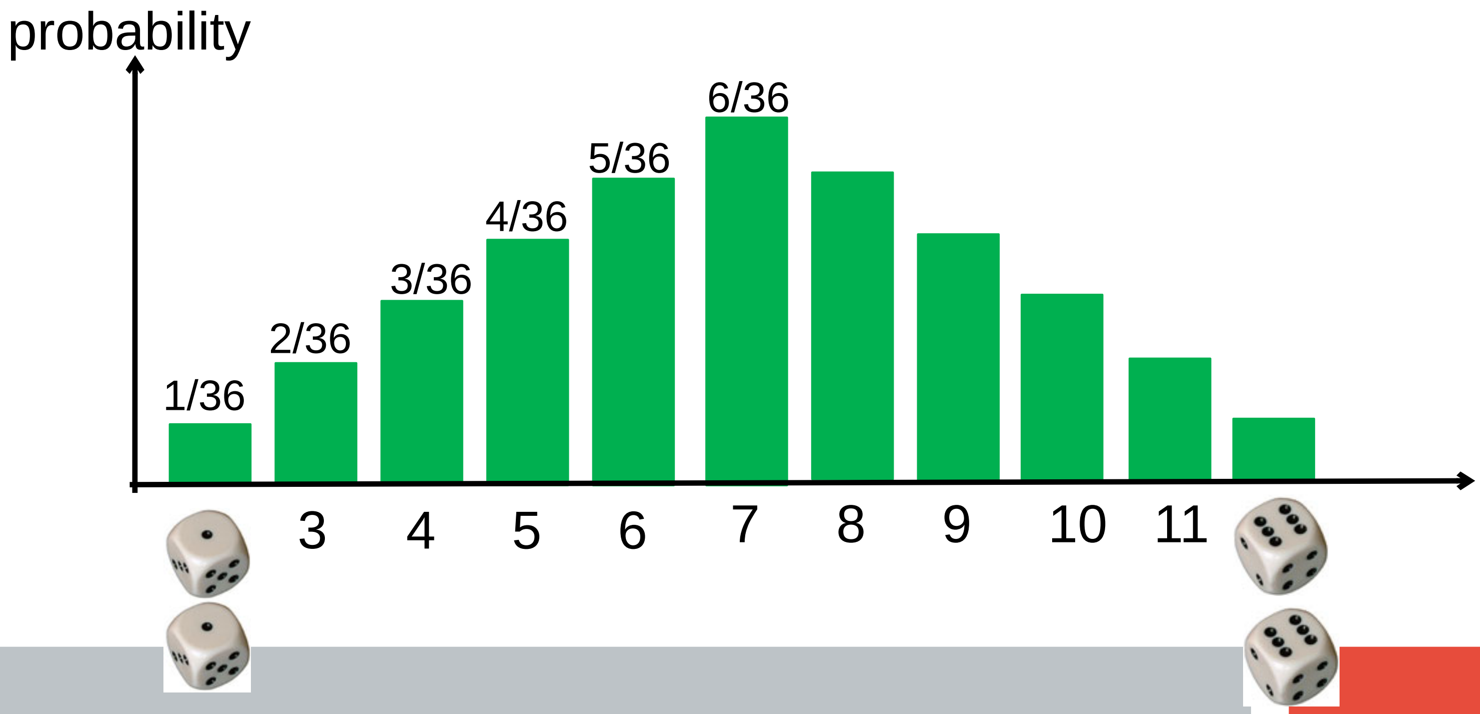
$P(Y=11) = P(X_1=5) \cdot P(X_2=6) + P(X_1=6) \cdot P(X_2=5) = 2/36$ etc.



Sum of two random variables

$$E(Y)=\frac{1\cdot 2}{36}+\frac{2\cdot 3}{36}+\frac{3\cdot 4}{36}+\frac{4\cdot 5}{36}+\frac{5\cdot 6}{36}+\frac{6\cdot 7}{36}+\frac{5\cdot 8}{36}+\frac{4\cdot 9}{36}+\frac{3\cdot 10}{36}+\frac{2\cdot 11}{36}+\frac{1\cdot 12}{36}=7=3.5+3.5=\boldsymbol{\mu 1}+\boldsymbol{\mu 2}$$

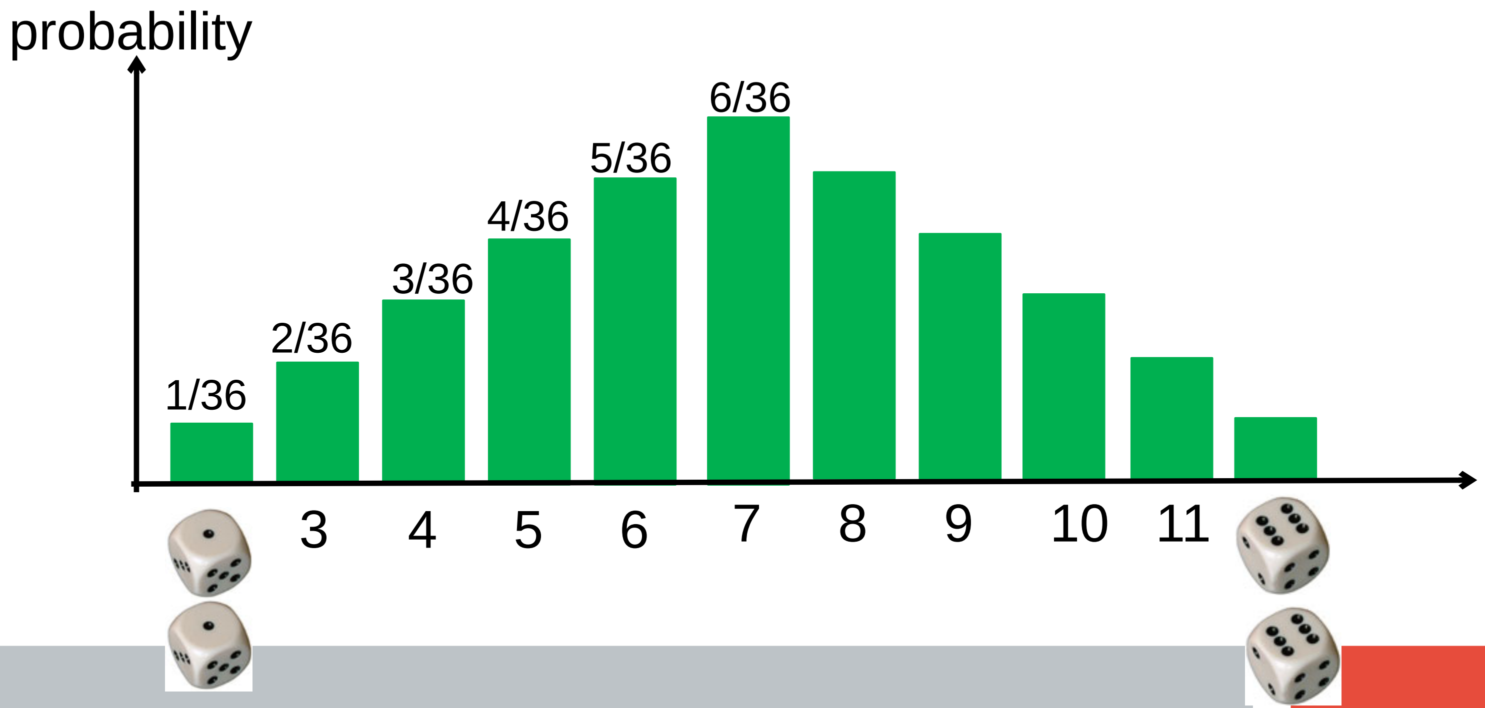
Mean sums linearly



Sum of two random variables

$$\boldsymbol{E}\left(\left(\boldsymbol{Y}-\mu\right)^{2}\right)=\mu_{1}^{2}+\mu_{2}^{2}$$

Variance also sums linearly



Expected Value: Use

Say, you built a house: Price = 15,000,000 ¥

Should you buy earthquake insurance?

