Lecture - Day 3 (part 2)
Normal distribution

MF9130E V24
2024.04.10

Chi Zhang
Oslo Center for Biostatistics and Epidemiology
chi.zhang@medisin.uio.no

Random variables

Discrete and continuous

Random variable: a quantity that can take random values, with a certain probability

Discrete variables
- coin tossing H,T
- birth boy, girl

Continuous variables
- weight and height
- age

Properties of probability:

Non-negative (0 or above), less than 1, prob of all outcomes sum up to 1.

Histogram and bar plot
X-axis is usually what **values** the variable can be - either a single value, or a range

Y-axis is the **frequency**; or **proportions** (probability) corresponding to that variable value (or range)

Histogram

Continuous variable

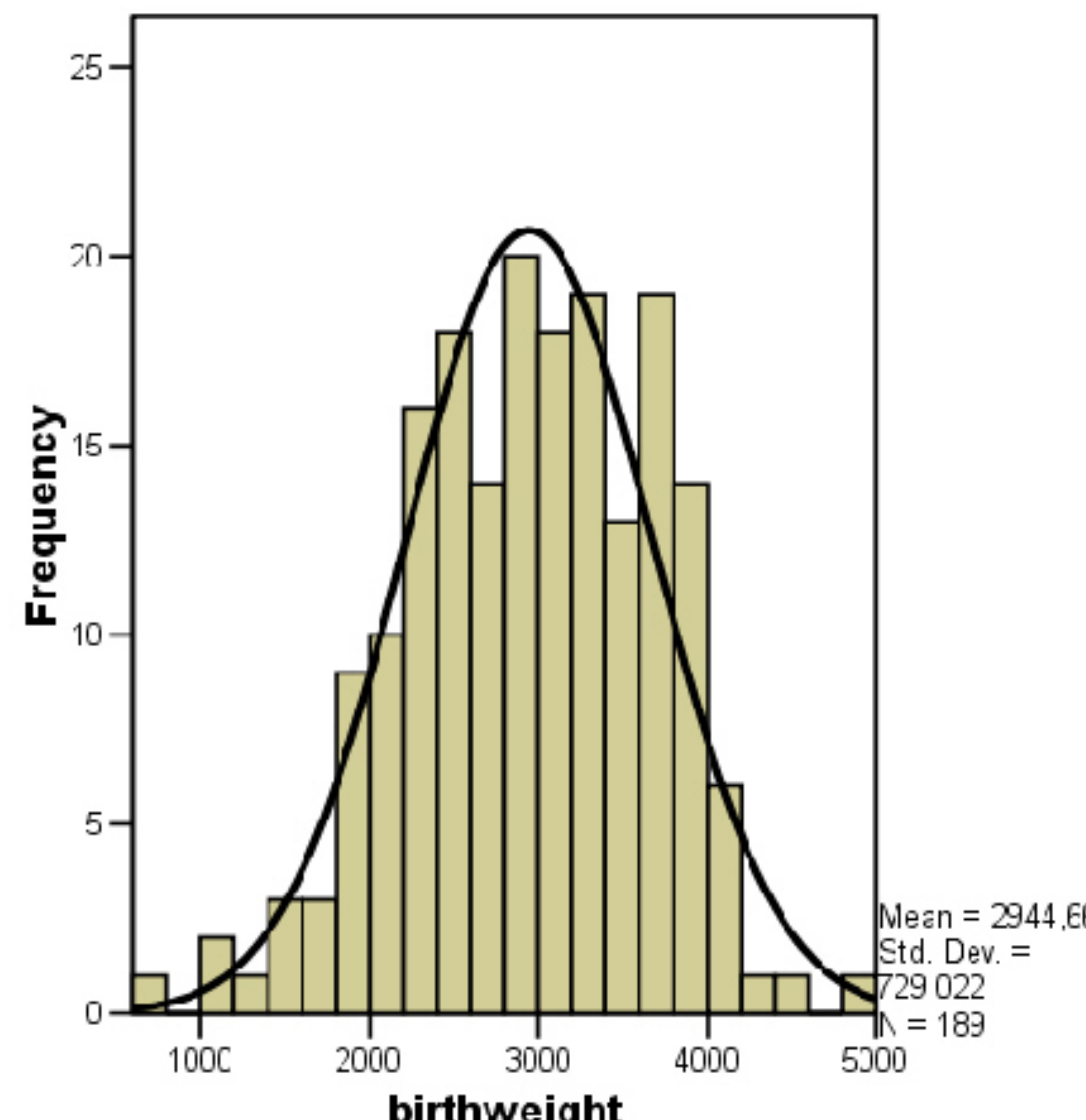
Probability distribution: relationship between outcome and probability

Outcome here is the **measurement (on a continuous scale)**, instead of categories

E.g. Birth weight of 189 newborns
3001, 2918, 3000, 3001,….

