Advanced Research Methods

NOT ONLY STATS

Paweł W. Woźniak

PLAN

- Today: More about experimental designs + some stats
- Next week: experimenting with stats
- In two weeks: Presentation and visualisation

How to use the literature

- Hornbæk is a crash course
- MacKenzie has more then you need to know (Ch 5,6,8)
- Use Field as R reference

Overview

- Discrete and Continuous Data
  - Modeling and Measuring
  - Binomial and Gaussian distributions
- Descriptive statistics
  - distribution of continuous data
  - computing the average (mean, median, mode)
  - computing the variance (sum of squares, variance, standard deviation)
- The normal distribution
  - z-scores

Tossing a coin twice

If I flip a coin two times,
how many “heads” will there be?

1. 0 head
2. 1 head
3. 2 heads

Tossing a coin twice

If I flip a coin two times,
how many “heads” will there be?

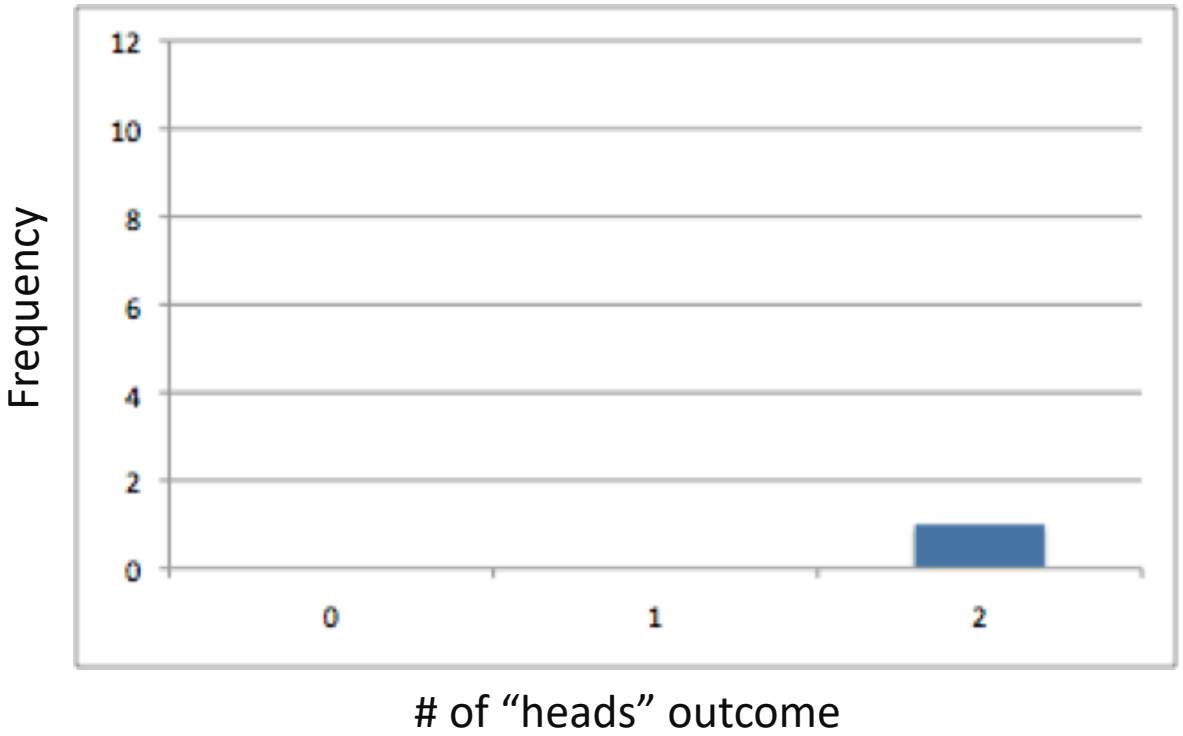
Answer is probabilistic, not deterministic.

Definition of probability:

the extent to which an event is likely to occur, measured by the ratio of the favorable cases to the whole number of cases possible.

Tossing a coin twice (20 experiments)

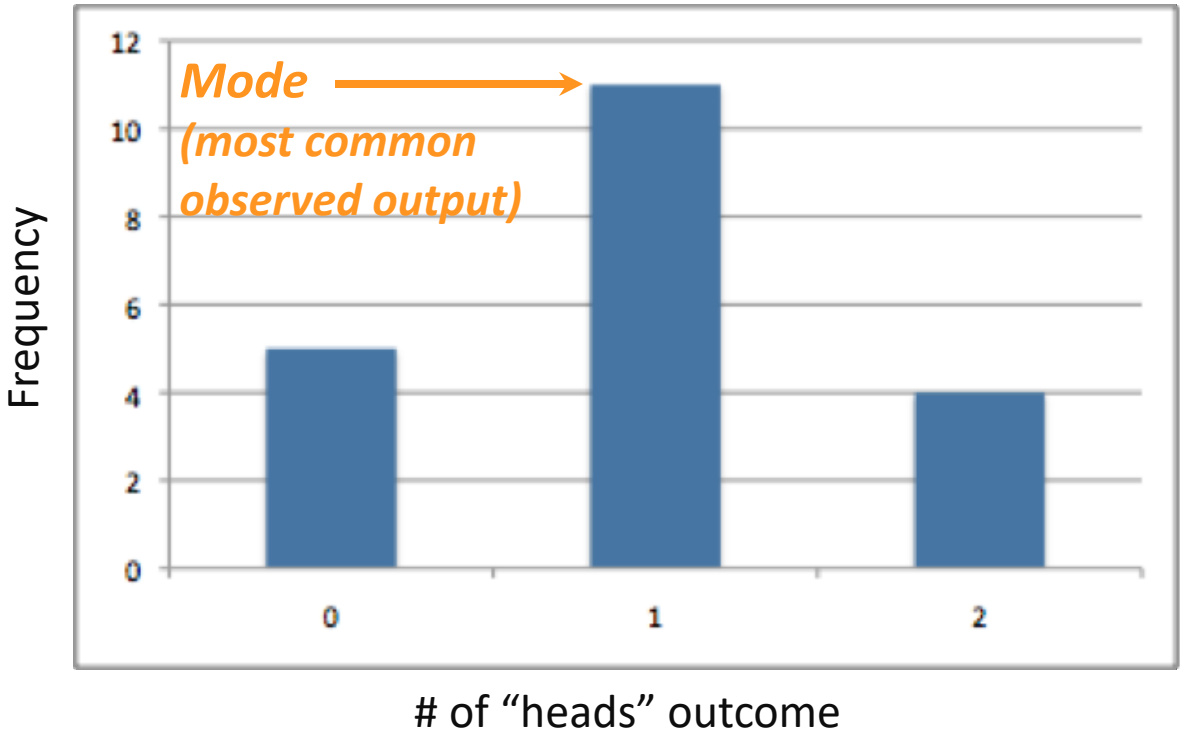
If I flip a coin two times,
how many “heads” will there be?



Observation	No. of Heads
1	2
2	0
3	1
4	0
5	0
6	1
7	0
8	1
9	1
10	2
11	1
12	1
13	1
14	1
15	0
16	2
17	1
18	1
19	2
20	1

Tossing a coin twice (20 experiments)

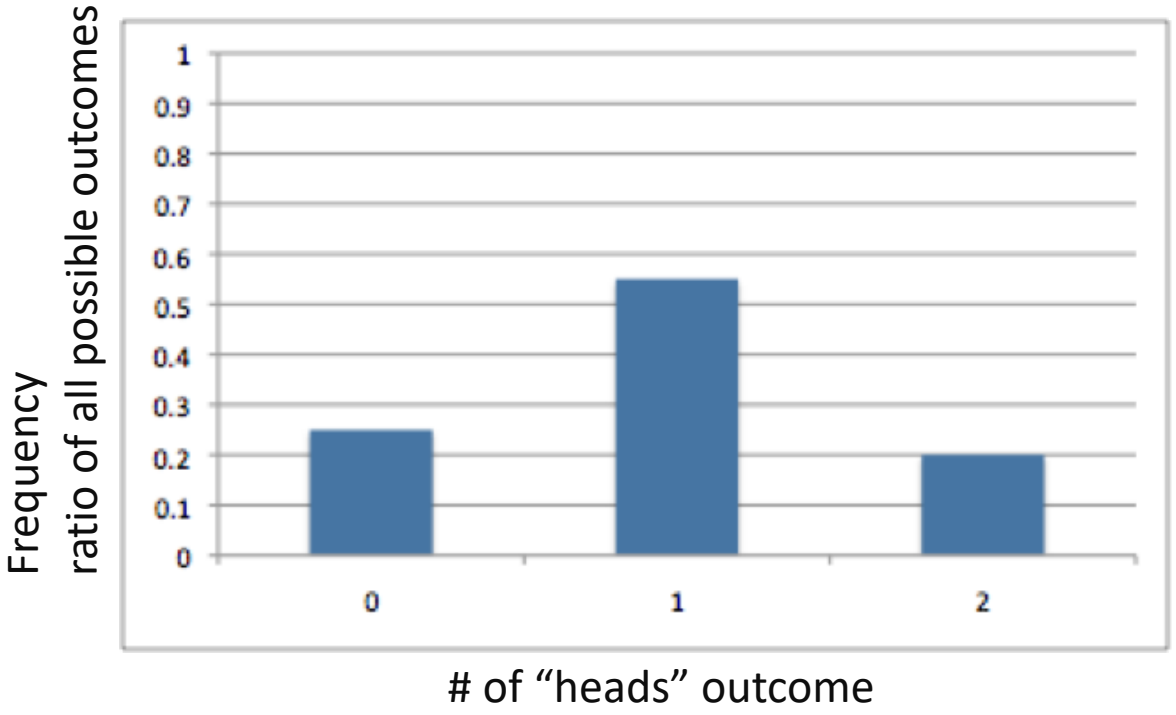
If I flip a coin two times,
how many "heads" will there be?



Observation	No. of Heads
1	2
2	0
3	1
4	0
5	0
6	1
7	0
8	1
9	1
10	2
11	1
12	1
13	1
14	1
15	0
16	2
17	1
18	1
19	2
20	1

Tossing a coin twice (20 experiments)

If I flip a coin two times,
how many "heads" will there be?



Definition of probability:
the extent to which an event is likely to occur, measured by the ratio of the favorable cases to the whole number of cases possible.

Only stupid people do experiments!!!

Developing a model for the "Tossing a coin twice" process

If I flip a coin two times,
how many "heads" will there be?

Outcome	First flip	Second flip
1	Heads	Heads
2	Heads	Tails
3	Tails	Heads
4	Tails	Tails

Developing a model for the "Tossing a coin twice" process

What is the ***probability*** of "heads"?

Outcome	First flip	Second flip
1	Heads	Heads
2	Heads	Tails
3	Tails	Heads
4	Tails	Tails

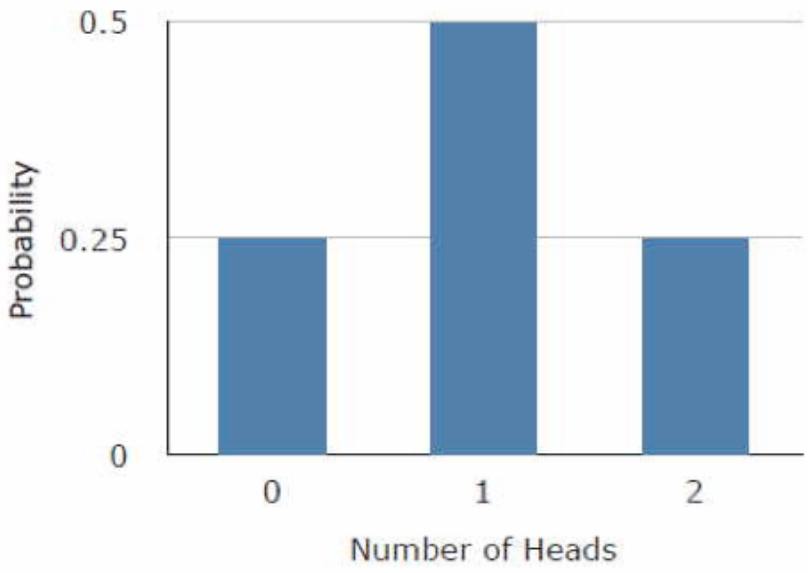
What is probability?

the extent to which an event is likely to occur, measured by the ratio of the favorable cases to the whole number of cases possible.