Probability of an earthquake to affect your house in the next 30 years: p_E=0.05

Premiums for 30 years: 1,200,000¥ (40000¥ / year)

Insurance will pay half the initial cost of the house

Can you calculate the expected costs after 30 years in the uninsured and insured case?

Would you recommend earthquake insurance?

Expected value: Earthquake Insurance

House price = 15,000,000¥ p_E=0.05

Premiums = 1,200,000¥ Refund = 7,500,000¥

Uninsured case:

No Earthquake: Costs X = House price = 15M¥
Earthquake: Costs X = 2·House price = 30M¥

E(X) = 15M¥ · (1-p_E)+ 30M¥ · p_E = **15.75M¥**

Insured case:

No Earthquake: Costs X = House price + Premiums = 16.2M¥
Earthquake: Costs X = 2·House price + Premiums - Refund= 23.7M¥

E(X) = 16.2M ￥ (1-p_E)+ 23.7M ￥ p_E = **16.575M¥**

Earthquake Insurance…

THE TEN MOST COSTLY WORLD EARTHQUAKES AND TSUNAMIS BY INSURED LOSSES, 1980-2014 (1)

($ millions)

				Losses when occurred		
Rank	Date	Event	Location	Overall	Insured (2)	Fatalities
1	Mar. 11, 2011	Earthquake, tsunami	Japan: Aomori, Chiba, Fukushima, Ibaraki, Iwate, Miyagi, Tochigi, Tokyo, Yamagata	$210,000	$40,000	15,880
2	Feb. 22, 2011	Earthquake	New Zealand: Canterbury, Christchurch, Lyttelton	24,000	16,500	185
3	Jan. 17, 1994	Earthquake	USA: CA: Northridge, Los Angeles, San Femando Valley, Ventura, Orange	44,000	15,300	61
4	Feb. 27, 2010	Earthquake, tsunami	Chile: Concepcion, Metropolitana, Rancagua, Talca, Temuco, Valparaiso	30,000	8,000	520
5	Sep. 4, 2010	Earthquake	New Zealand: Canterbury, Christchurch, Avonside, Omihi, Timaru, Kaiapoi, Lyttelton	10,000	7,400	NA
6	Jan. 17, 1995	Earthquake	Japan: Hyogo, Kobe, Osaka, Kyoto	100,000	3,000	6,430
7	Jun. 13, 2011	Earthquake	New Zealand: Canterbury, Christchurch, Lyttelton	2,700	2,100	1
8	May 20 and May 29, 2012	Earthquake (series)	Italy: Emilia-Romagna, San Felice del Panaro, Cavezzo, Rovereto di Novi, Carpi, Concordia	16,000	1,600	18
9	Dec. 26, 2004	Earthquake, tsunami	Sri Lanka, Indonesia, Thailand, India, Bangladesh, Myanmar, Maldives, Malaysia	10,000	1,000	220,000
10	Oct. 17, 1989	Earthquake	USA: CA: Loma Prieta, Santa Cruz, San Francisco, Oakland, Berkeley, Silicon Valley	10,000	960	68

Insurance Information Institute: www.iii.org

However: Expectation versus Utility

Let’s say, I’d offer the following bet:
Your wager is 1000¥

I will flip a coin, for “Heads” I’ll give you 3000¥, for “Tails” I’ll take your 1000¥.

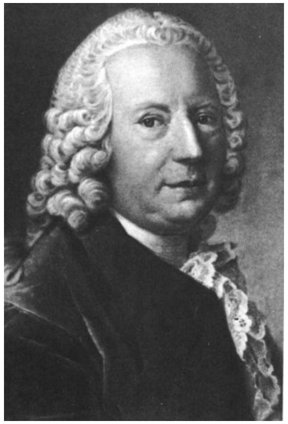
$E(X) = -1000¥ \cdot 1/2 + 3000¥ \cdot 1/2 = 1000¥$

Now, let’s say, I’d offer the following bet:
Your wager is 10,000,000¥

I will flip a coin, for “Heads” I’ll give you 30M¥, for “Tails” I’ll take your 10M¥.

$E(X) = -10M¥ \cdot 1/2 + 30M¥ \cdot 1/2 = 10M¥$

Would you do it?



Daniel Bernoulli

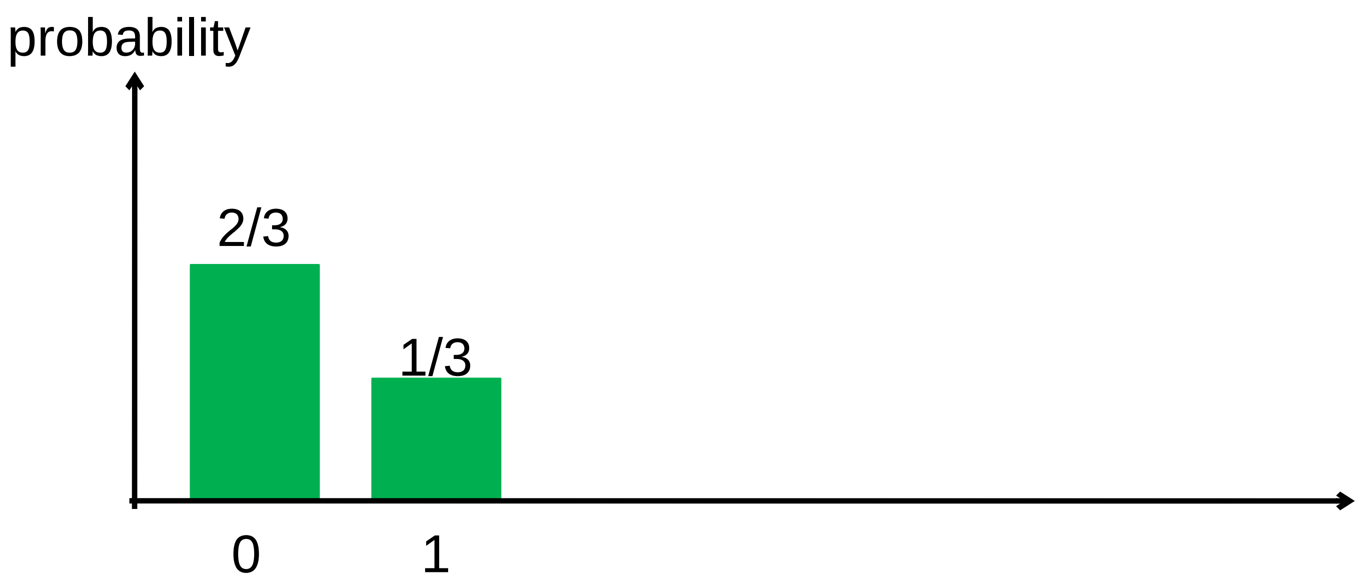
Important if you will gamble…

Suppose you throw a die once:

"1", "2", "3", "4" -> You lose = 0

"5", "6" -> You win = 1

Bernoulli experiment



Important if you will gamble…

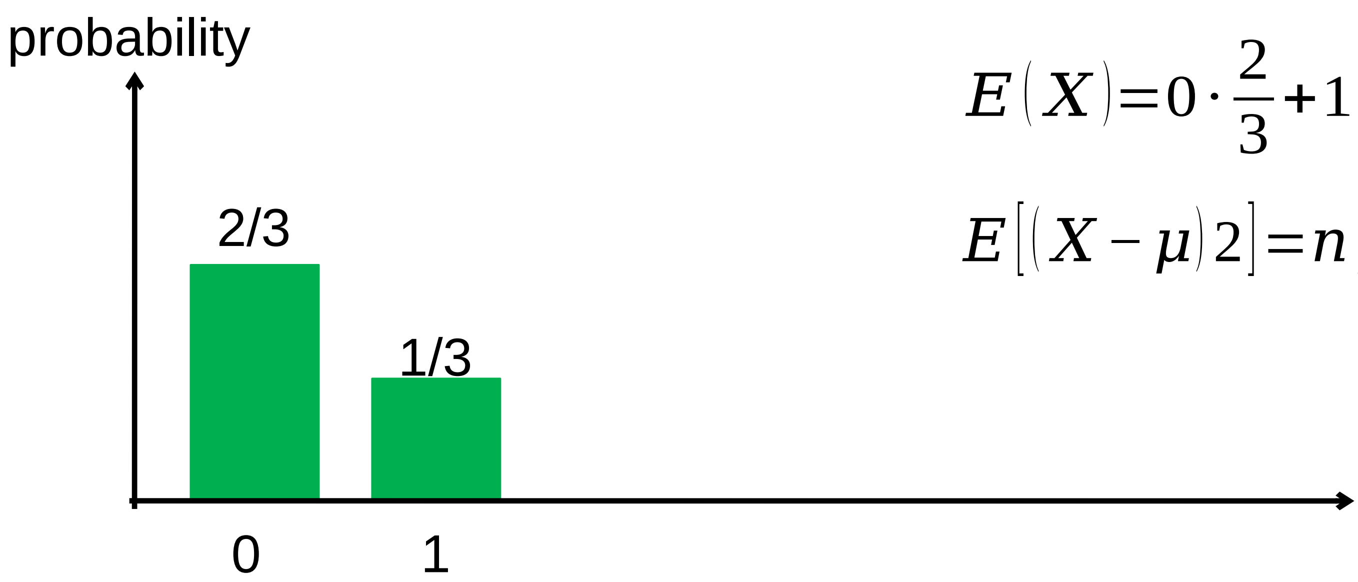
Suppose you throw a die once:

"1", "2", "3", "4" -> You lose = 0

"5", "6" -> You win = 1

Bernoulli experiment

$$P(X=x)=\binom{n}{x}p^{x}q^{(n-x)}$$



$$E(X)=0\cdot\frac{2}{3}+1\cdot\frac{1}{3}=np=\frac{1}{3}$$

$$E\left[(X-\mu)2\right]=npq=1\cdot pq=\frac{2}{9}$$

$$\sigma\approx 0.47$$

Binomial Distribution

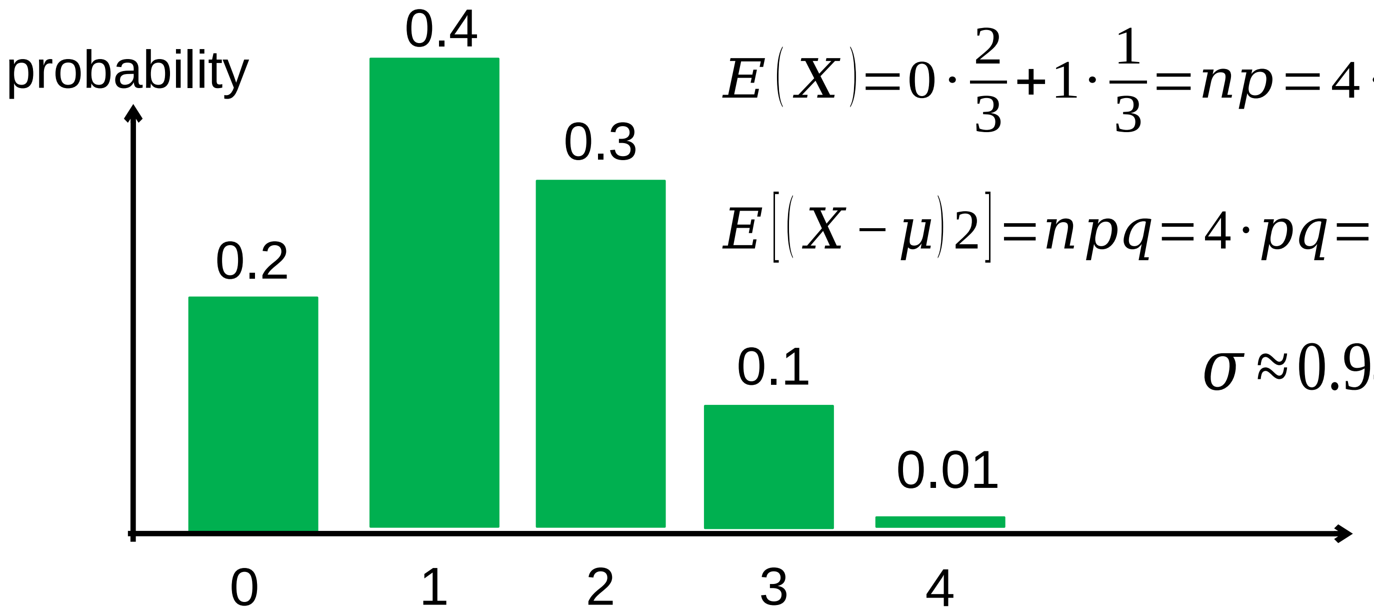
Suppose you throw a die <u>four times</u>:

"1", "2", "3", "4" -> You lose = 0

"5", "6" -> You win = 1

Binomial distribution:

$$P(X=x)=\binom{n}{x}p^{x}q^{(n-x)}$$



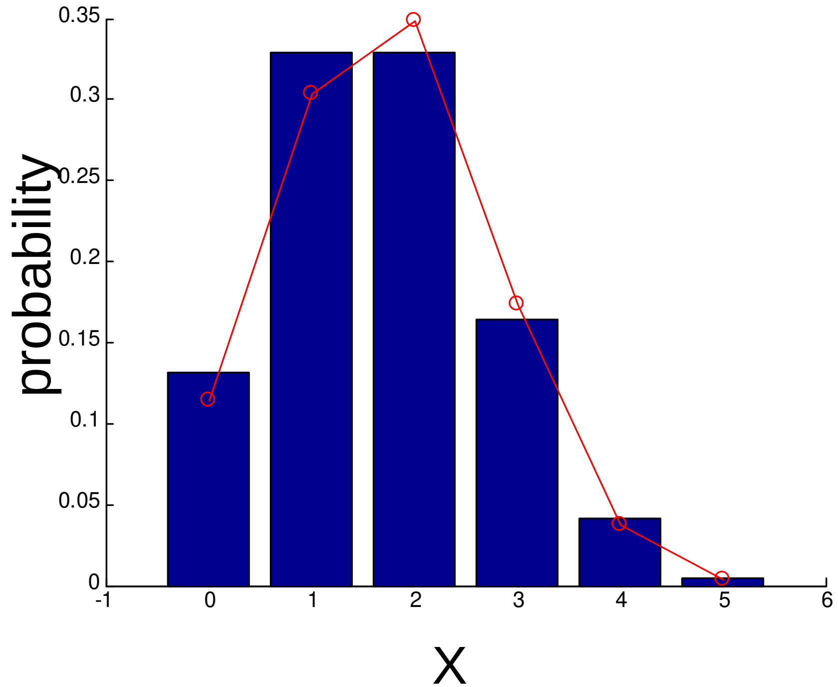
$$E(X)=0\cdot\frac{2}{3}+1\cdot\frac{1}{3}=np=4\cdot p\approx 1.33$$

