The Normal Distribution & Descriptive Statistics

Kin 304W

Week 2: May 14, 2013

Writer's Corner

Grammar Girl
Quick and Dirty Tips
For Better Writing

http://grammar.quickanddirtytips.com/

Writer's Corner

What is wrong with the following sentence?

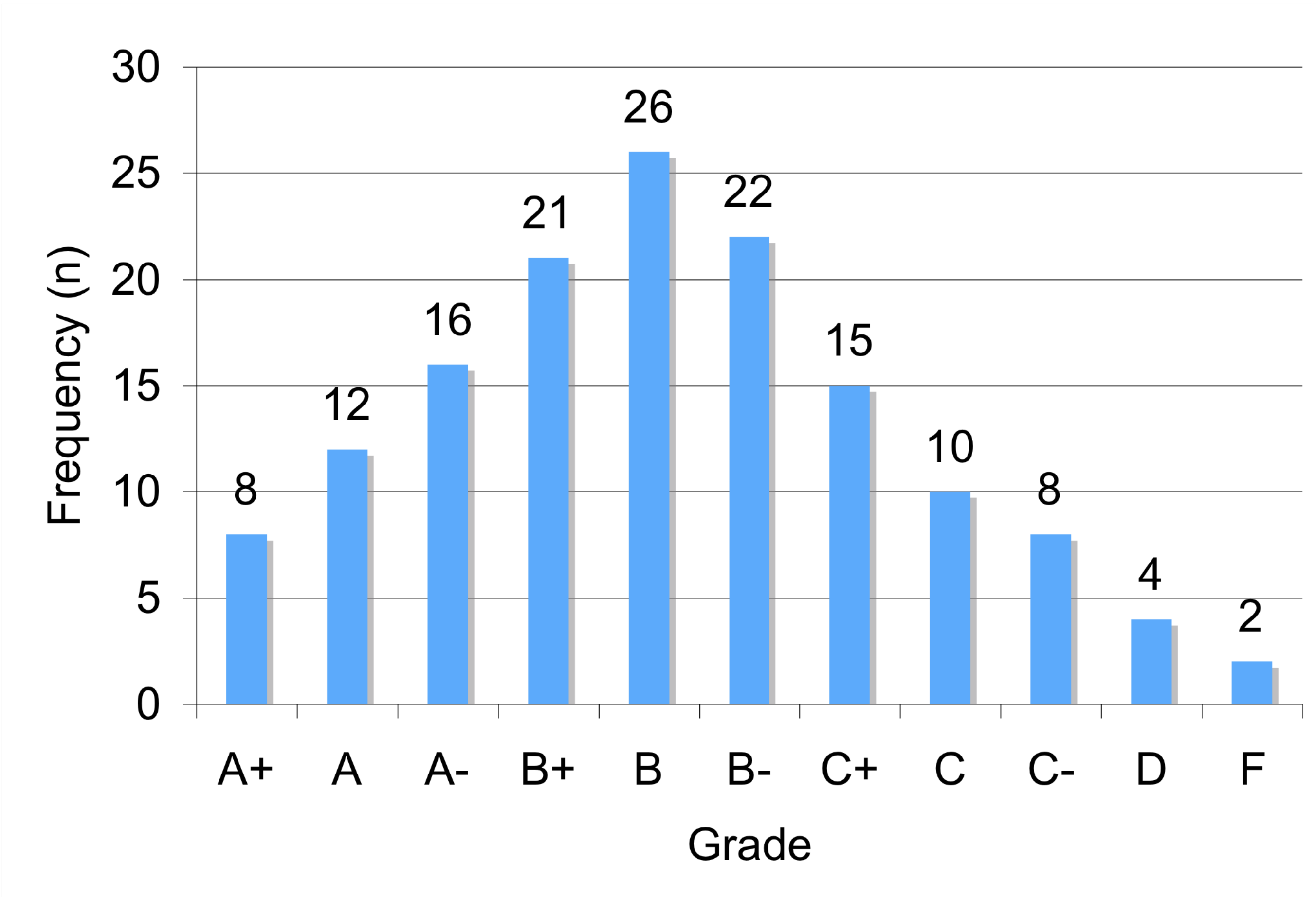
"This data is useless because it lacks specifics."