Developing a model for the "Tossing a coin twice" process

What is the probability of "heads"?

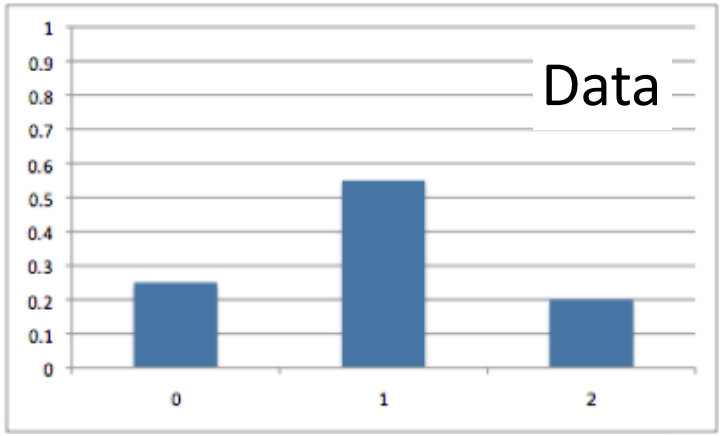
Outcome	First flip	Second flip
1	Heads	Heads
2	Heads	Tails
3	Tails	Heads
4	Tails	Tails

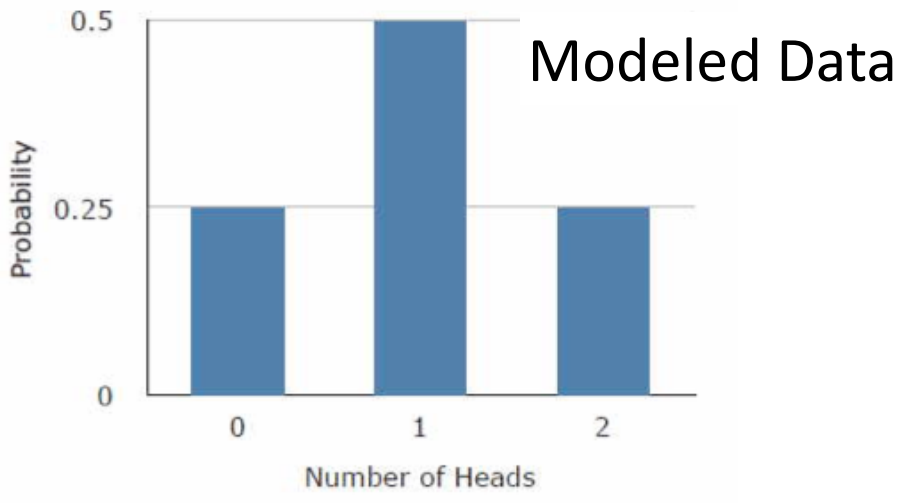


Tossing a coin twice

What is the probability of “heads”?

Outcome	First flip	Second flip
1	Heads	Heads
2	Heads	Tails
3	Tails	Heads
4	Tails	Tails





Which is the better approach?

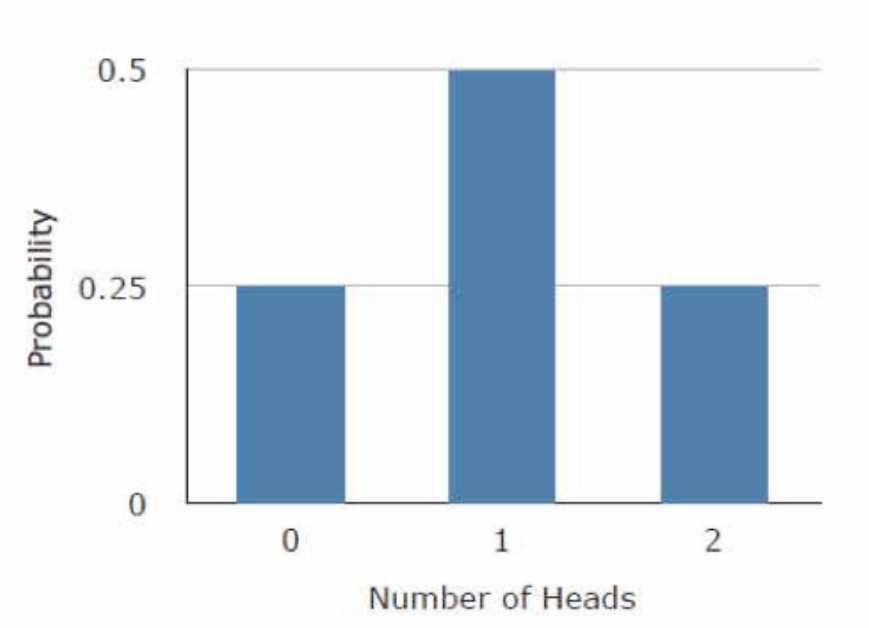
Neither!!! Models can be used to approximate true estimates from empirical data that will always be imprecise and noisy.

Data distributions

- Knowing how data is distributed is the first step towards understanding how one variable is likely to predict another variable.
- This distribution can serve as a model
- We can accept or reject this model, after collecting empirical evidence (i.e., experimental data).
- Models are important because they help us to organize our data meaningfully, even while we are collecting them.

Binomial Distribution

… is the discrete probability distribution of the number of successes (x) in a sequence of *N* independent yes/no experiments, each of which yields success with probability *p*

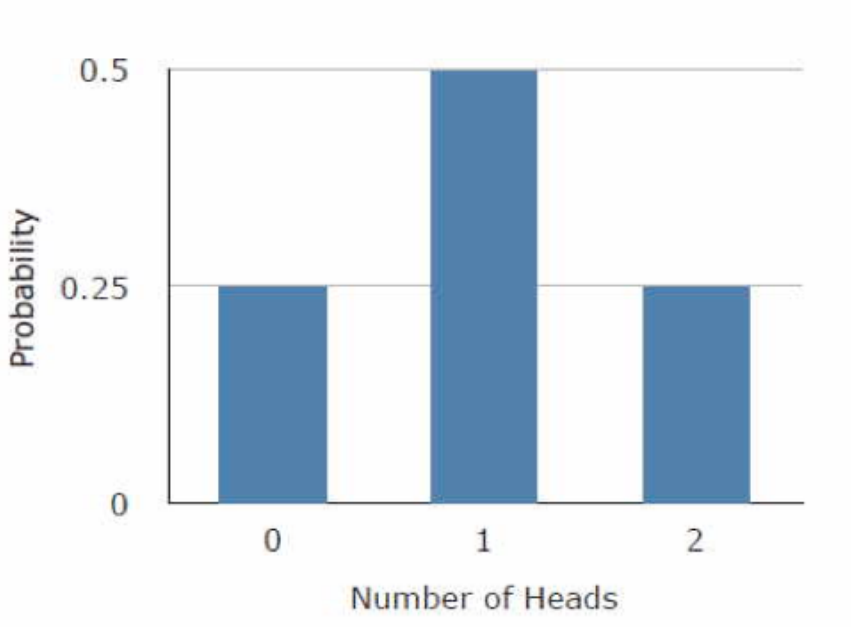


$$P(x) = \frac{N!}{x!(N-x)!}\pi^{x}(1-\pi)^{N-x}$$

N: number of flips
x: number of desired outcome
π: probability of the desired outcome (0.5=chance)

Binomial Distribution

… is the discrete probability distribution of the number of successes (x) in a sequence of *N* independent yes/no experiments, each of which yields success with probability *p*



$$P(x) = \frac{N!}{x!(N-x)!}\pi^{x}(1-\pi)^{N-x}$$

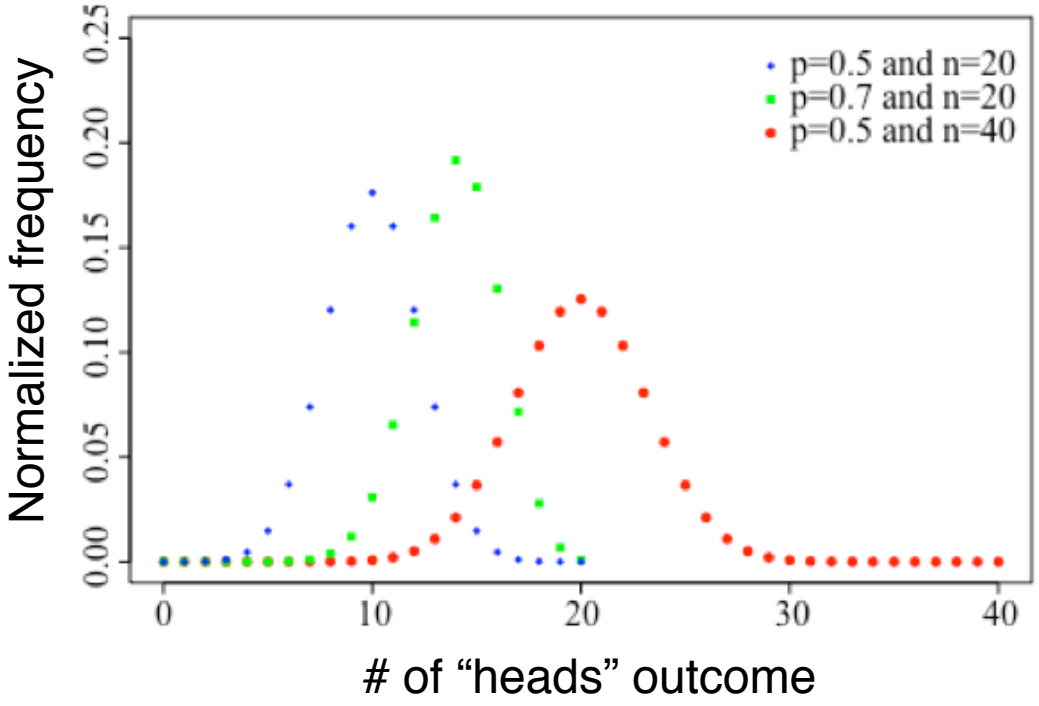
$$P(0) = \frac{2!}{0!(2-0)!}(.5^{0})(1-.5)^{2-0} = \frac{2}{2}(1)(.25) = 0.25$$

$$P(1) = \frac{2!}{1!(2-1)!}(.5^{1})(1-.5)^{2-1} = \frac{2}{1}(.5)(.5) = 0.50$$

$$P(2) = \frac{2!}{2!(2-2)!}(.5^{2})(1-.5)^{2-2} = \frac{2}{2}(.25)(1) = 0.25$$

Binomial Distribution

… is the discrete probability distribution of the number of successes (x) in a sequence of *N* independent yes/no experiments, each of which yields success with probability *p*