$$E\left[(X-\mu)2\right]=npq=4\cdot pq=\frac{8}{9}$$

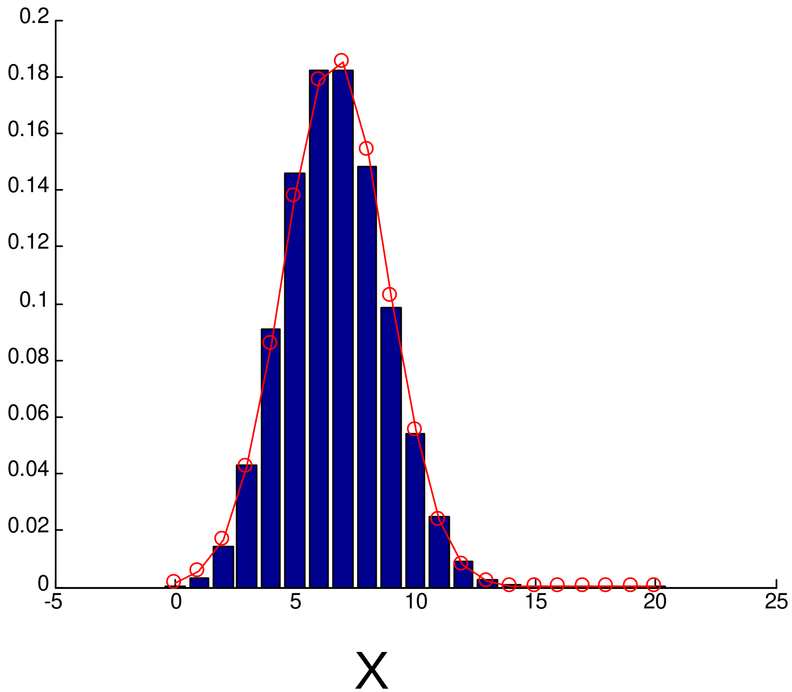
$$\sigma\approx 0.94$$

Binomial ↔ Normal

Suppose you throw a die 5 times



20 times



The red line and circles show the approximation with a normal distribution **N(μ,σ²) = N(np,npq).** With increasing number of n, the binomial distribution converges toward the normal distribution.

(formal proof: de Moivre-Laplace theorem)

Binomial ↔ Normal

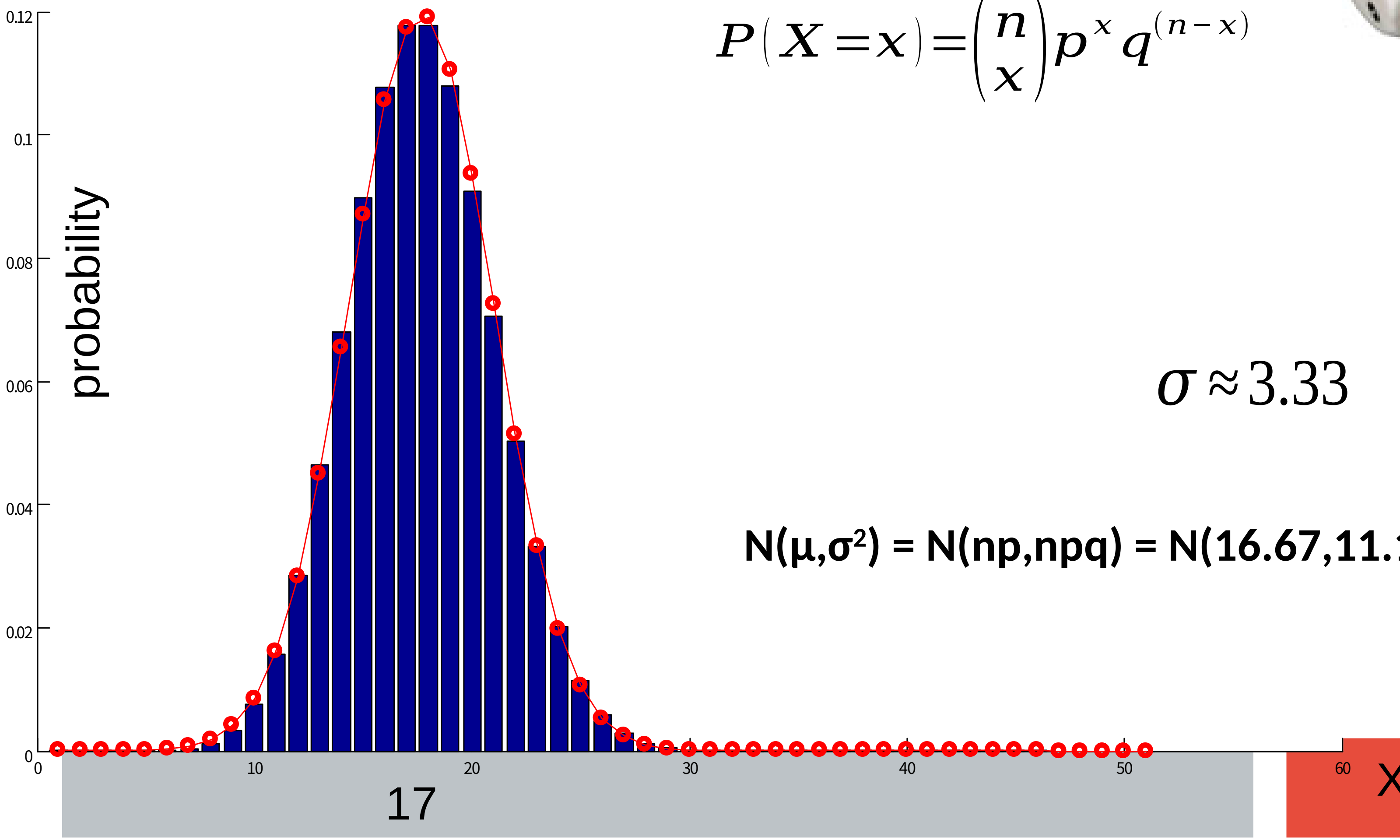
Suppose you throw a die 50 times:

"1", "2", "3", "4" -> Failure = 0

"5", "6" -> Success = 1

Binomial distribution:

$$P(X=x)=\binom{n}{x}p^x q^{(n-x)}$$



$$\sigma \approx 3.33$$

N(μ,σ²) = N(np,npq) = N(16.67,11.11)

Central Limit Theorem (CLT)

Suppose is the mean of a random sample X_1, X_2, …, X_n (sample size n). The random variables X_i have a distribution with a mean μ and variance σ^2.

for n-> ∞:

$$\bar{X} \sim N\left(\mu, \sigma^2/n\right)$$

i.e., for large n, Xbar is normally distributed with parameters μ, σ^2/n

or:

$$\bar{X} \sim N\left(0,1\right)$$

i.e., for large n, a standardized Xbar follows a standard normal distribution

$$N(\mu,\sigma^2): f\left(x\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Central Limit Theorem (CLT)