Outline

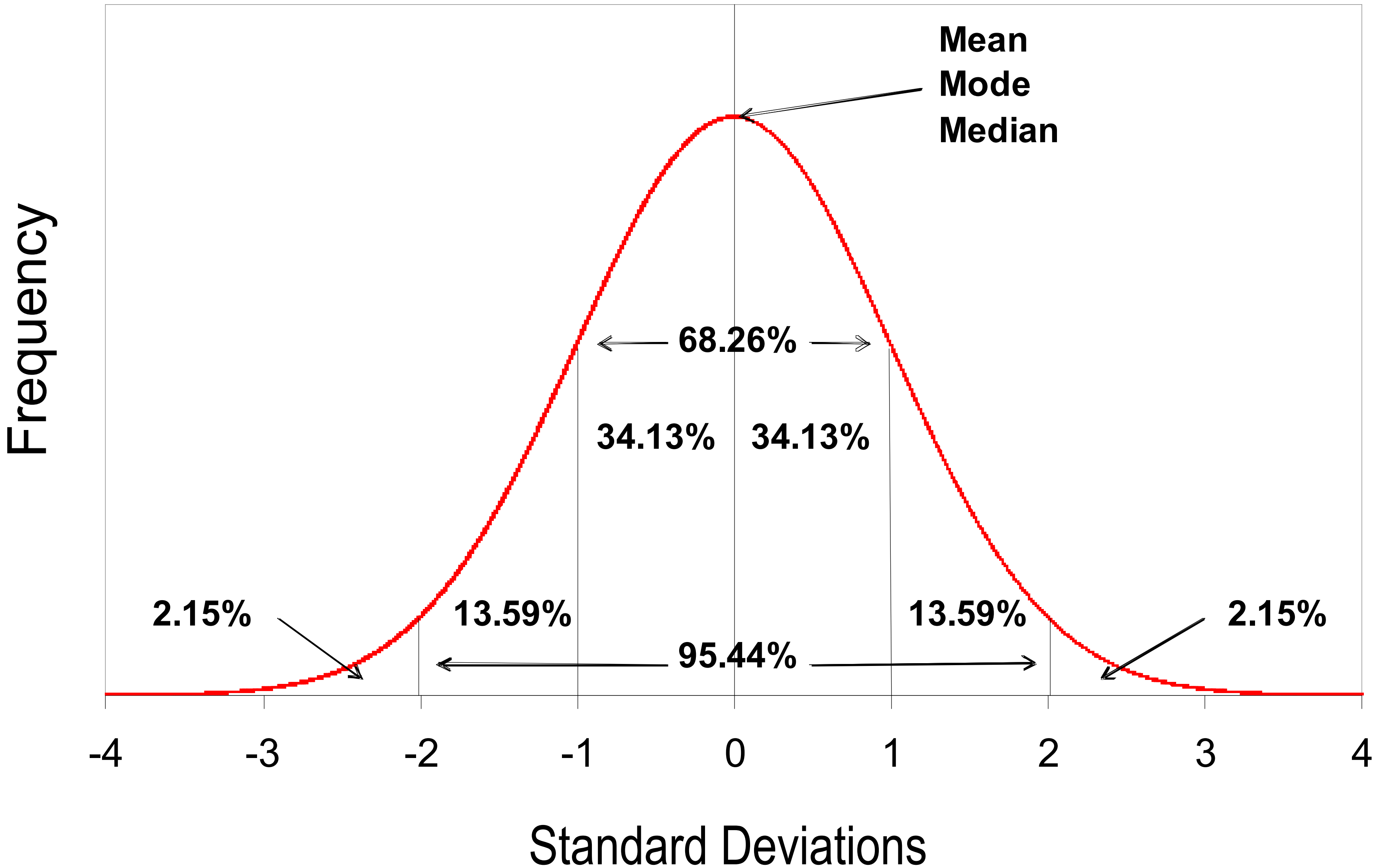
- Normal Distribution
- Testing Normality with Skewness & Kurtosis
- Measures of Central Tendency
- Measures of Variability
- Z-Scores
- Arbitrary Scores & Scales
- Percentiles

Frequency Distribution

Histogram of hypothetical grades from a second-year chemistry class (n=144)