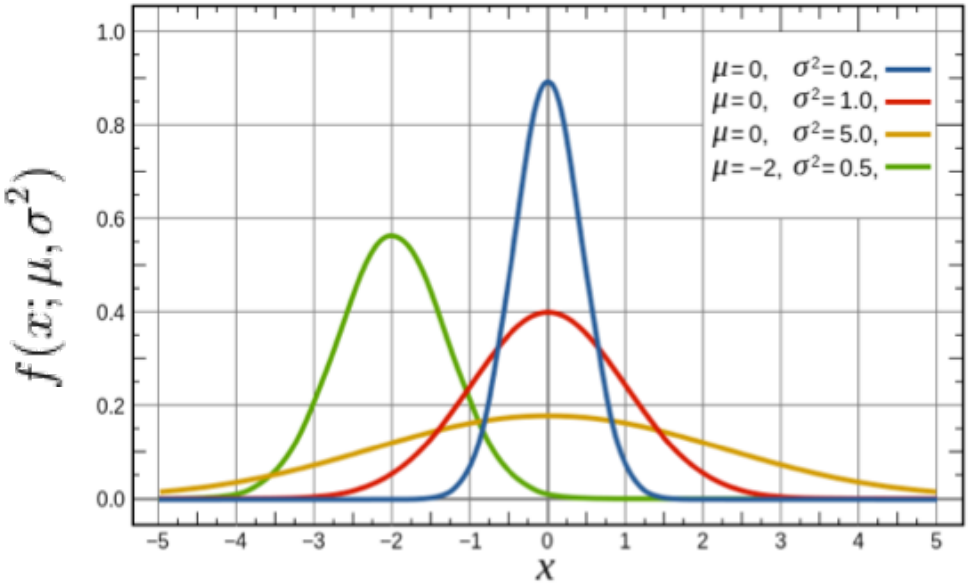


$$P(x) = \frac{N!}{x!(N-x)!}\pi^{x}(1-\pi)^{N-x}$$

N: number of flips
x: number of desired outcome
π: probability of the desired outcome

Gaussian/Normal Distributions

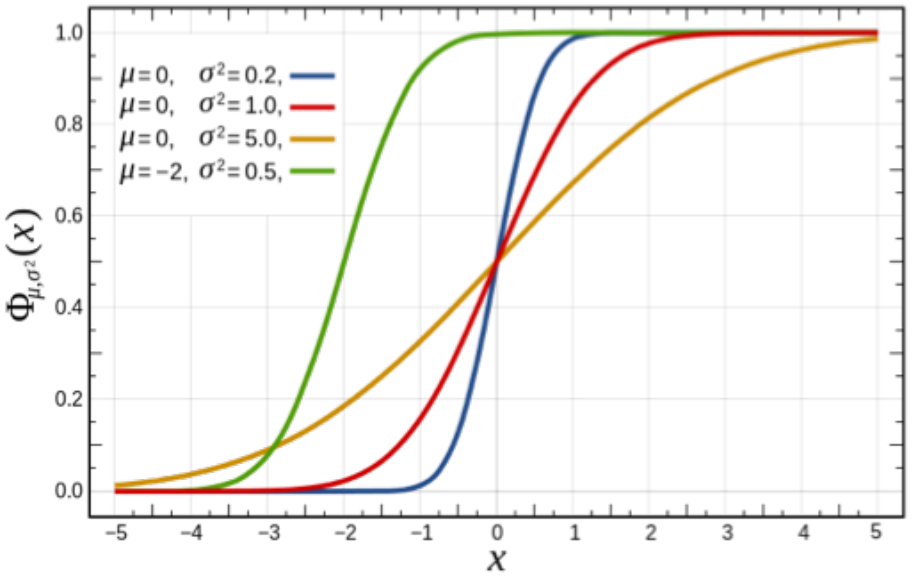
Carl Friedrich Gauss
(1777 - 1855)



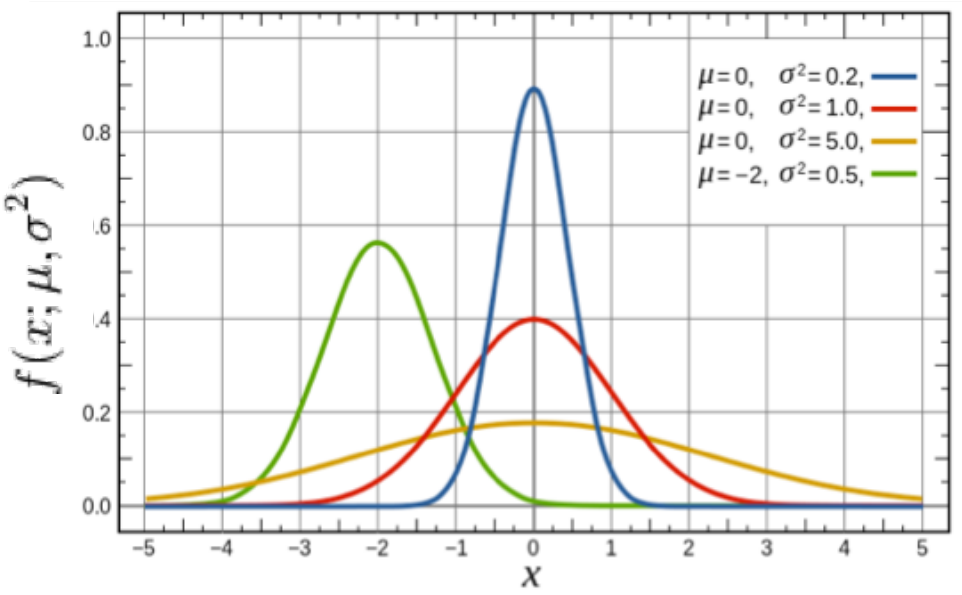
$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ = true mean of the distribution
σ^2 = variance of the observations
x = value of the observation(s)

Cumulative Distribution



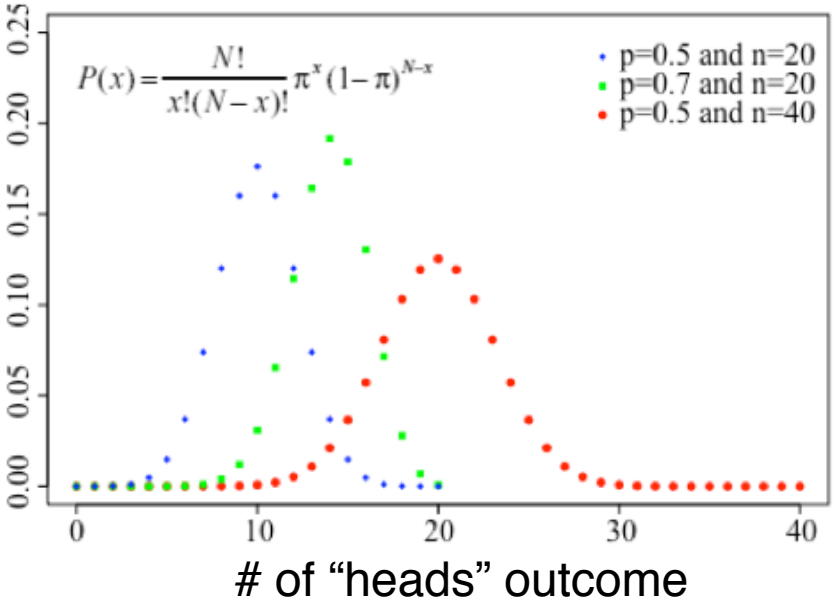
Normal Distribution

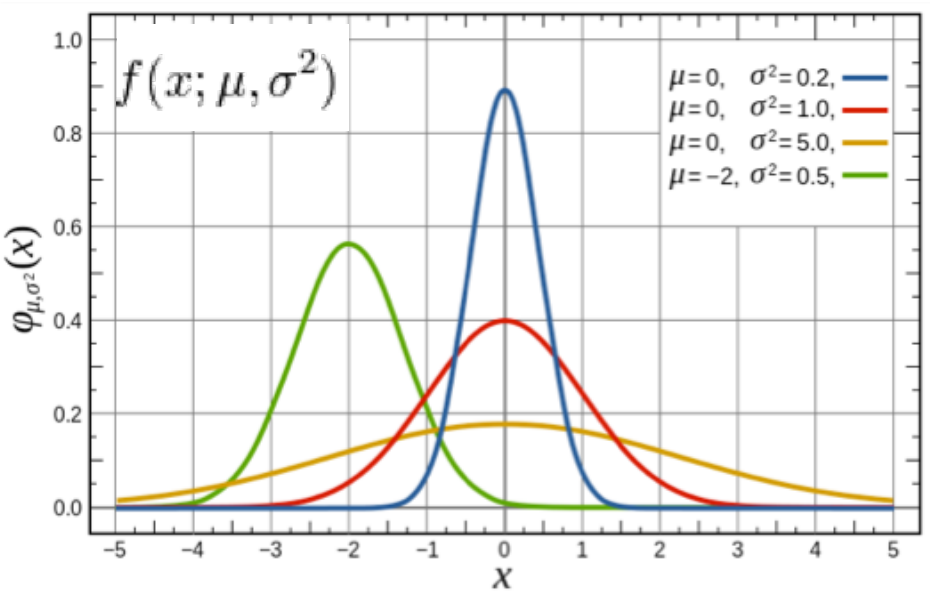


$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

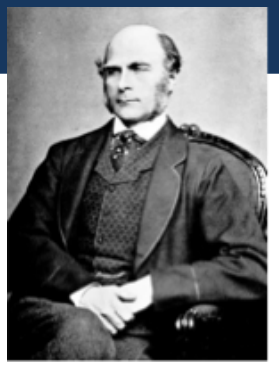
μ = true mean of the population
σ^2 = variance of the observations
x = value of the observation(s)

Binomial and Normal distributions



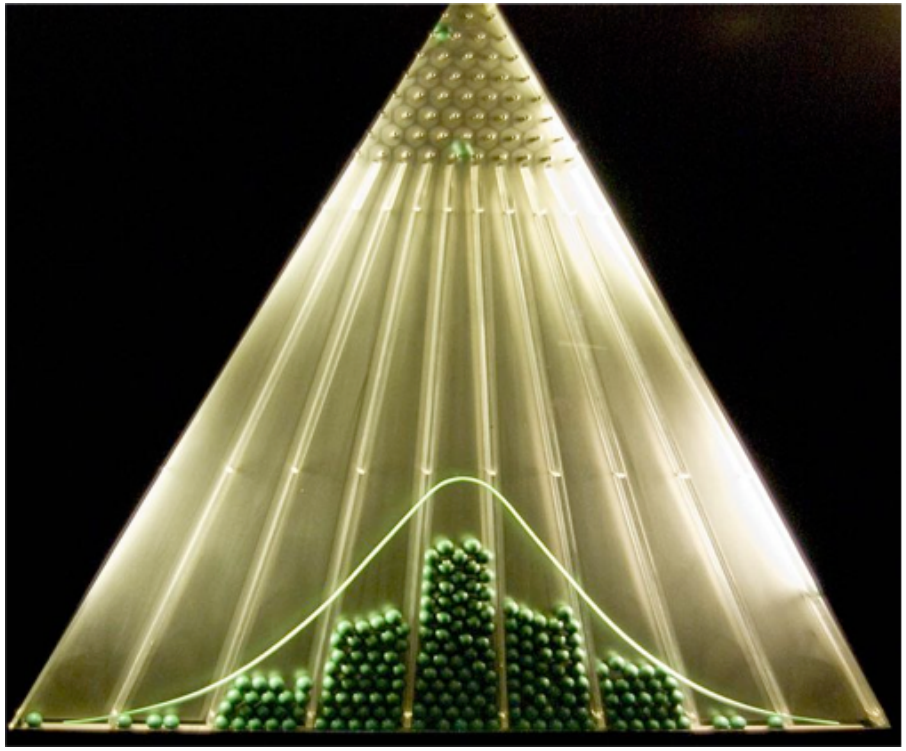


- binomial distributions deal with discrete data
- normal distributions deal with continuous data
- the bean-machine elegantly shows how the discrete data at the basic levels can grow to be characterized as a continuous distribution.