How to find the probability? Try **counting elements in an interval**
- 20 measurement between 2800 and 3000;
- 30 between 3000 and 3200; for example.
- the interval (bin) can be bigger, or smaller

You can get the proportion for each interval by dividing the counts over N



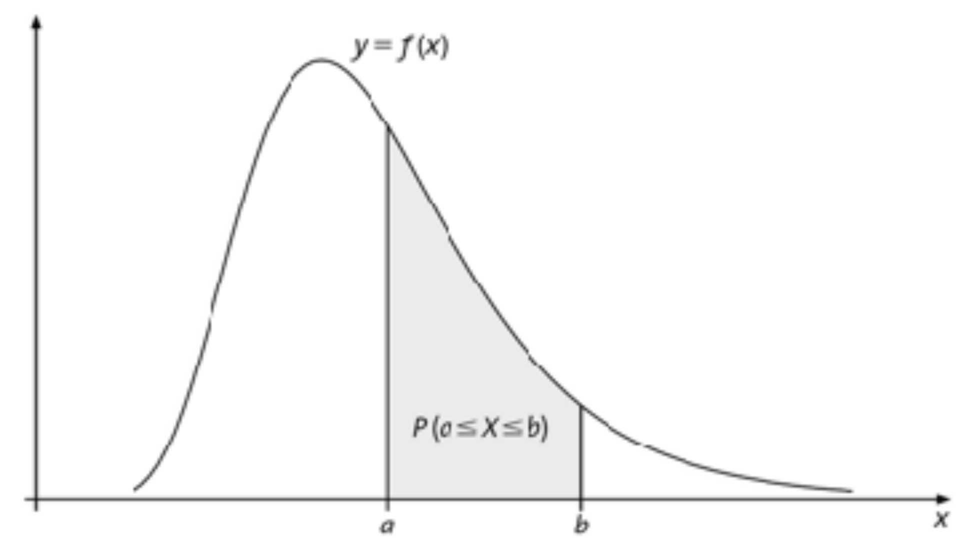
Probability distribution

Probability density

A continuous probability distribution is defined by the **probability densitify function** f(x) (a relationship between the measurement x and its **density** at this point)

NOT a probability itself! P(x) = 0. Need integration

The following properties are important
f(x) >= 0
The area under the curve in total is 1
P(a <= X <= b) is the **area under the curve** from a to b



Figur 5.1 Sannsynlighetstetthet for en målevariabel

We are usually more interested in the probability of X **greater or smaller** than a certain value; or **between** two values.
e.g. we care about probability of birth weight between 3000 to 3010, but not equal to 3000 exactly.