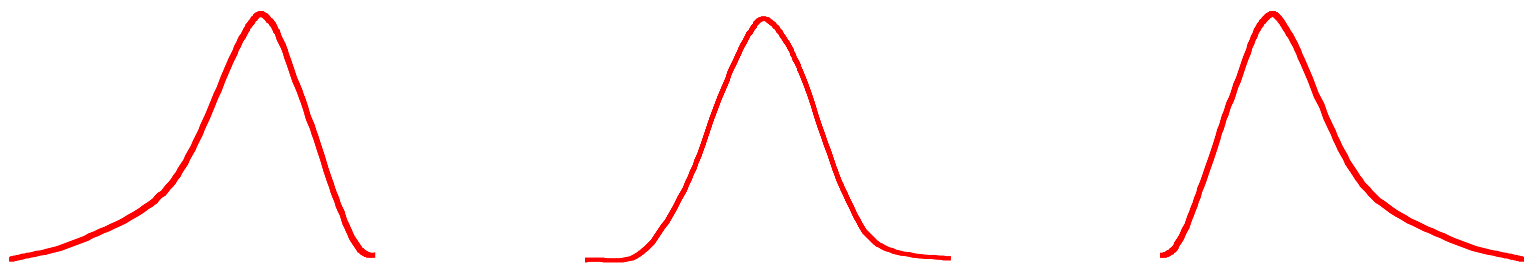
Normal Frequency Distribution



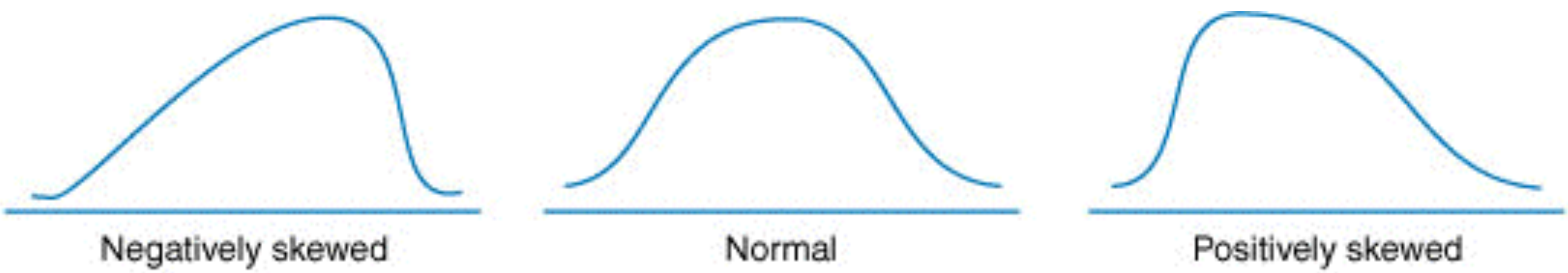
Skewness & Kurtosis

- Deviations in shape from the Normal distribution.
- Skewness is a measure of symmetry, or more accurately, lack of symmetry.
  - A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.
- Kurtosis is a measure of peakedness.
  - A distribution with high kurtosis has a distinct peak near the mean, declines rather rapidly, and has heavy tails.
  - A distribution with low kurtosis has a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case.

Skewness - Measure of Symmetry



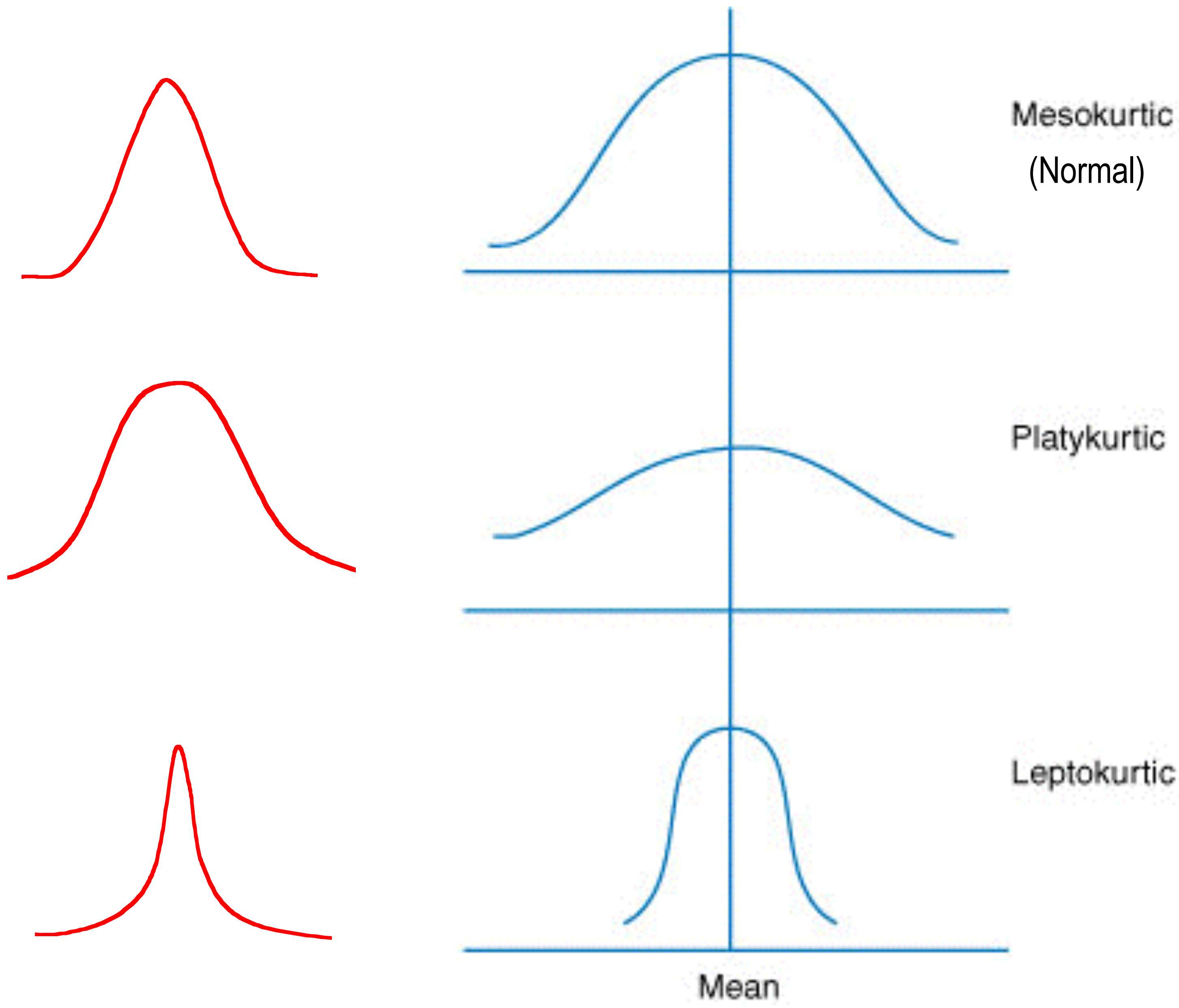
Negatively skewed | Normal | Positively skewed