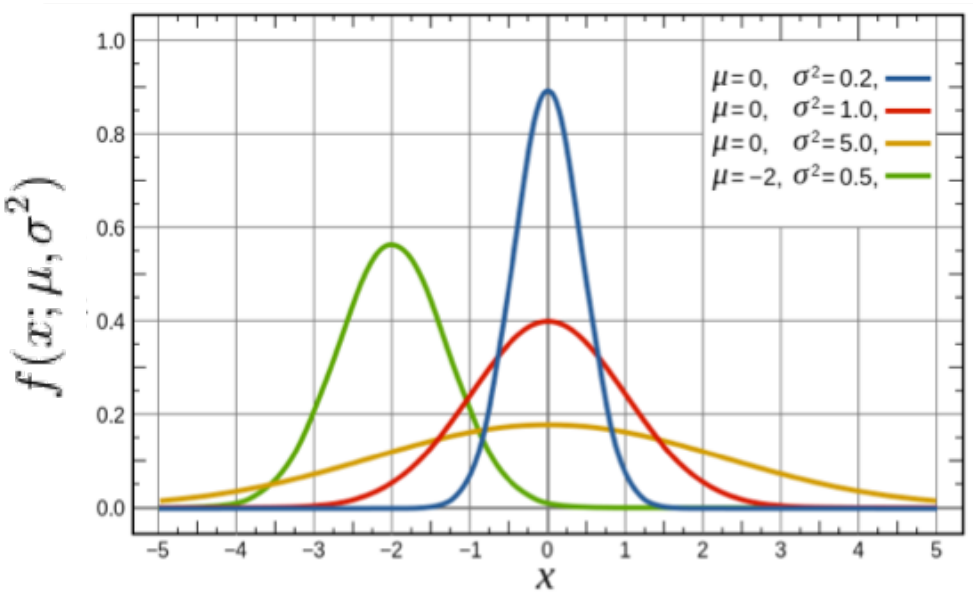
Francis Galton

Galton's bean machine



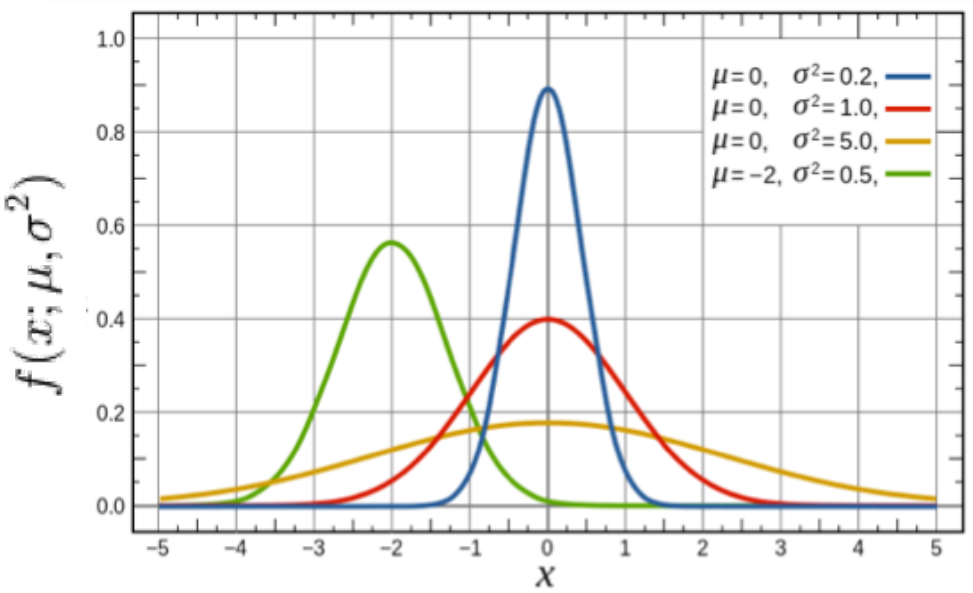
Video: https://youtu.be/Bampgm0HKDU

Normal Distribution



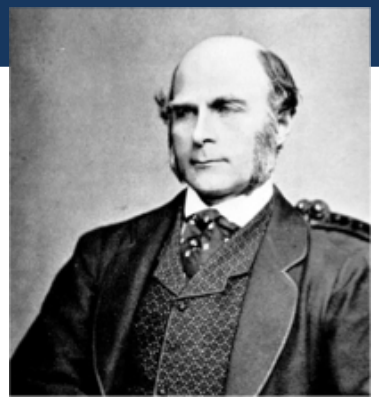
I can only recognize the occurrence of the normal curve – the Laplacian curve of errors – as a very abnormal phenomenon. It is roughly approximated to in certain distributions; for this reason, and on account for its beautiful simplicity, we may, perhaps, use it as a first approximation, particularly in theoretical investigations.
Pearson (1901)

Normal Distribution



Average & Variance

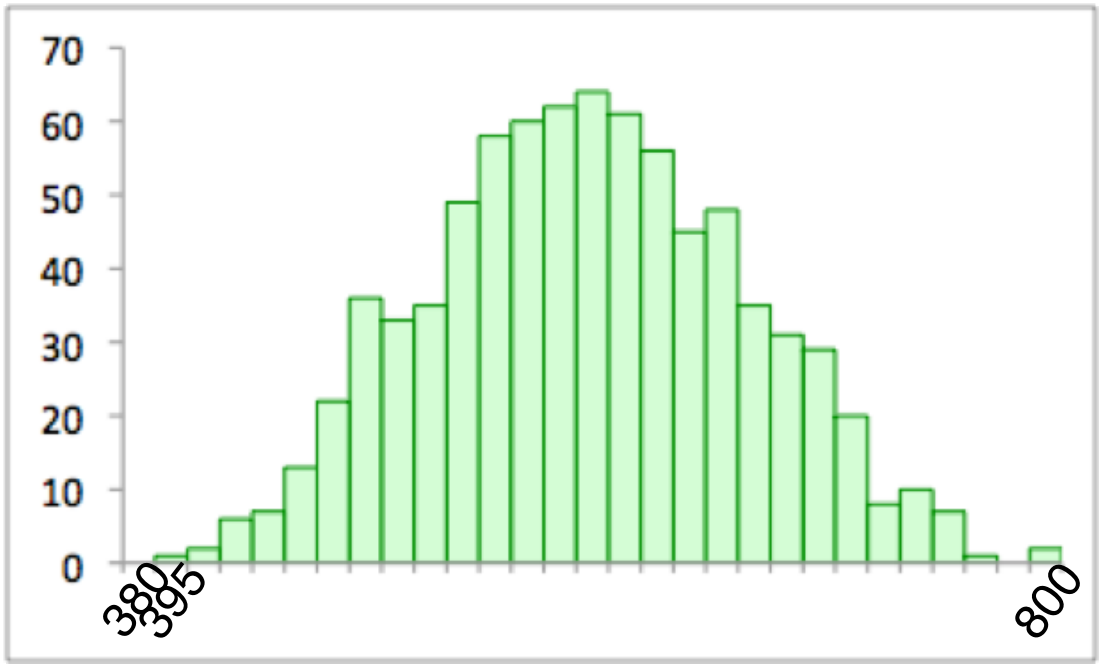
A story about Francis Galton (1906), an Ox & 801 normal observers



Based on a real story. The numbers are made up but the characters are real. No animals were harmed in this production.

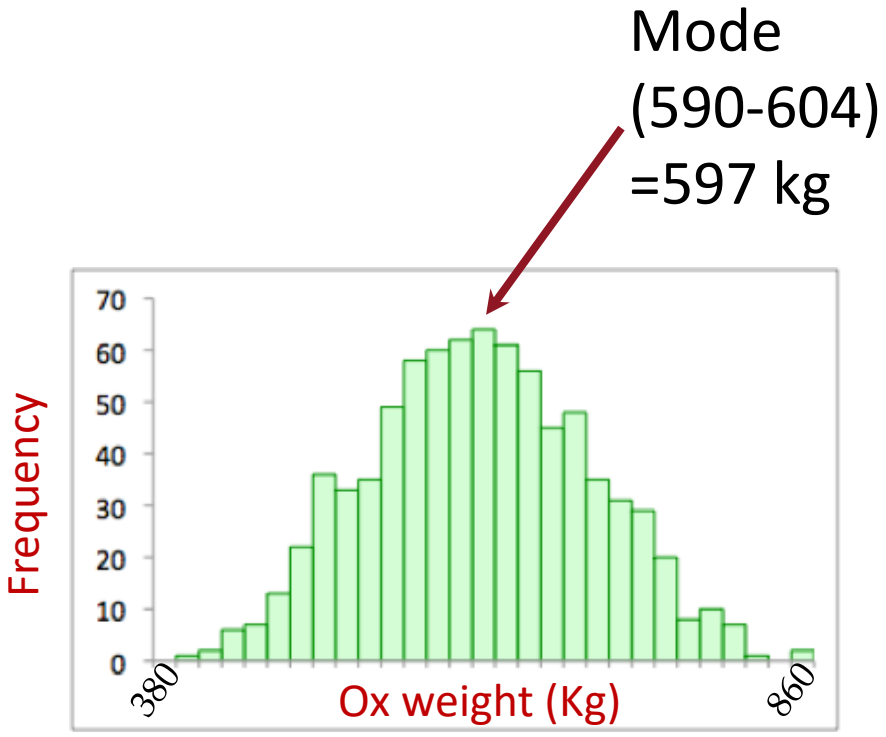
class intervals (kg)	class frequency
380	0
395	1
410	2
425	6
440	7
455	13
470	22
485	36
500	33
515	35
530	49
545	58
560	60
575	62
590	64
605	61
620	56
635	45
650	48
665	35
680	31
695	29
710	20
725	8
740	10
755	7
770	1
785	0
800	2

Creating a histogram

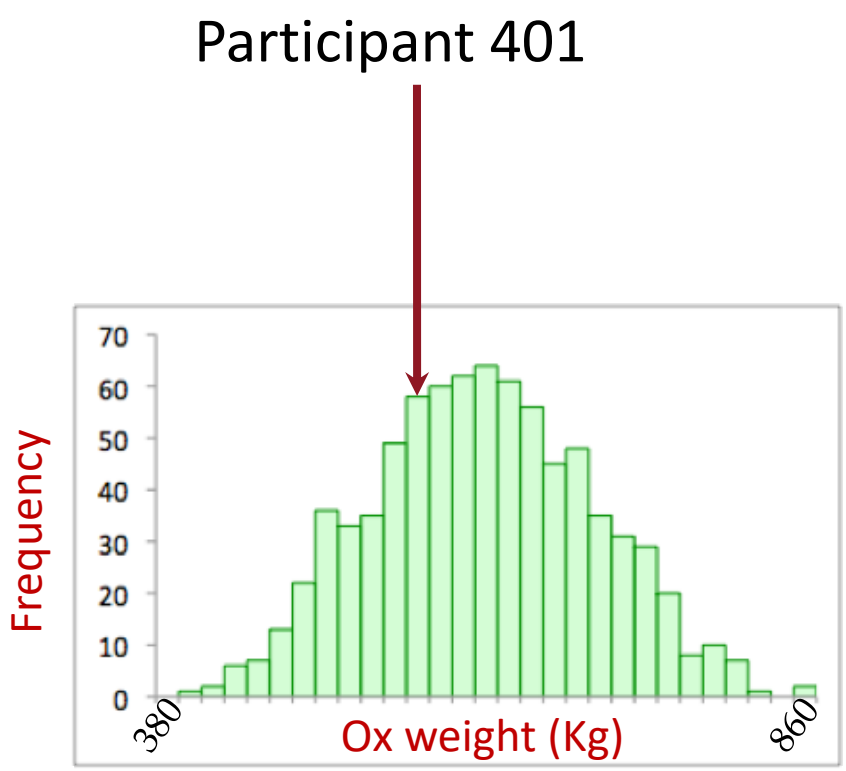


- determine class intervals (15 kg)
- count the number of times each score occurs for each class interval
- enter this data into a “frequency table”
- plot the data as a histogram

Observers' estimates



Median: the number separating the higher half from the lower half of sampled data.
Since we have 801 observations, it is observation #401.



x_1	380.21 kg
$x_{...}$	...
x_{400}	543.78 kg
x_{401}	547.48 kg ← Median
x_{402}	547.87 kg
$x_{...}$	...
x_{801}	858.1 kg

Sample Mean ($\bar{x}$):
estimate of the population mean

x_1	380.21 kg
$x_{...}$	...
x_{400}	543.78 kg
x_{401}	547.48 kg
x_{402}	547.87 kg
$x_{...}$	...
x_{801}	858.1 kg
$\bar{X} = (\sum x)/n$	**542.95 kg** ← Mean

Estimating the Central tendency from data

- Mode
  this is the most frequent observation, ideal for discrete data
- Median (M)
  this is the middle observation of distribution of data, when they are ranked in order of magnitude
- Mean of sample ($\bar{x}$)
  this is the the sum of all observed data, divided by the number of observations

Estimating the Central tendency from data

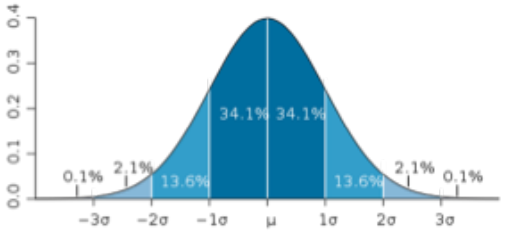
- Mode
  this is the most fre
- Median
  this is the middle
  ranked in order of
- Mean
  this is the the sum
  observations

These estimated central tendencies are NOT the true center (ie, μ) of the data distribution.