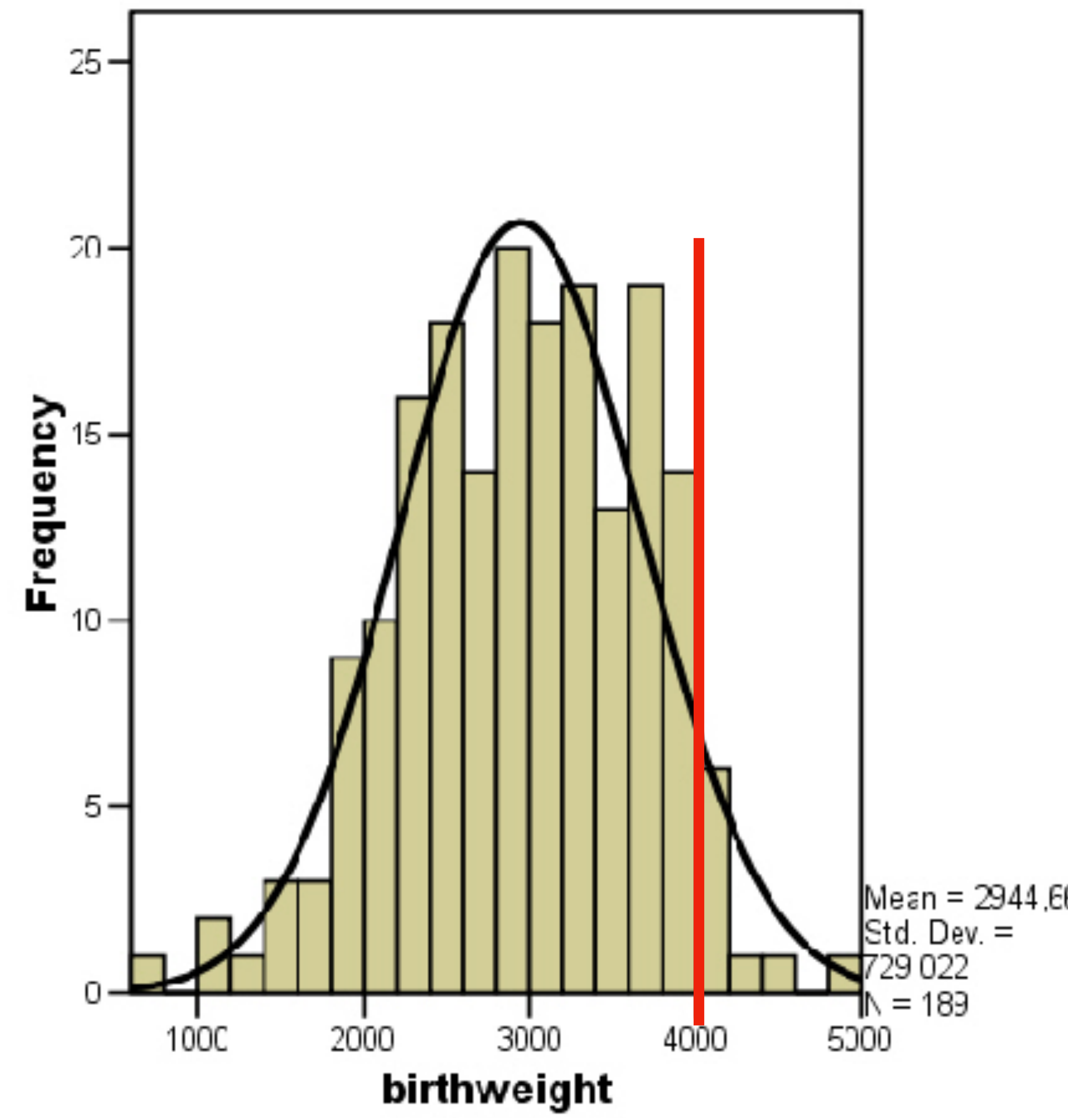
Probability distribution

Probability density

Birthweight example: N=189

The area under the curve in total is 1
- if you sum all the bars (2 + 3 + 1 + …) the total is 189
- (the **height** of each bar)

P(a <= X <= b) is the area under the curve from a to b
- P(2800 <= X <= 3000) = 20/189
- P(X > 4000) =9/189

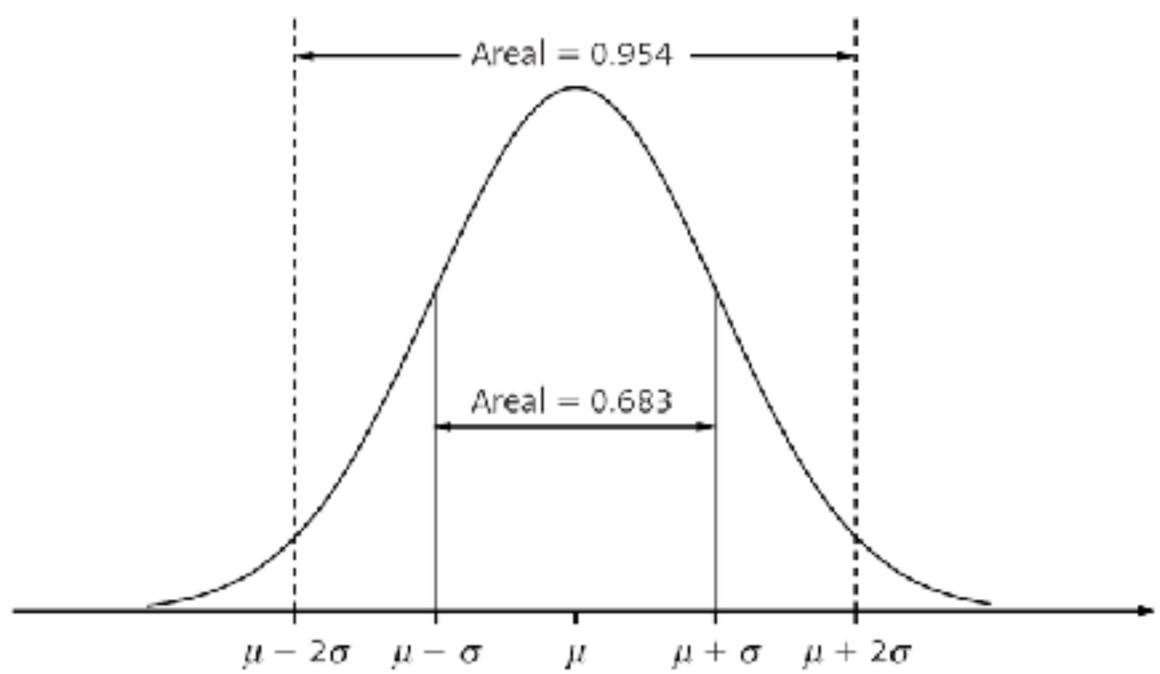


Normal distribution

Probability densifity function of the normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

where μ is the mean, σ the standard deviation and $\exp(a) = e^a$



Figur 5.4 Tegning av en normalfordeling. Det er avmerket at $\mu \pm \sigma$ dekker 68 %, mens $\mu \pm 2\sigma$ dekker 95 %.