Many variables in BPK are positively skewed.

Can you think of examples?

Kurtosis - Measure of Peakedness



Coefficient of Skewness

$$skewness = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^3}{(N-1)s^3}$$

Where: $\overline{X}$ = mean, X_i = X value from individual i, N = sample size, s = standard deviation

A perfectly Normal distribution has Skewness = 0

If -1 ≤ Skewness ≤ +1, then data are Normally distributed

Coefficient of Kurtosis

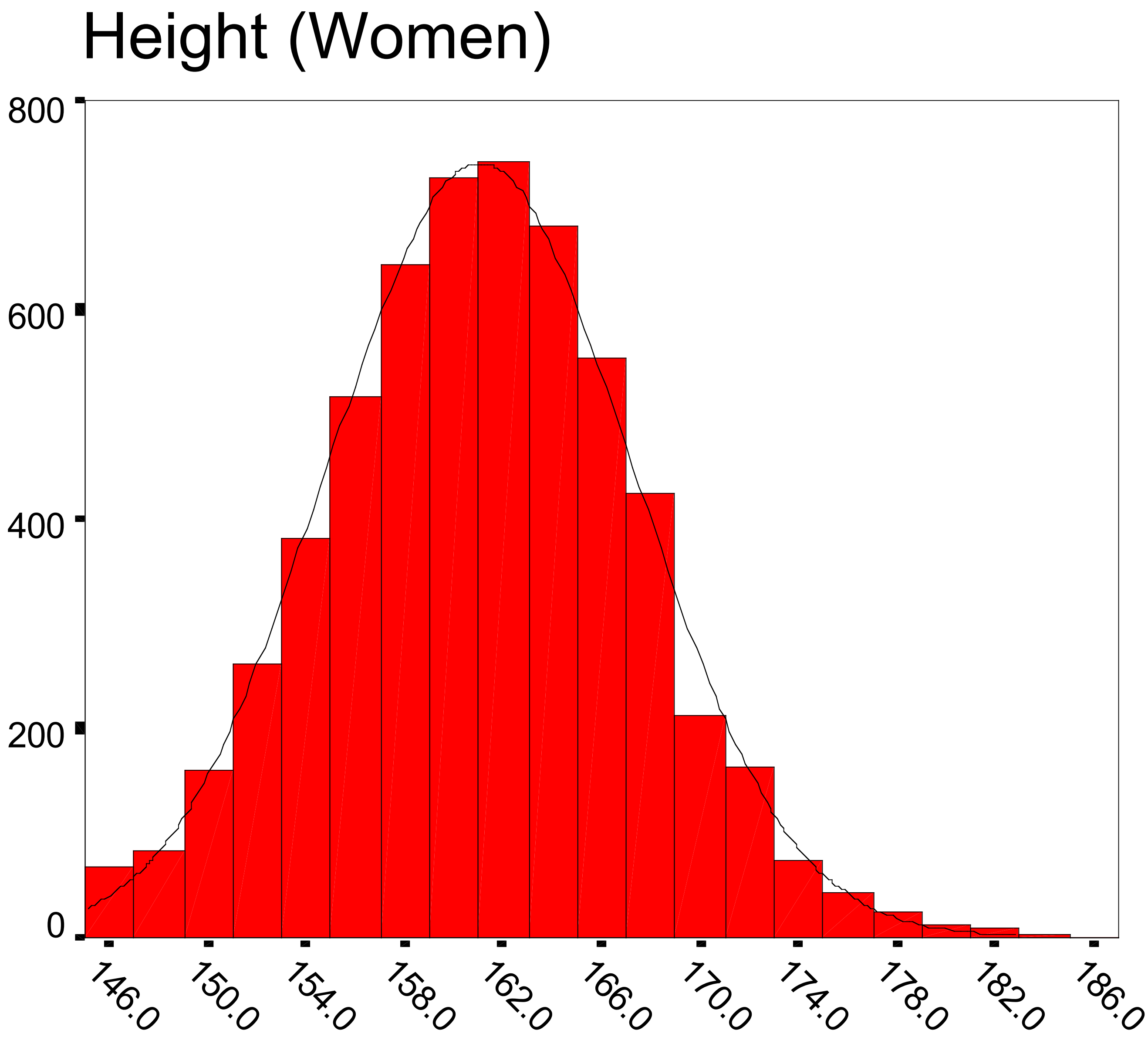
$$kurtosis = \frac{\sum_{i=1}^{N} (X_i - \overline{X})^4}{(N-1)s^4}$$

Where: $\overline{X}$ = mean, X_i = X value from individual i N = sample size, s = standard deviation

A perfectly Normal distribution has Kurtosis = 3 based on the above equation.