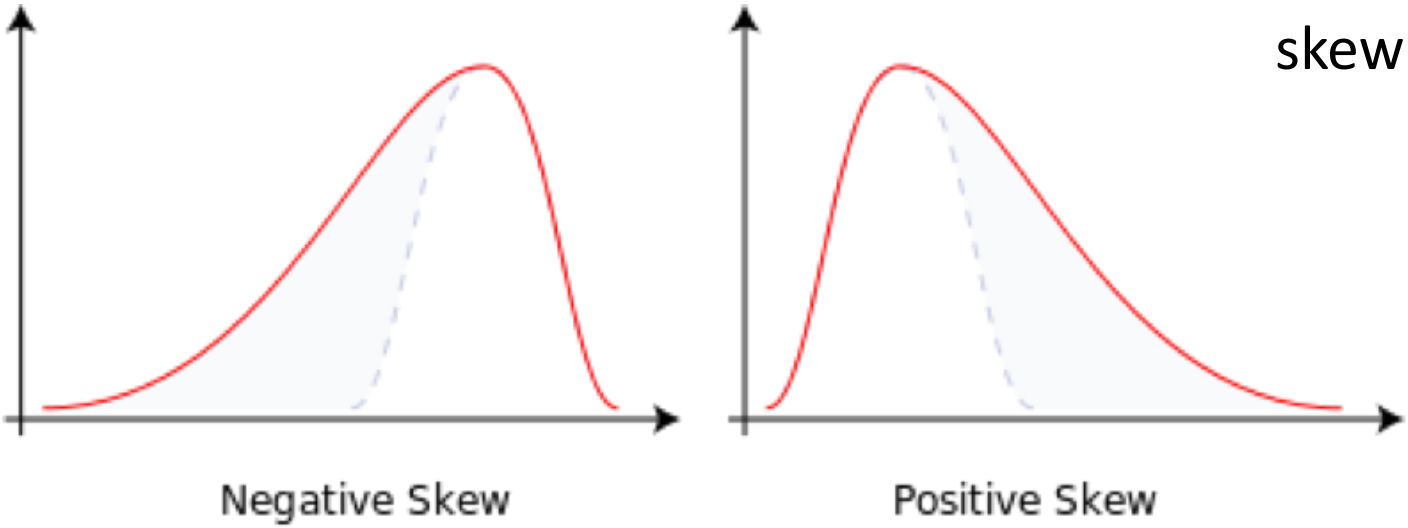
They are only estimates from the observed empirical data.

The more data you have, the closer your estimate will be to the true center.

Skewness



$$\text{skew} = \sum \frac{(X-\mu)^3}{\sigma^3}$$

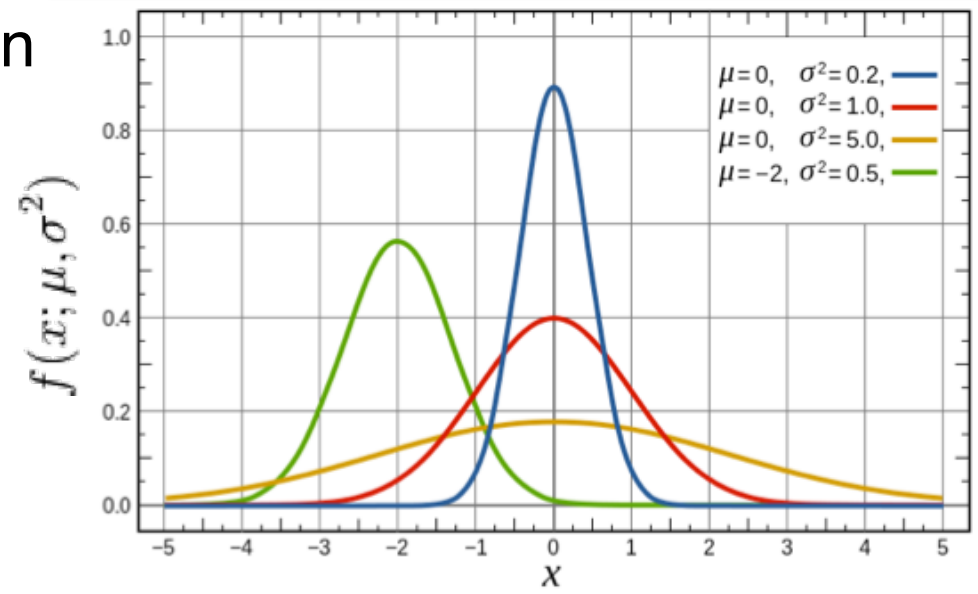


- refers to the asymmetry of the probability distribution
- if mean=median, if the distribution is symmetric
- if mean < median, distribution has a negative skew
- if mean > median, distribution has a positive skew

Estimating the variability of our observed data

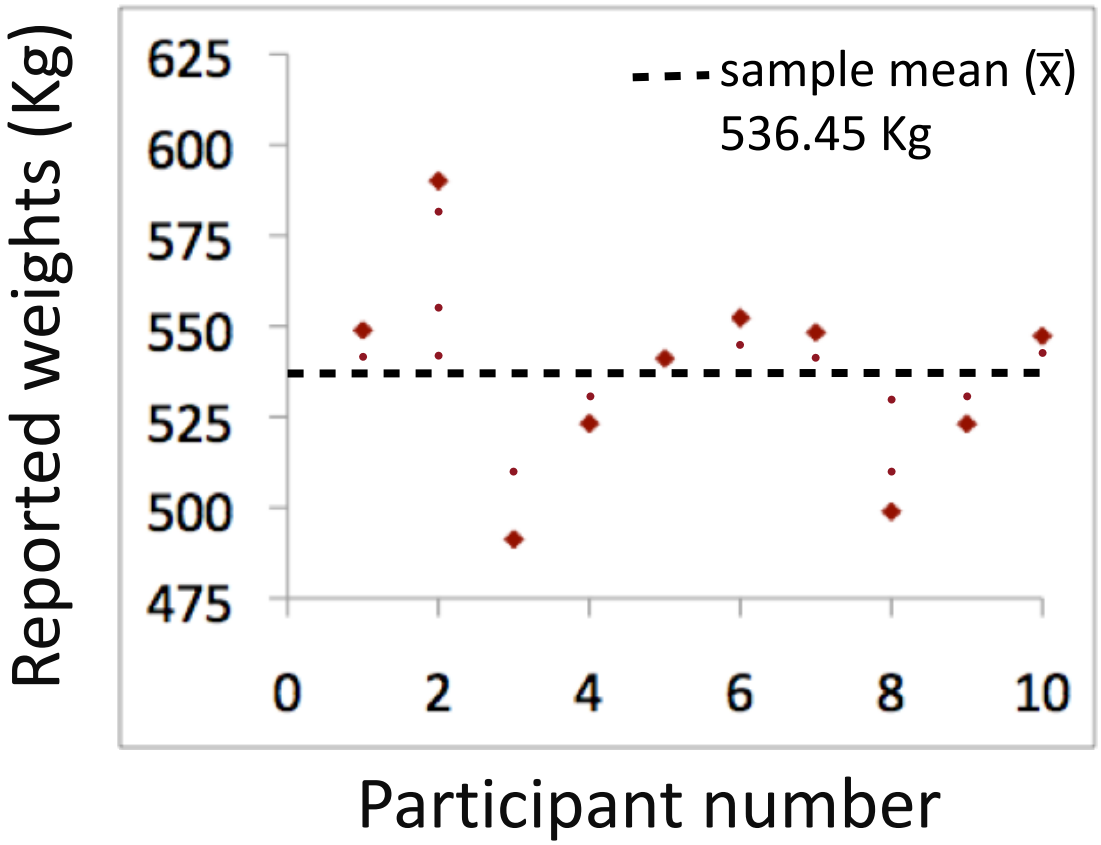
The variance of a distribution indicates the spread of the data.

That is, how different observations are to the mean.