Suppose is the mean of a random sample X_1, X_2, …, X_n (sample size n). The random variables X_i have a distribution with a mean μ and

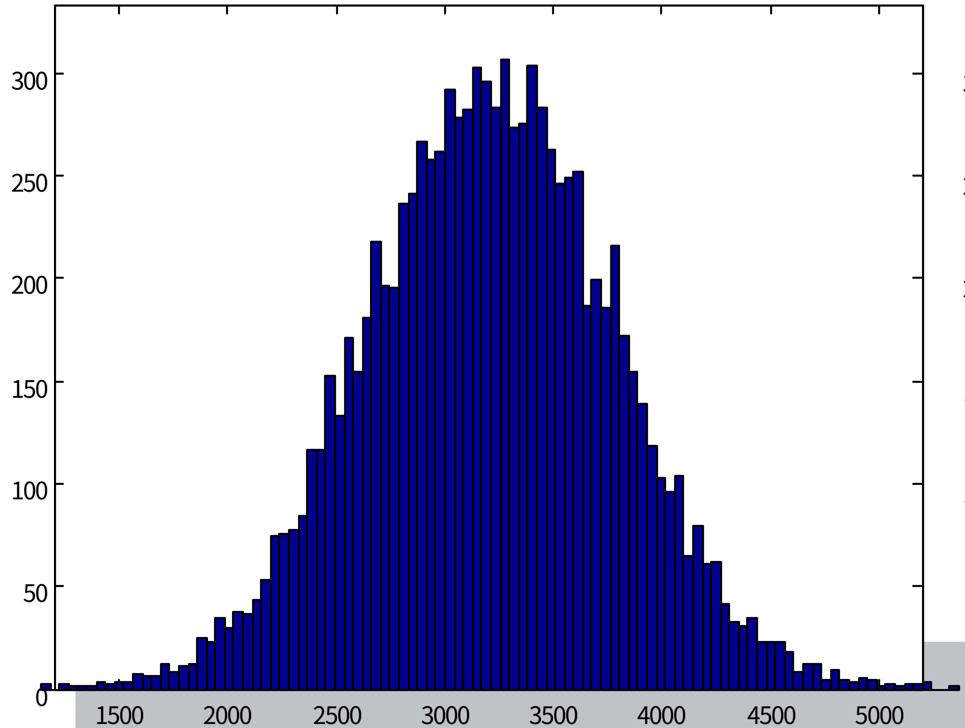
This works no matter the distribution of your Xs

This is amazing.

Example: babies

<u>Population distribution</u>

N(3209, 564^2) μ = 3209g
σ = 564g



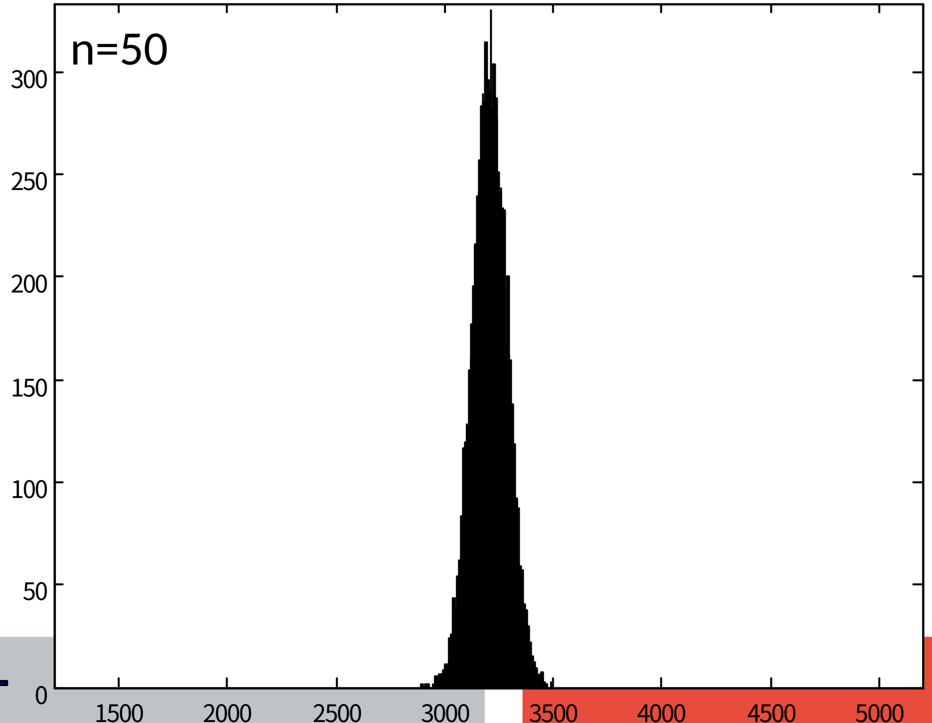
Birth weight [g]

<u>Sampling distribution</u>

is approximately normally distributed:

$$N\left(\mu,\sigma^2/n\right)=N\left(3209,562^2/50\right)$$

Usually, we do not know σ, so we approximate by N(Xbar, s^2/2)



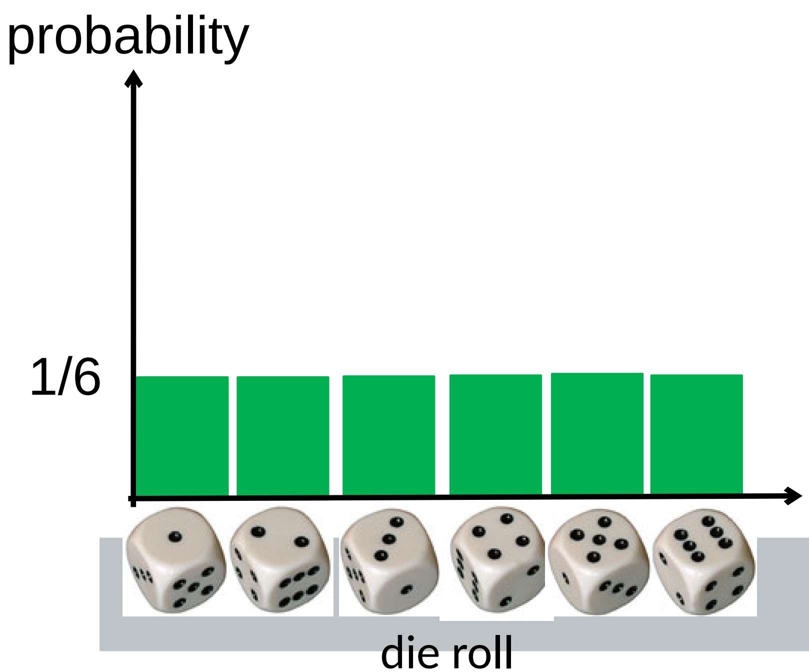
sample mean [g]