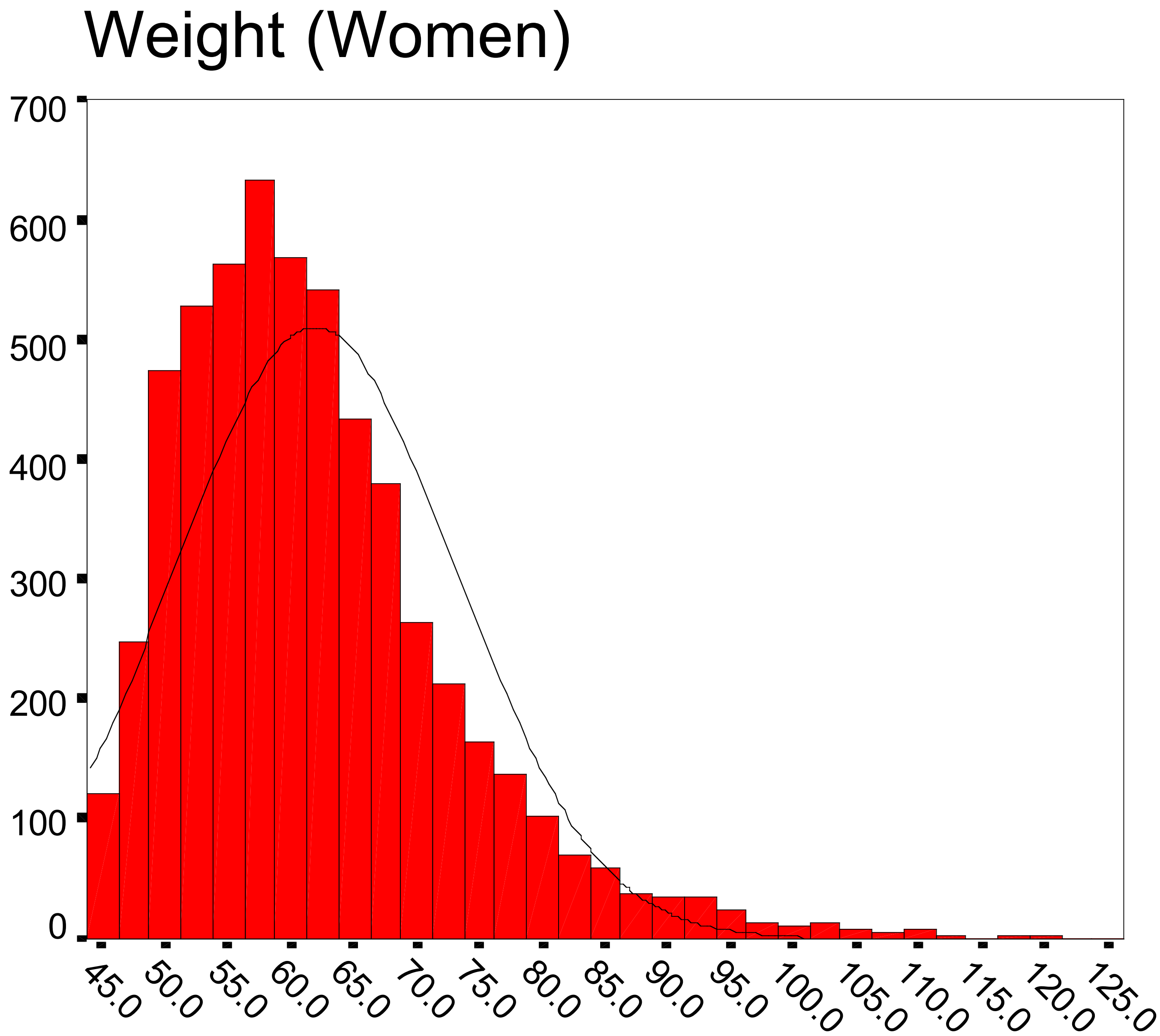
However, SPSS and other statistical software packages subtract 3 from kurtosis values. Therefore, a kurtosis value of 0 from SPSS indicates a perfectly Normal distribution.

Is Height in Women Normally Distributed?