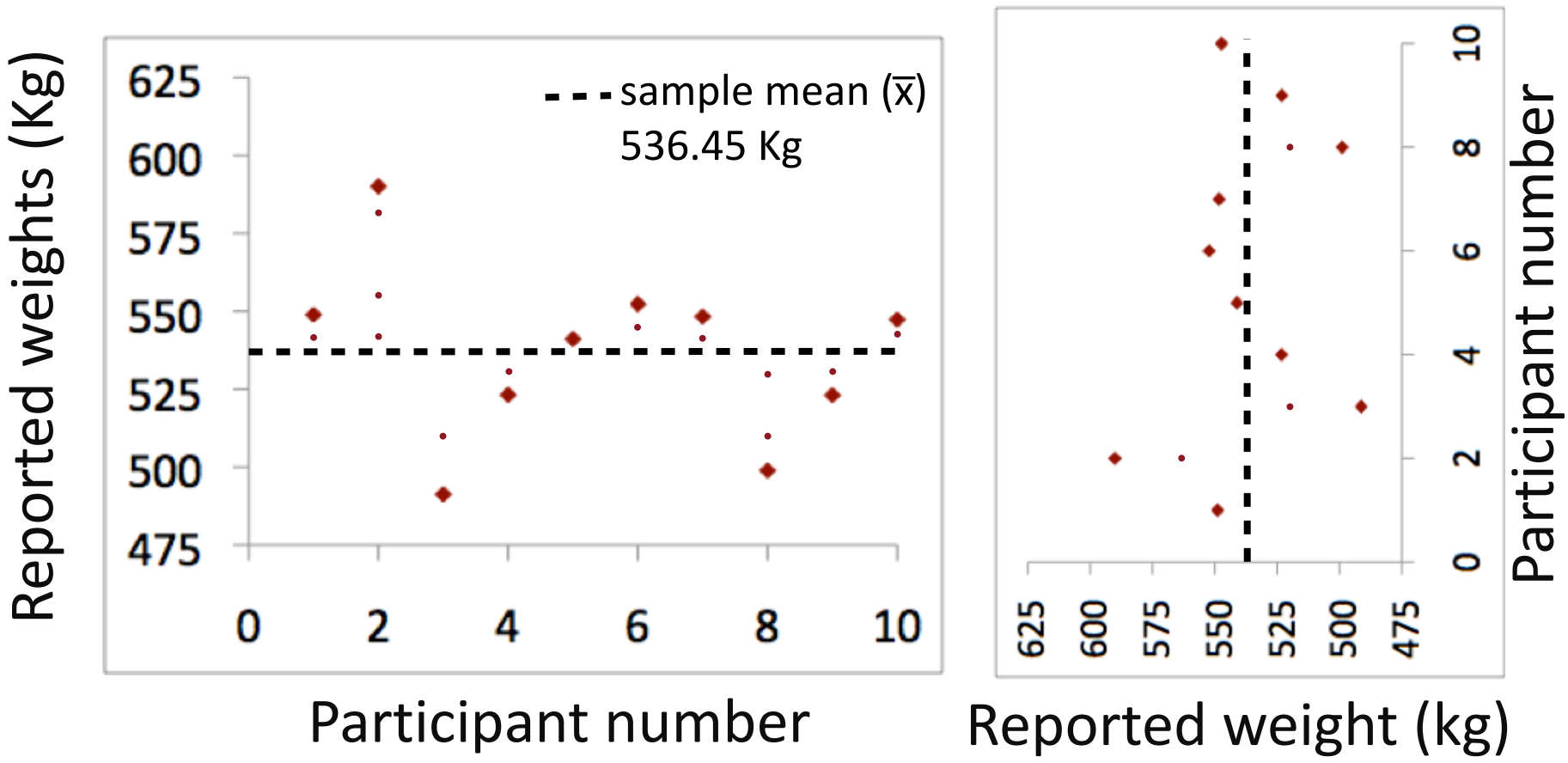


$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

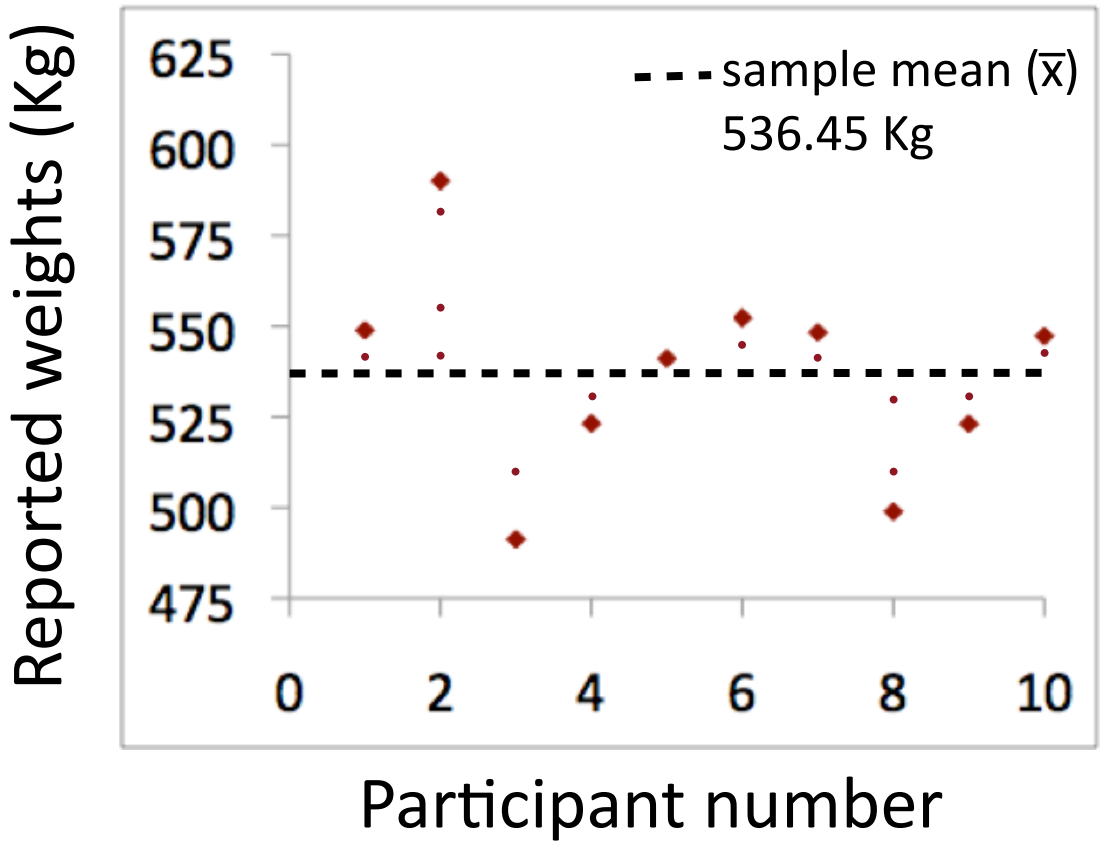
Estimating the variability of our observed data (n=10)



Estimating the variability of our observed data (n=10)

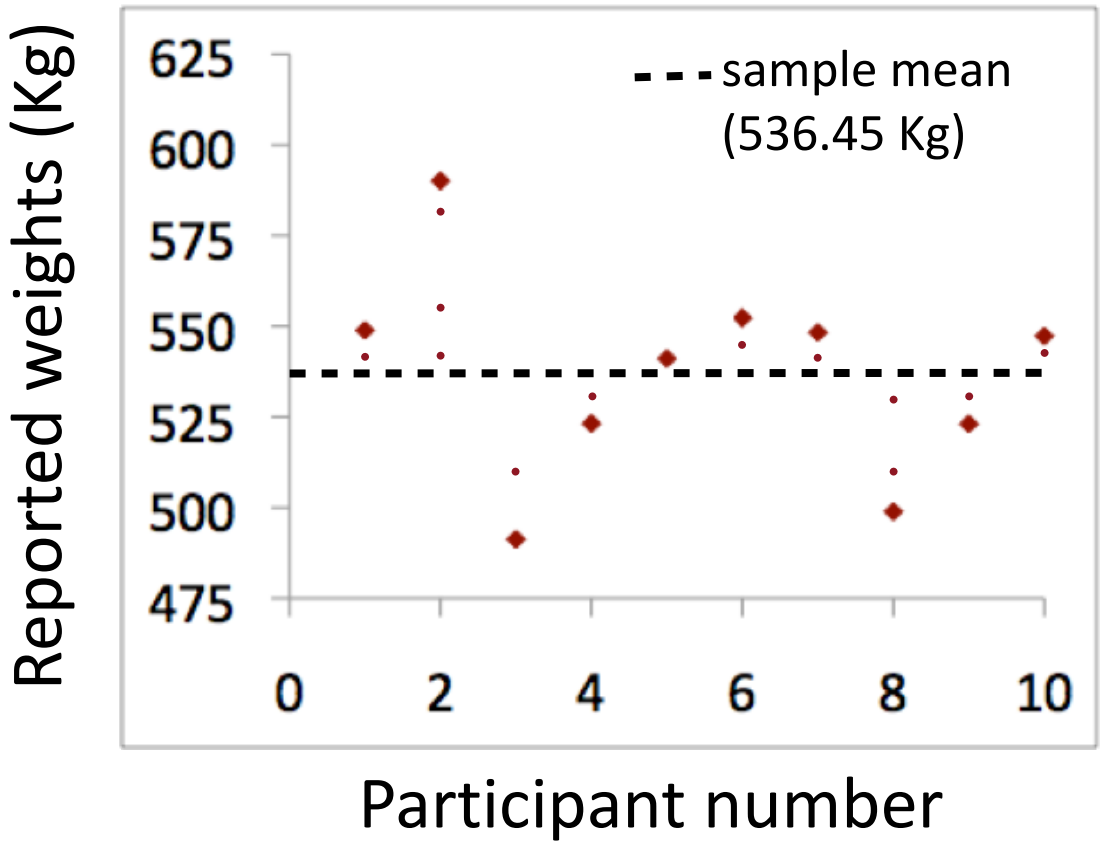


Estimating the variability of our observed data (n=10)



Raw score (kg)	Deviation (x_i-x_{mean}, [kg])
548,9	*548.9-536.45=* 12,45
590,1	53,65
491,3	-45,15
523,2	-13,25
541,1	4,65
552,3	15,85
548,3	11,85
498,9	-37,55
523,1	-13,35
547,3	10,85
$\sum (x_i - x_{mean}) = 0$ Kg	

Estimating the variability of our observed data



Raw score (kg)	(Deviation)^2 $(x_i-x_{mean})^2$, kg^2
548,9	155,0025
590,1	2878,3225
491,3	2038,5225
523,2	175,5625
541,1	21,6225
552,3	251,2225
548,3	140,4225
498,9	1410,0025
523,1	178,2225
547,3	117,7225
$\sum (x_i-x_{mean})^2 = 7366.625$ kg^2	

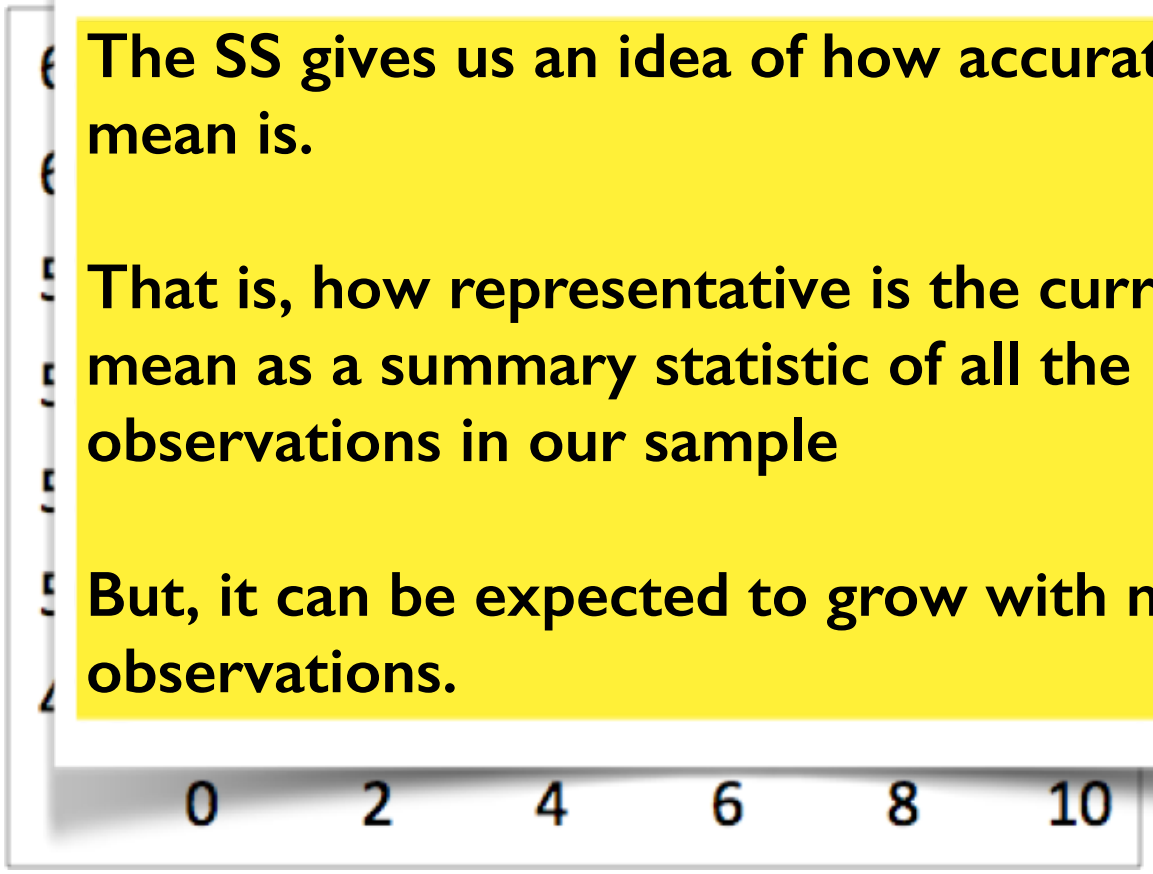
sum of squared errors (SS)

Estimating the variability of our observed data

The SS gives us an idea of how accurate the mean is.