Example: dice

Population distribution

uniform, P(X=x)=1/6

μ = 3.5
σ2 = 35/12

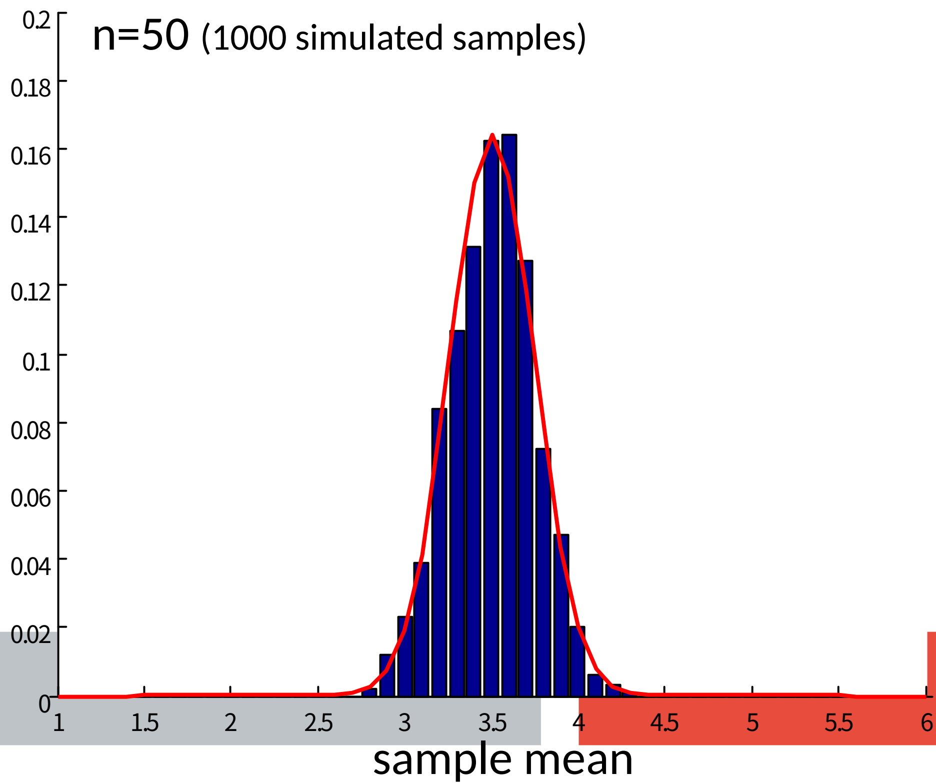


Sampling distribution

is approximately normally distributed:

N(,σ) = N(3.5,(35/12)/50)

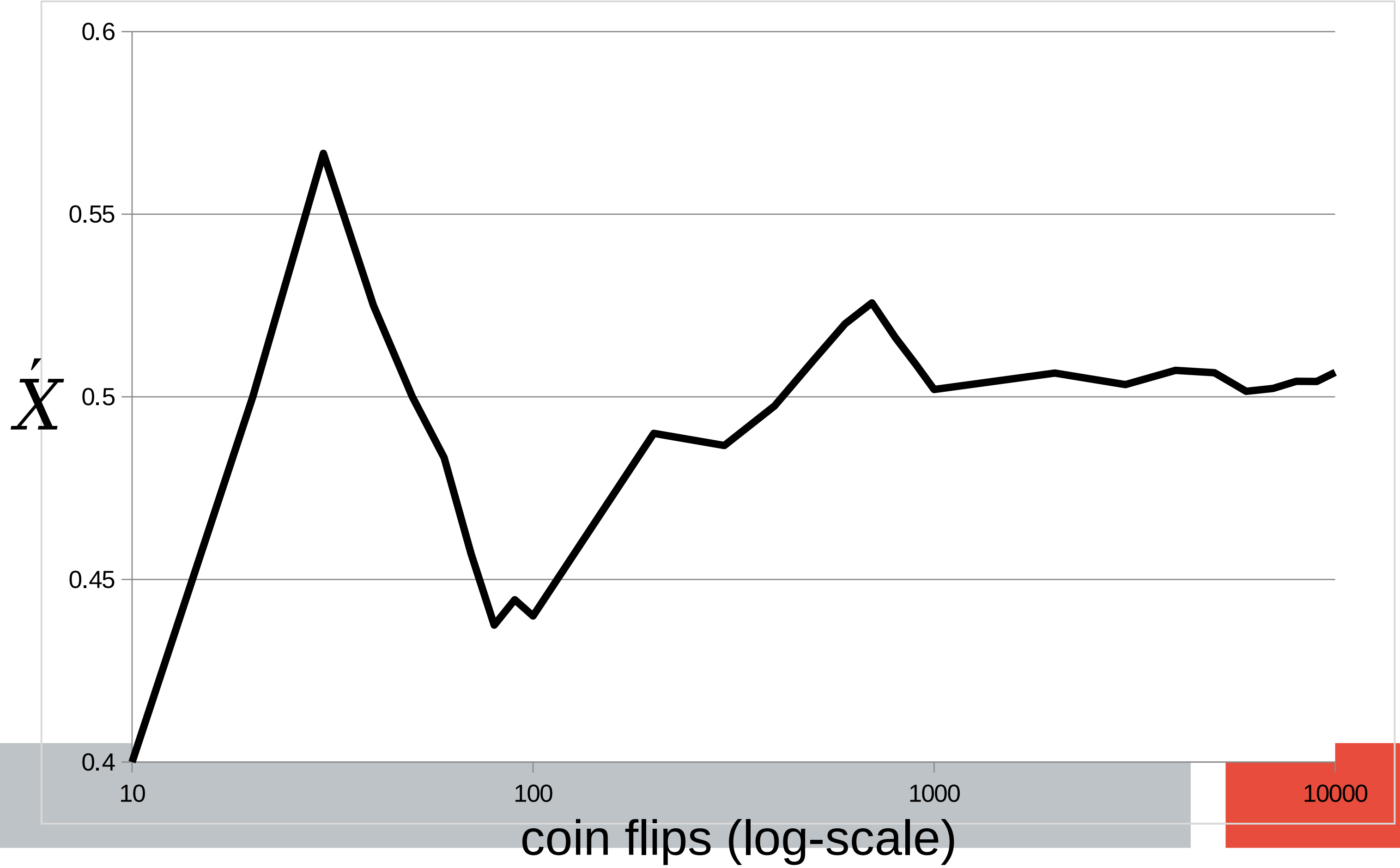
$$N(\mu, \sigma^2/n) = N(3.5,\ (35/12)/50)$$



Law of Large Numbers

Another important limit theorem is the law of large numbers: if we repeat an experiment very often, the observed mean should converge to the theoretical ("true") mean.

John Kerrich, a South African mathematician, was interned 1940 in Denmark. He flipped a coin 10,000 times…