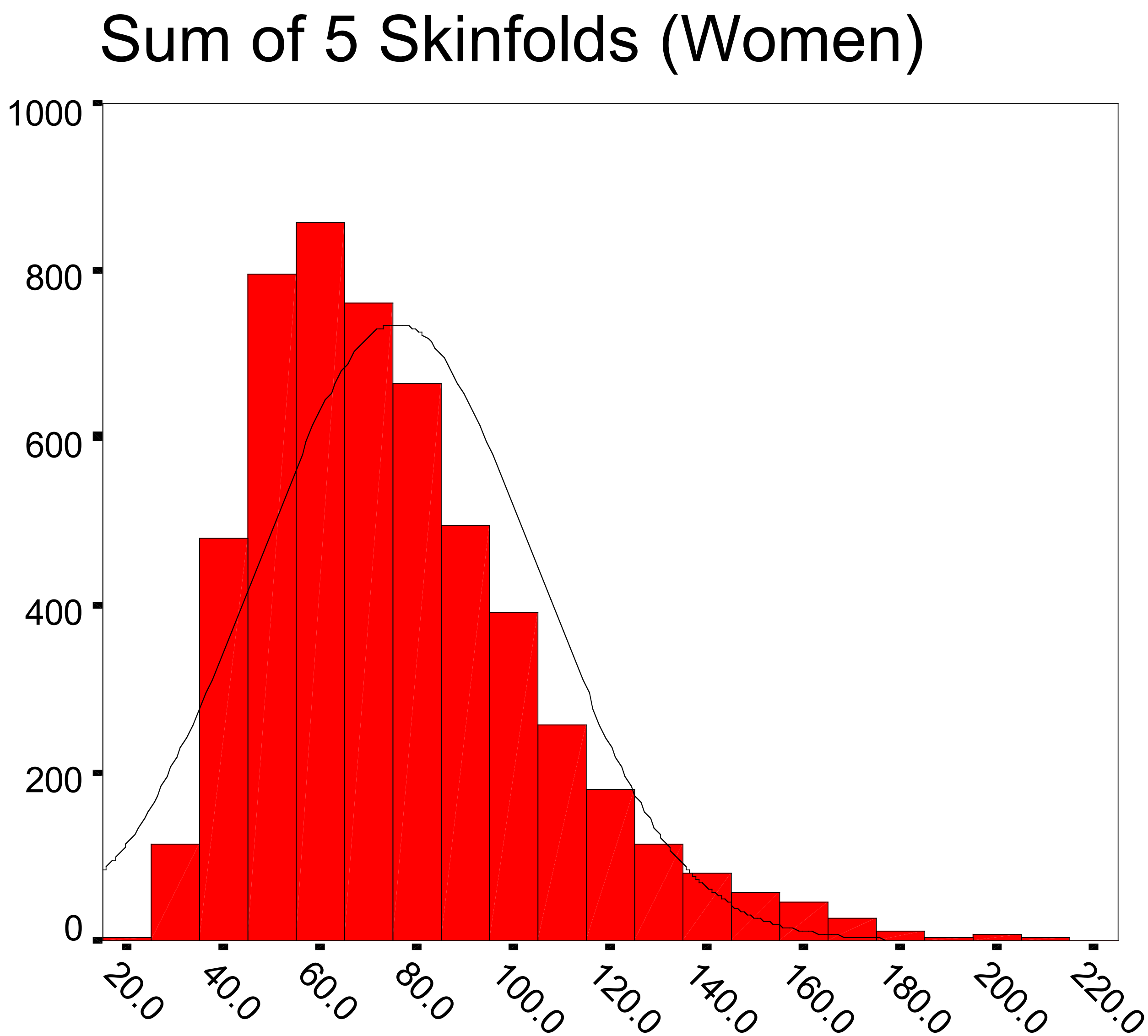
N = 5782
Mean = 161.0 cm
SD = 6.2 cm
Skewness = 0.092
Kurtosis = 0.090

Is Weight in Women Normally Distributed?