That is, how representative is the current mean as a summary statistic of all the observations in our sample

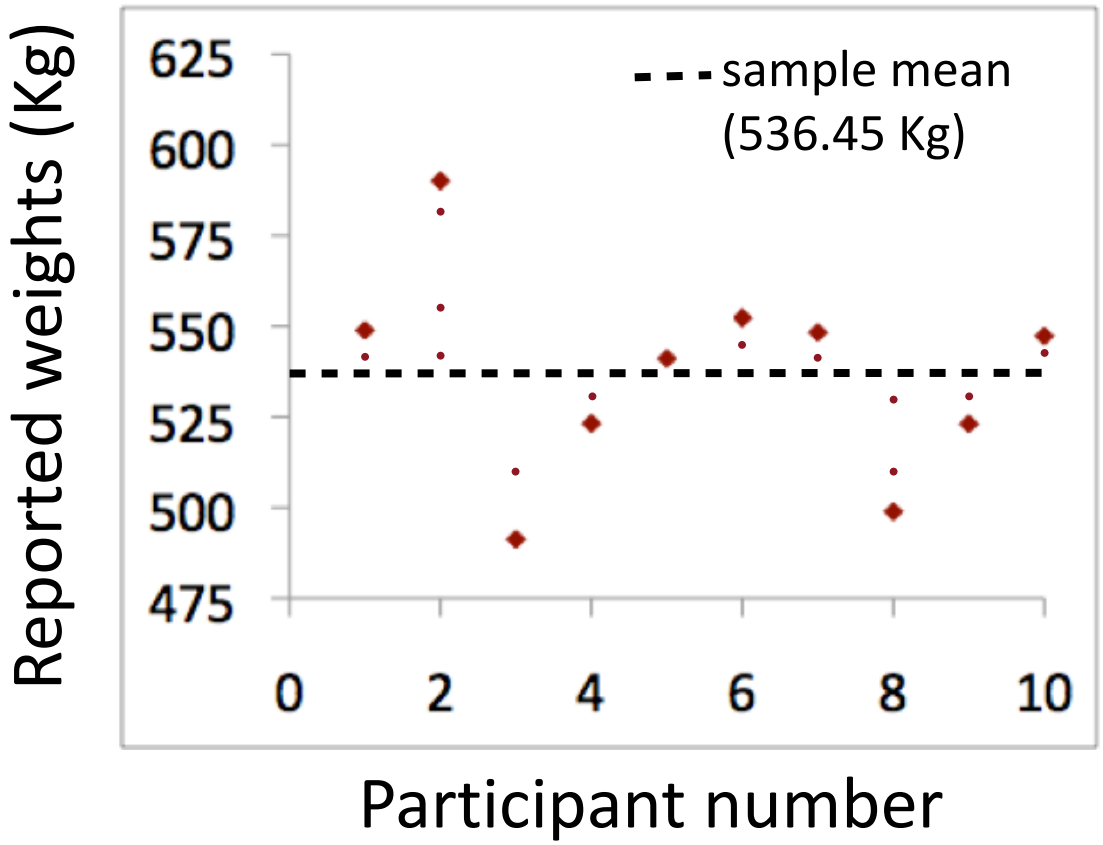
But, it can be expected to grow with more observations.



$(\text{Deviation})^2$ $(x_i - x_{mean})^2$, kg^2
155,0025
2878,3225
2038,5225
175,5625
21,6225
251,2225
140,4225
1410,0025
178,2225
547,3 117,7225
$\sum (x_i - x_{mean})^2 = 7366.625\ kg^2$

sum of squared errors (SS)

Estimating the variability of our observed data



Raw score (kg)	(Deviation)^2 $(x_i-x_{mean})^2$, kg^2
548,9	155,0025
590,1	2878,3225
491,3	2038,5225
523,2	175,5625
541,1	21,6225
552,3	251,2225
548,3	140,4225
498,9	1410,0025
523,1	178,2225
547,3	117,7225
$\sum (x_i-x_{mean})^2/N = 736{,}66$ kg^2	

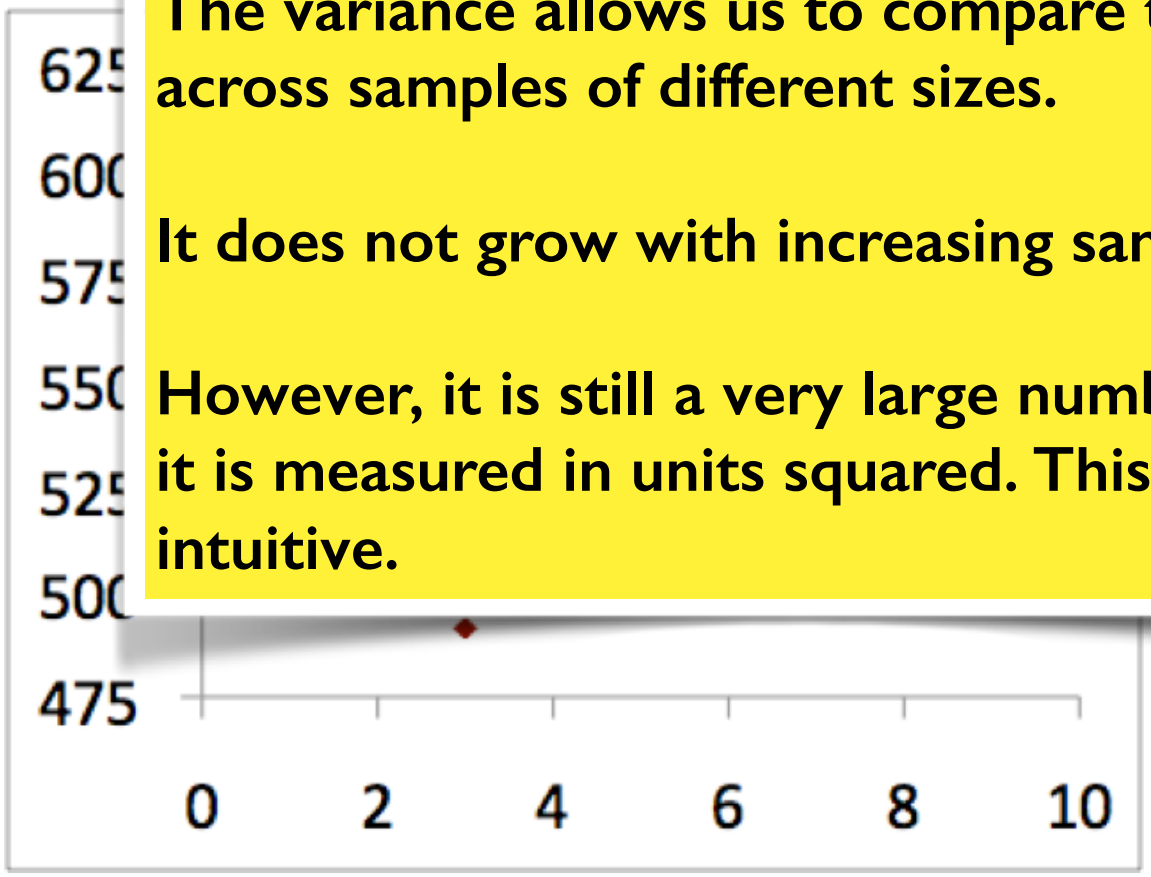
Variance of the sample (SS/N)

Estimating the variability of our observed data

The variance allows us to compare the "error" across samples of different sizes.

It does not grow with increasing sample size.

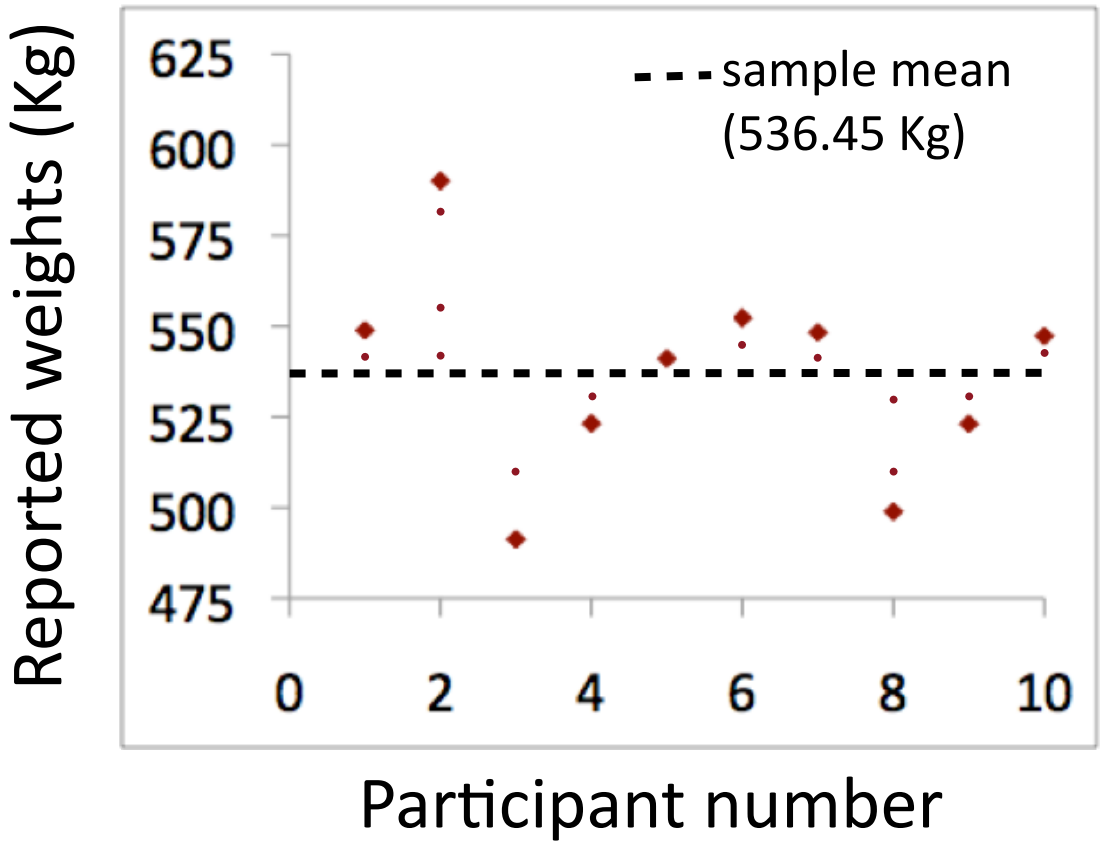
However, it is still a very large number because it is measured in units squared. This is not intuitive.