$$X \sim N(\mu, \sigma^2)$$

Read: mu, sigma

Normal distribution

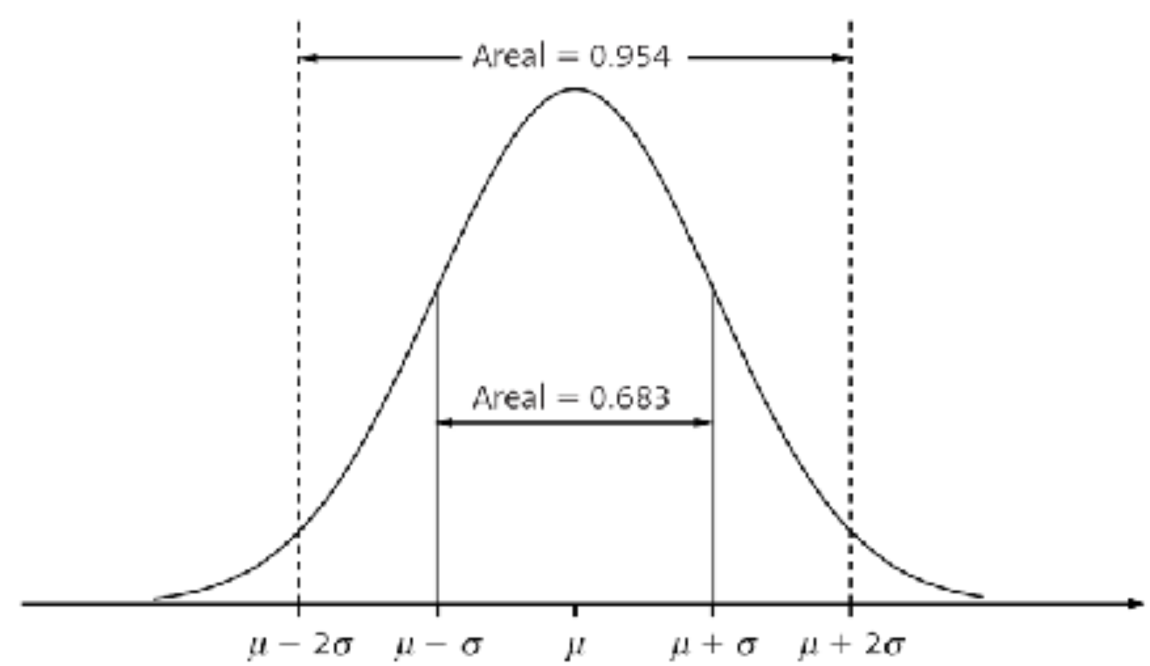
$$X \sim N(\mu, \sigma^2)$$

Symmetric, bell shape

Mean defines the location (where it is centered)

SD (sigma) defines the variation, i.e. spread

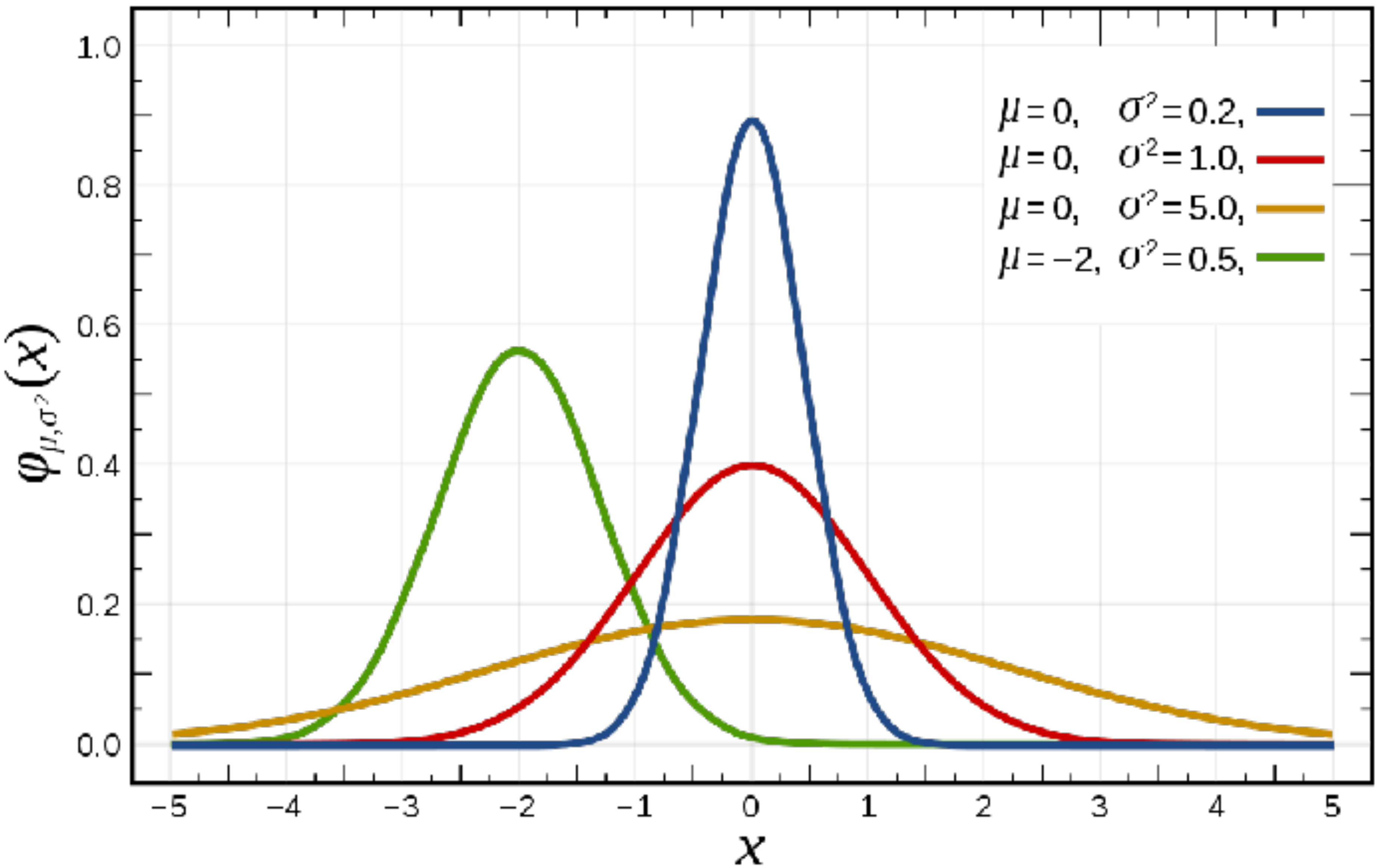
An interval with center at the mean value, going 2 standard deviation each way covers **approximately 95%** of the distribution



Figur 5.4 Tegning av en normalfordeling. Det er avmerket at $\mu \pm \sigma$ dekker 68 %, mens $\mu \pm 2\sigma$ dekker 95 %.