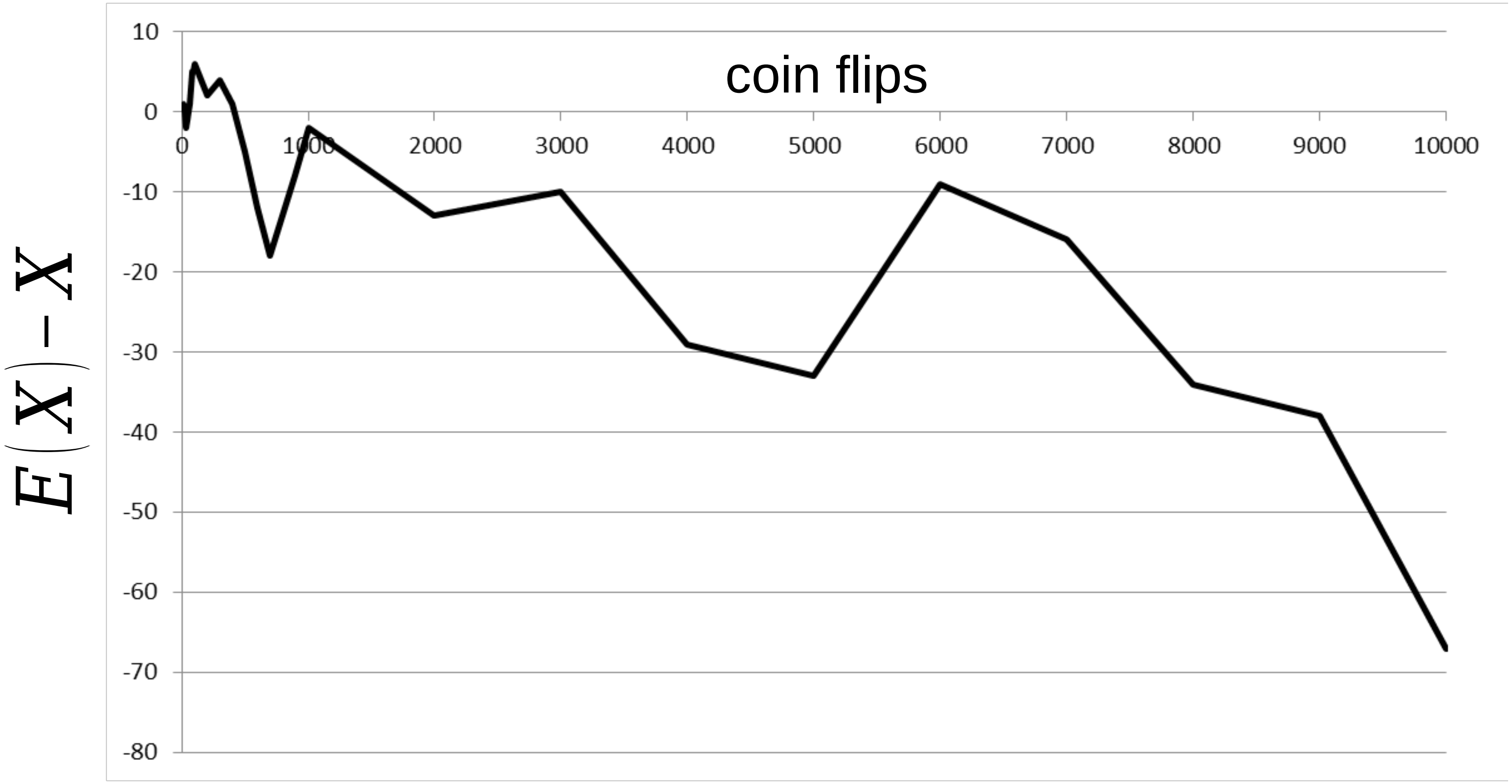
Law of Large Numbers

Question #1: Suppose, I'd give you 1000¥ if you get more than 60% heads. You can choose if you flip 10 or 1000 times.
Which is better?

Question #2: Suppose, I'd give you 1000¥ if you get 5 more heads than expected. You can choose if you flip 10 or 1000 times.
Which is better?

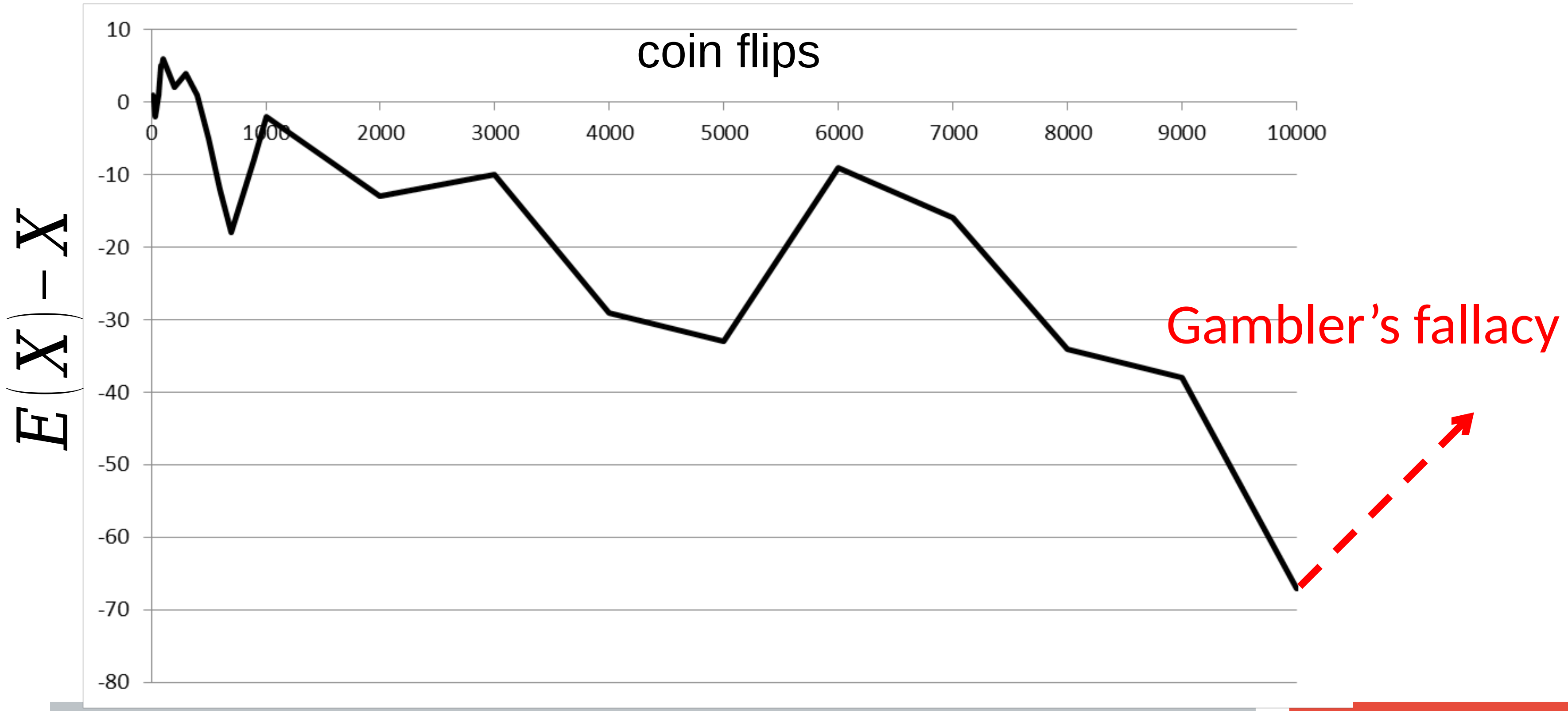
Law of Large Numbers

This shows John Kerrich's data expressed as absolute difference to the expected value (for 1000 coin flips: 500 heads).