Deviation)^2 x_i-x_{mean})^2, kg^2
155,0025
2878,3225
2038,5225
175,5625
21,6225
251,2225
140,4225
498,9	1410,0025
523,1	178,2225
547,3	117,7225
$\sum (x_i\text{-}x_{mean})^2 / N = 736{,}66$ kg^2

Variance of the sample (SS/N)

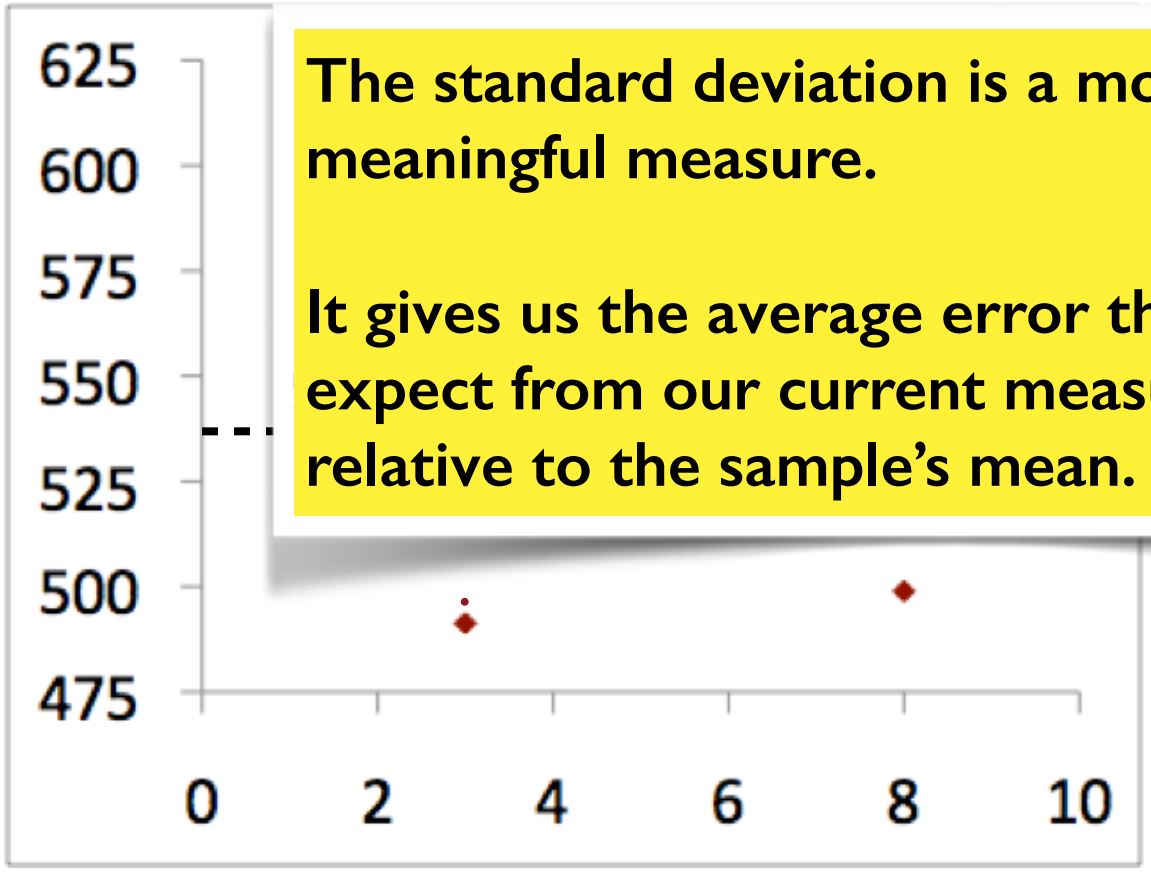
Estimating the variability of our observed data



Raw score (kg)	(Deviation)^2 $(x_i-x_{mean})^2$, kg^2
548,9	155,0025
590,1	2878,3225
491,3	2038,5225
523,2	175,5625
541,1	21,6225
552,3	251,2225
548,3	140,4225
498,9	1410,0025
523,1	178,2225
547,3	117,7225
$\sqrt{\sum (x_i-x_{mean})^2/N} = 27.14$ kg	

Standard deviation of the sample

Estimating the variability of our observed data



The standard deviation is a more intuitive and meaningful measure.

It gives us the average error that we can expect from our current measurements, relative to the sample's mean.

Raw score	(Deviation)^2
	^{2}, kg^2
	55,0025
	78,3225
	38,5225
	75,5625
	21,6225
	51,2225
548,3	140,4225
498,9	1410,0025
523,1	178,2225
547,3	117,7225
$\sqrt{\sum (x_i - x_{mean})^2 / N} = 27.14$ kg	

Standard deviation of the sample

Estimating the variability of our data sample

Population

Often, we are interested in the mean (μ) and the variance of the **population**.

These formulas allow us to calculate the variance (σ^2) and standard deviation (σ) of the population.

- Variance of the population (σ^2)
  $(\sum (x_i-\mu)^2)/N$
- standard deviation of the population (σ)
  $[(\sum (x_i-\mu)^2)/N]^{0.5}$

Terms

- μ is the population mean
- σ^2 is the population variance
- σ is the population standard deviation

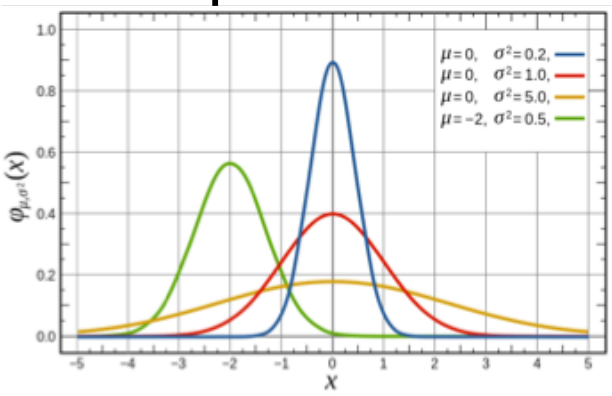
- $\bar{x}$ is the sample mean
- s^2 is the corrected sample variance *(i.e. with Bessel's correction)*
- s is the corrected sample standard deviation *(i.e. with Bessel's correction)*

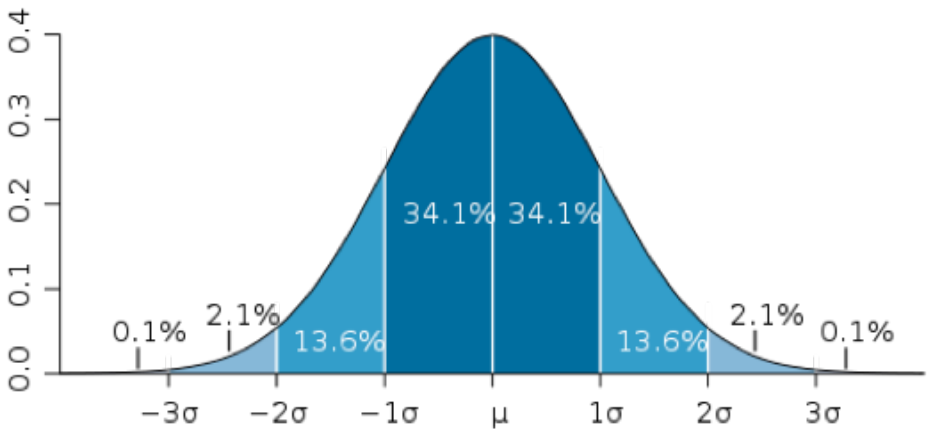
R

- Paweł, this is a good time to do descriptives in R

The usefulness of knowing the mean and the standard deviation of the population

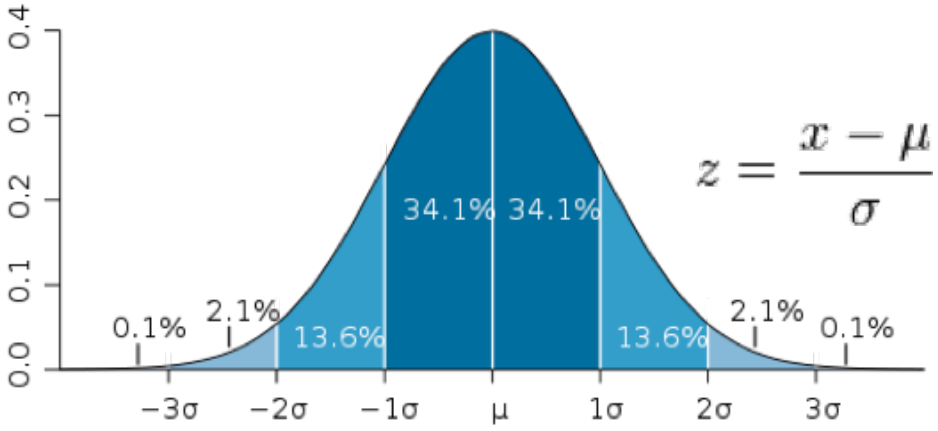
- 68.2% of the population should fall within ±1σ around the μ
- 95.4% of the population should fall within ±2σ around the μ
- 99.7% of the population should fall within ±3σ s.d. around μ





Standardizing the normal distribution

- all normal distributions can be treated the same if we treat measures in units of σ, around the mean μ as the center.
- critical z-scores
  - 1.645 (95.0%:5.0%)
  - 1.960 (97.5%:2.5%)
  - 2.576 (99.5%:0.5%)

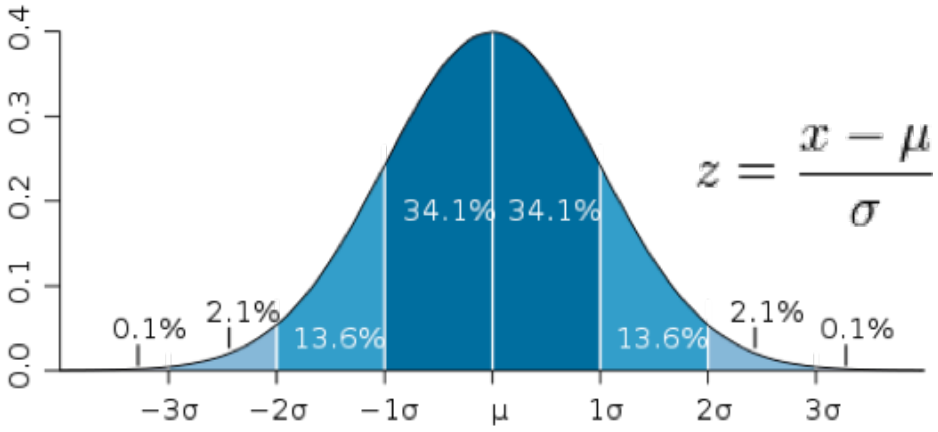


Standardizing the normal distribution

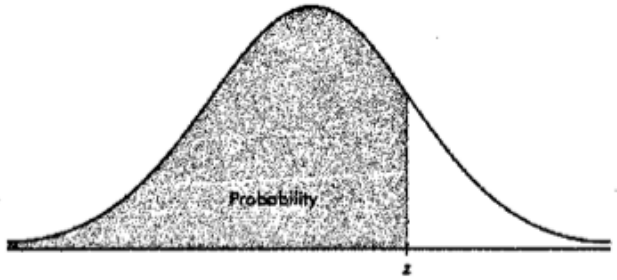
- all normal distributions can be treated the same if we measure in units of σ, around the mean μ as the center.

- These units of σ are referred to as z-scores

If we know that the N(180.17, 9.35) describes the height distribution of German males, how much taller/shorter are you relative to German men?



The z-table



- the z-score is the multiple of σ (i.e., z=2.53 is 2.53σ)
- if my height is 203, my z-score is:
  (203-180.17)/9.35
  =2.44
- to read the z-table,
  1) look down the left column (2.4)
  2) look across the row (0.04)
  3) p=0.9927
  4) a z-score of 2.44 is a value that is larger than 99.27% of the population

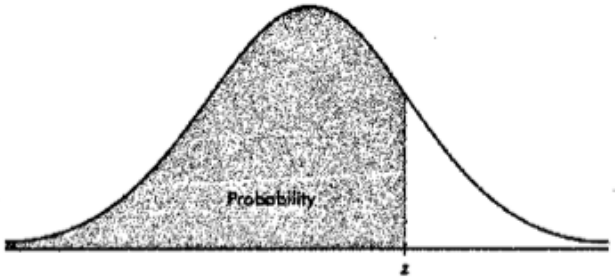
Rows are for 1st decimal number, columns for 2nd decimal number

TABLE A: STANDARD NORMAL PROBABILITIES (CONTINUED)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.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	.7704	.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

The z-table

- the z-score is the multiple of s (i.e., z=2.53 is 2.53*s)
- if my height is 203, my z-score is 2.44
- to find out the % of German men whom I'm taller than:
  - look down the row (2.4)
  - look across the col (0.04)
- I am taller than 99.21% of German men!!!



Rows are for 1st decimal number, columns for 2nd decimal number

TABLE A: STANDARD NORMAL PROBABILITIES (CONTINUED)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.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	.7704	.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