Normal distribution

Different locations and sd



Standard normal distribution

Definition

A **standard** normal distribution is a normal distribution with mean 0, variance 1

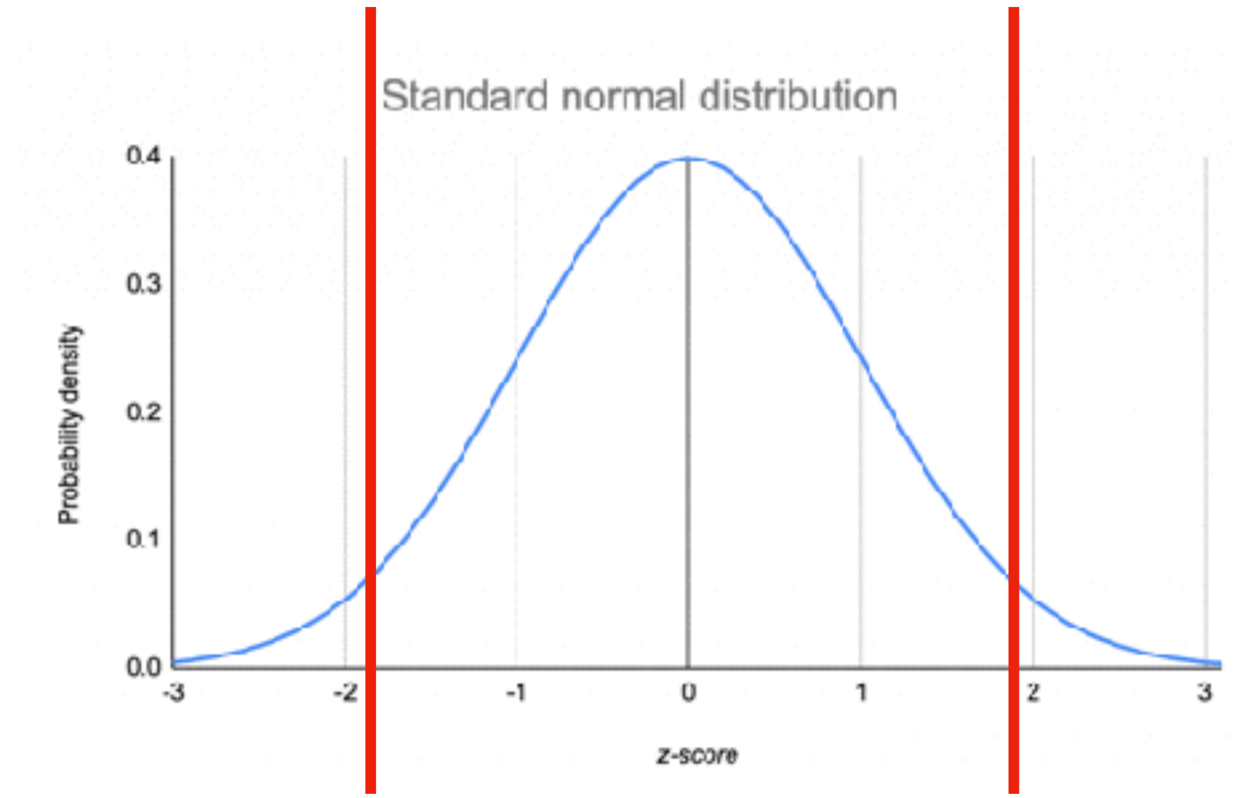
Noted as N(0, 1)

Any normal distribution can be transformed into a standard normal …

By substracting the mean, and dividing by the standard deviation

$$X \sim N(\mu, \sigma^2)$$

$$Y = \frac{X - \mu}{\sigma} \sim N(0, 1)$$



-1.96, 1.96 are 2.5% and 97.5% quantile for N(0,1)
Frequently used for computing 95% confidence intervals