Gambler’s Fallacy…

The belief that if someone has a “streak of luck”, probabilities for a success change, is called gambler’s fallacy. So, if somebody had a series of tails, it should become more likely to get heads next:

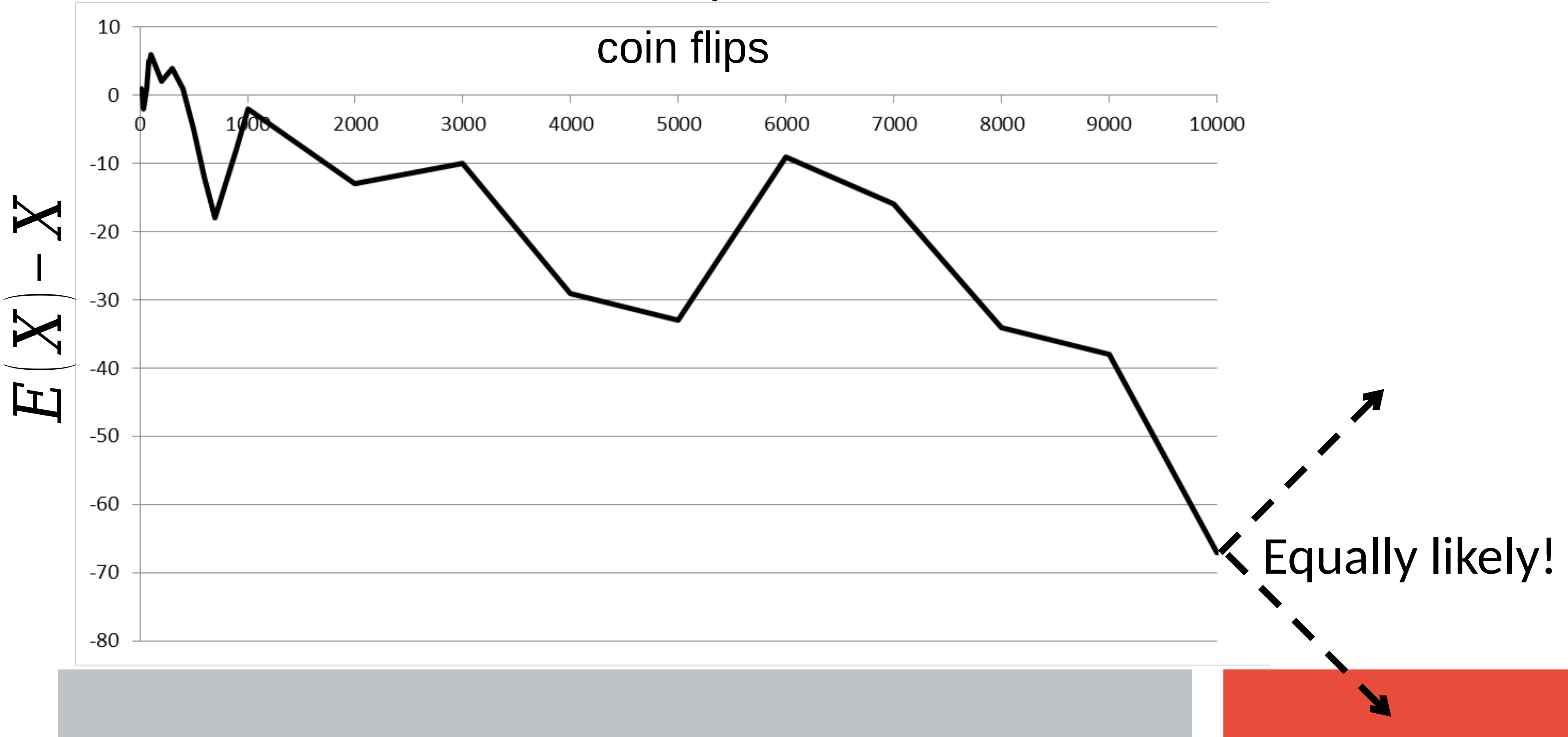


Law of Large Numbers

But: $P(X_{n+1}=0+X_n | X_n<E(X)) = P(X_{n+1}=1+X_n | X_n<E(X))=$

$P(X_{n+1}=0+X_n | X_n>E(X)) = P(X_{n+1}=1+X_n | X_n>E(X))$

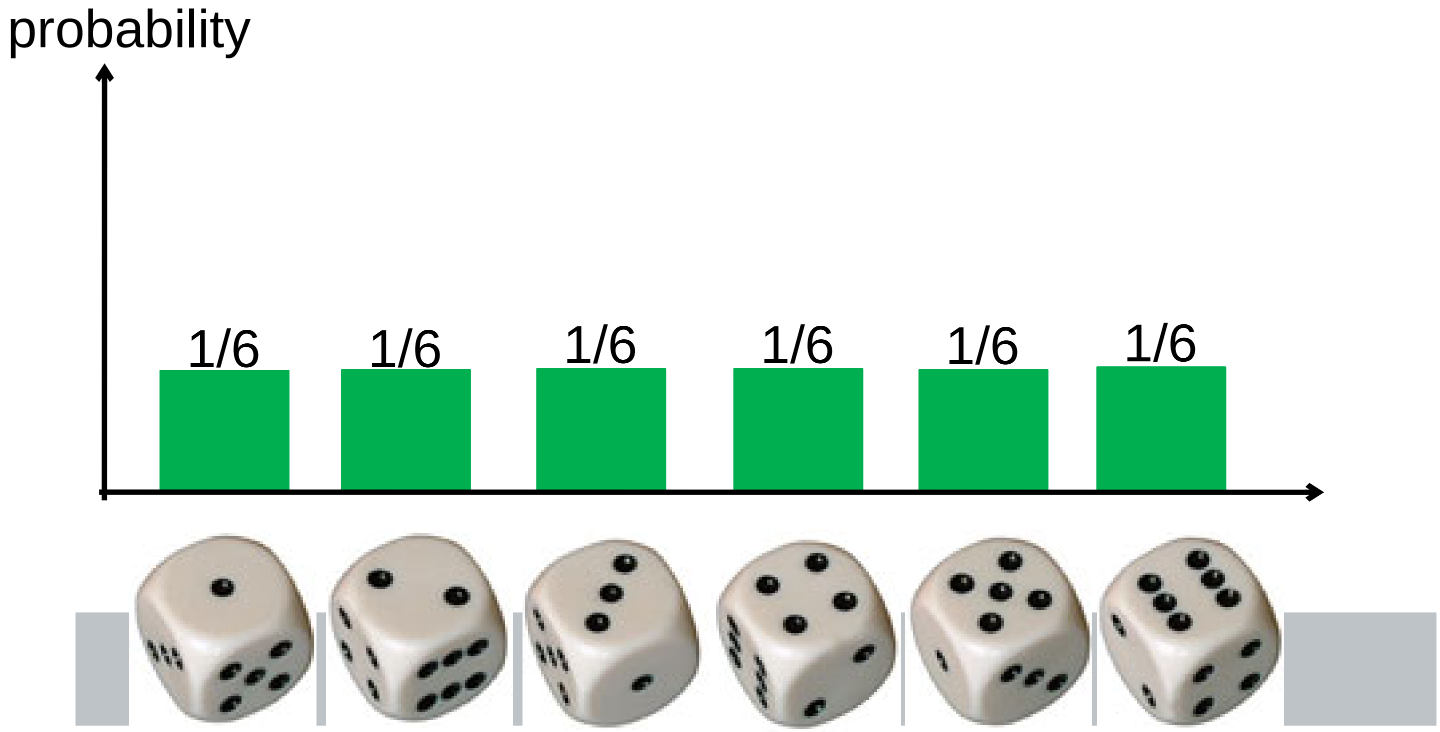
The dice/coins have no memory!



Law of Large Numbers: Dice

$x \in \{1,2,3,4,5,6\}; P(X=x) = \frac{1}{6}$ Expected result: $E(X) = \mu = 3.5$

The law of large number means that mean approaches 3.5 if we throw the dice long enough.



Law of Large Numbers

$$\lim_{n \to \infty} P(|\bar{X} - \mu| \geq \epsilon) = 0$$

$$\lim_{n \to \infty} P(|\bar{X} - \mu| < \epsilon) = 1$$

With as the observed mean of a random sample of the size from a distribution with mean .