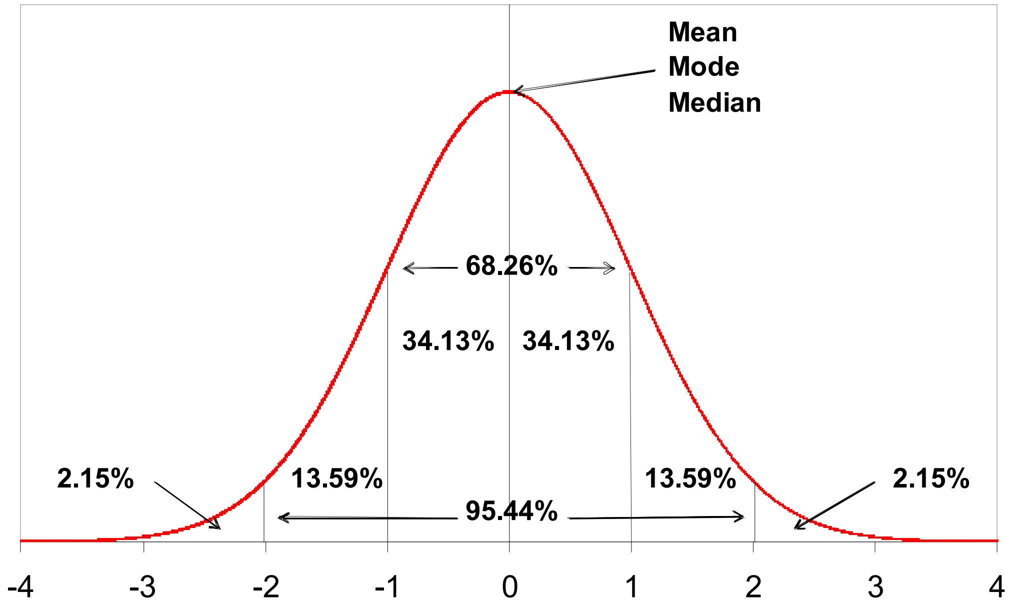


N = 5704
Mean = 61.9 kg
SD = 11.1 kg
Skewness = 1.30
Kurtosis = 2.64

Is Sum of 5 Skinfolds in Women Normally Distributed?