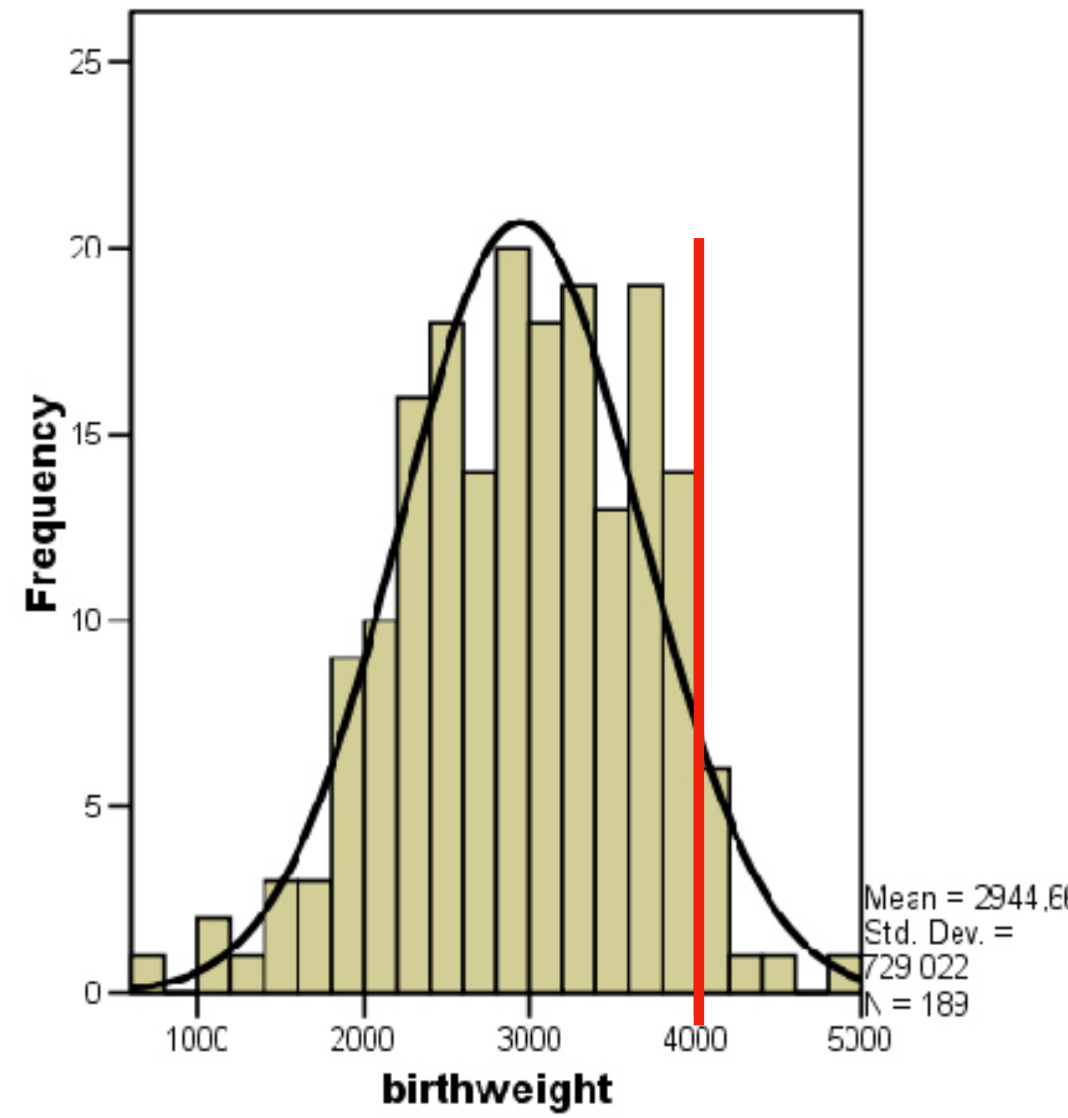
Standard normal distribution

We calculated the 'empirical' probability of weights greater than 4000 by **counting**:
P(X > 4000) =9/189 = 4.7%, approximately

Try using the theoretical distribution (**curve**)

- mean (mu) = 2945, sd (sigma) = 729
- transform X into Y, which is **N(0, 1)**
- then find the area for P(Y>y)

P(X>4000)
= P[(X - 2945)/729 > (4000-2945)/729]
= P[(X - 2945)/729 > 1.45]
= P(Y>1.45)

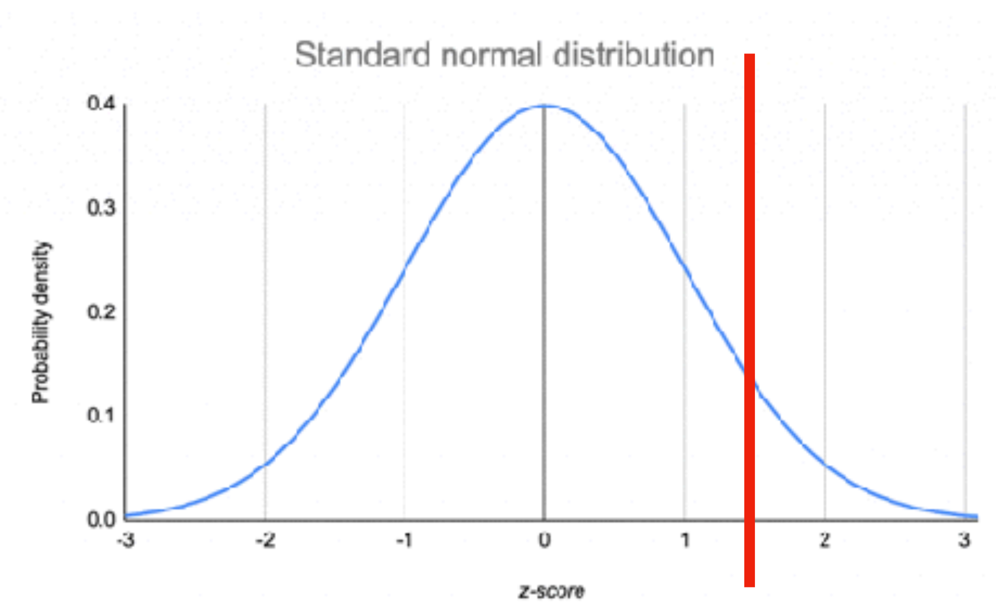


Standard normal distribution

Find P(Y>1.45)

Method I: check a table
P(Y>1.45) = **1 - P(Y <=1.45)** = 0.073

Need to be careful what the table shows you, remember complement rule!



Tabell over normalfordelingen

(standardisert til forventning 0 og standardavvik 1)

Tabellen gir sannsynligheten $\boxed{P(Y \leq y)}$ der Y er standard normalfordelt

Eksempel: $P(Y \leq 0.23) = 0.5910$

For negative tall kan du bruke: $P(Y \leq -y) = 1 - P(Y \leq y)$

y	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998

Standard normal distribution

Find P(Y>1.45)

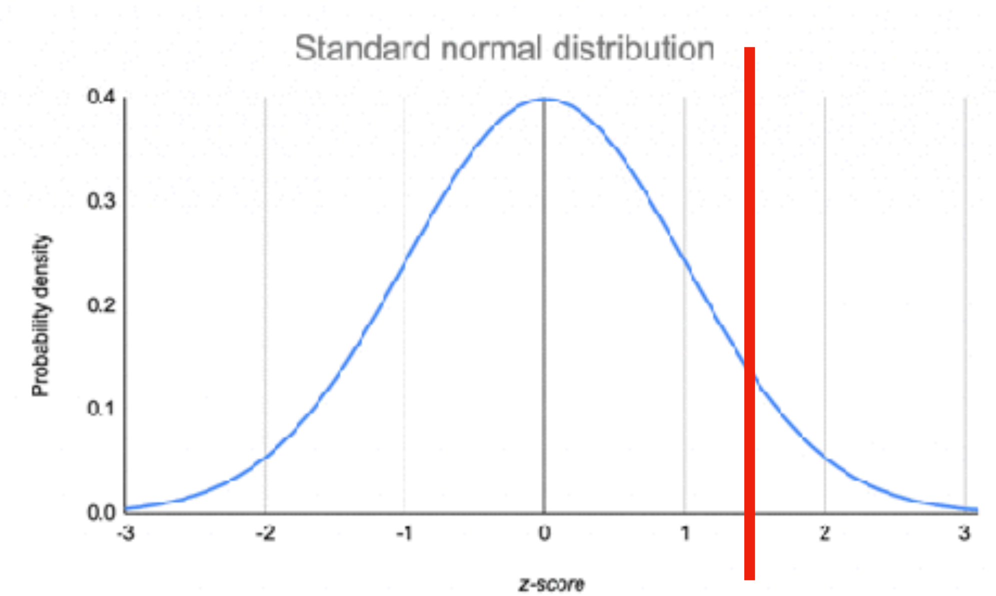
Method II: in R

```
> pnorm(1.45, 0, 1)
[1] 0.9264707
> 1-pnorm(1.45, 0, 1)
[1] 0.07352926
```

Method III: in STATA

```
. display normal(1.45)
.92647074

. display 1-normal(1.45)
.07352926
```



Central limit theorem CLT

Important conclusion about the mean

Sample mean is normally distributed around the true, **population mean,** with 1 over n times the variance

Regardless of which distribution the population is.
- Binomial
- Poisson
- Uniform
- …

This also applies for **sum** (n times mean; because mean is sum divided by n).

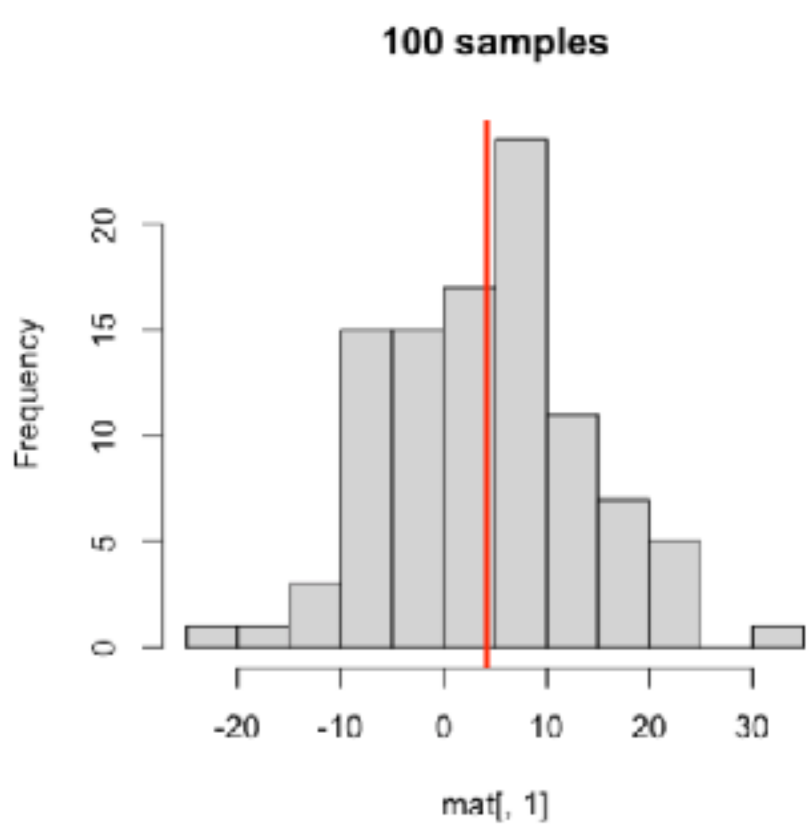
You will be able to make **inference** with CLT (relevant for confidence interval, t-test and more)

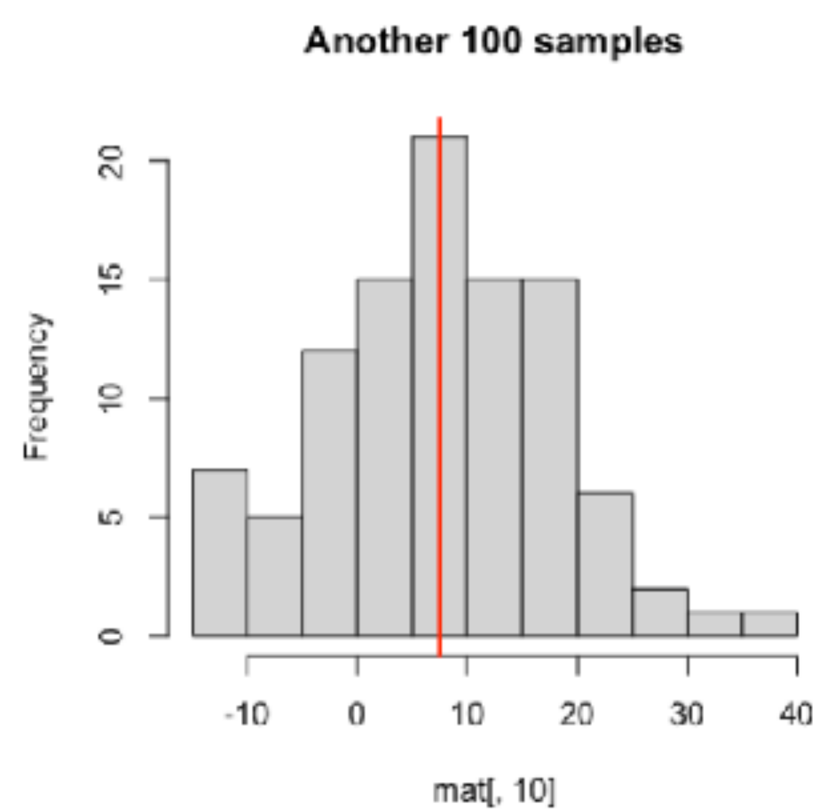
Central limit theorem CLT

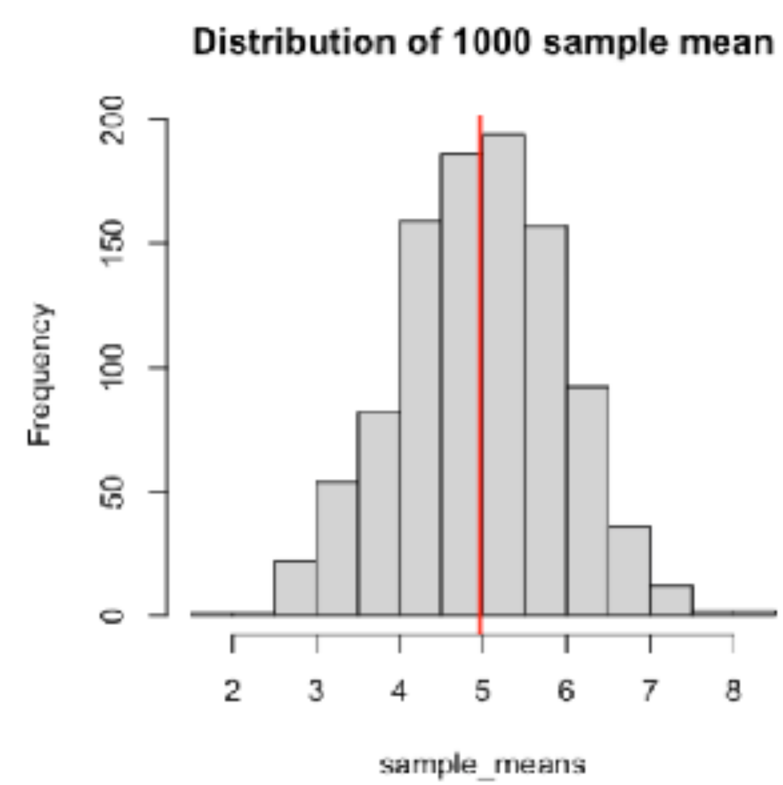
Draw 100 random samples with a known theoretical distribution: N(5,100) - centered at 5, with variance 100 (standard deviation 10)
Mark their mean with red in figure 1, 2

Repeat the procedure 1000 times, compute 1000 means, plot the **sample means**

Centered at 5, with variance 100/100 = 1 - a Normal Distribution!
First 100 is the sample size (100 random samples); second 100 is the variance







Central limit theorem CLT

Your data does not need to be normally distributed; it can be any shape

Their sample mean is normally distributed

This result is the foundation for day 3 and day 4, you will see that you can compute the range where your mean can vary!