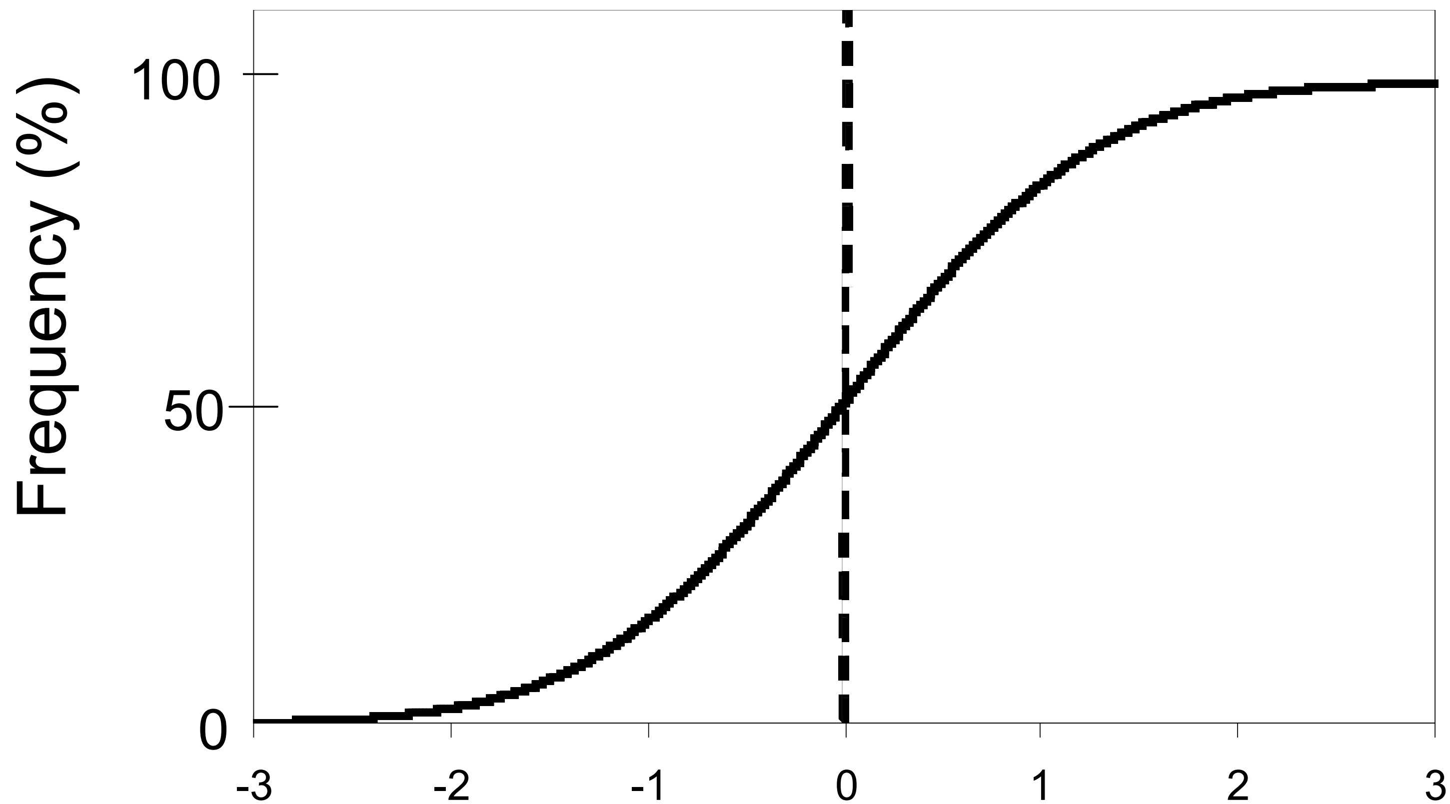


N = 5362
Mean = 75.8 mm
SD = 29.0 mm
Skewness = 1.04
Kurtosis = 1.30



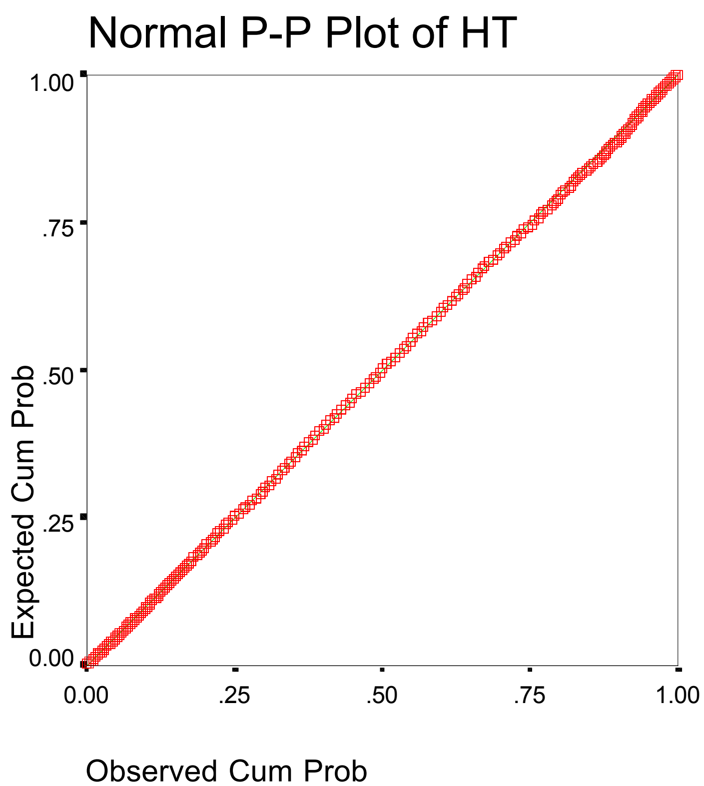
Normal Frequency Distribution

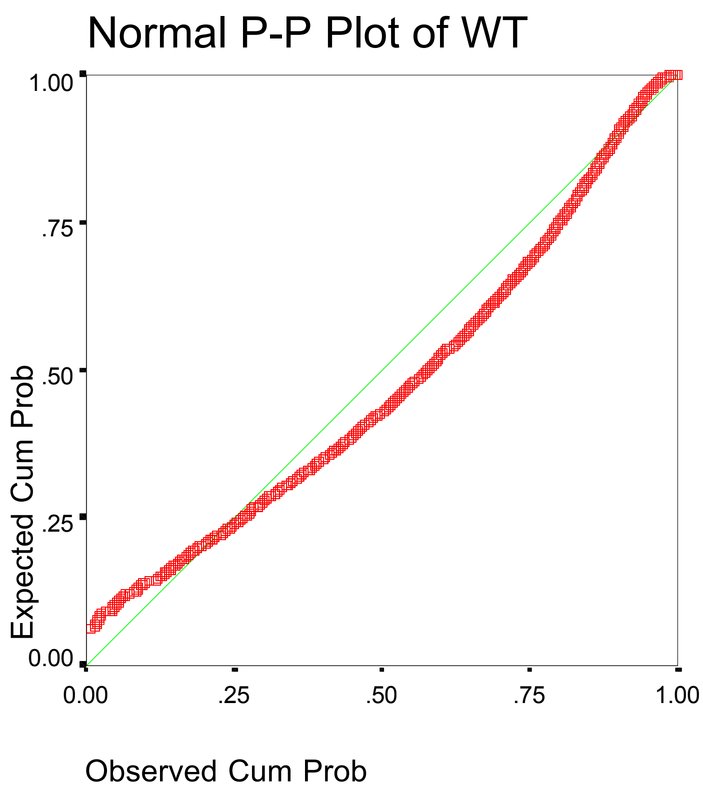
Cumulative Frequency Distribution (CFD)



Normal Probability Plots

Correlation between observed and expected cumulative probability is a measure of the deviation from normality.





Descriptive Statistics

- Measures of Central Tendency
  - Mean, Median, Mode
- Measures of Variability (Precision)
  - Variance, Standard Deviation, Interquartile Range
- Standardized scores (comparisons to the Normal distribution)
- Percentiles

Measures of Central Tendency

- Mean: "centre of gravity" of a distribution; the "weight" of the values above the mean exactly balance the "weight" of the values below it. Arithmetic average.
- Median (50th %tile): the value that divides the distribution into the lower and upper 50% of the values
- Mode: the value that occurs most frequently in the distribution

Measures of Central Tendency

- When do you use mean, median, or mode?
  - Height
  - Skinfolds
  - House prices in Vancouver
  - Vertical jump
  - 100 meter run time
- Study design: how many repeat measurements do you take on individuals to determine their true (criterion) score?
    - Discipline specific
    - Research design specific
    - Objective vs. subjective tests

Measures of Variability

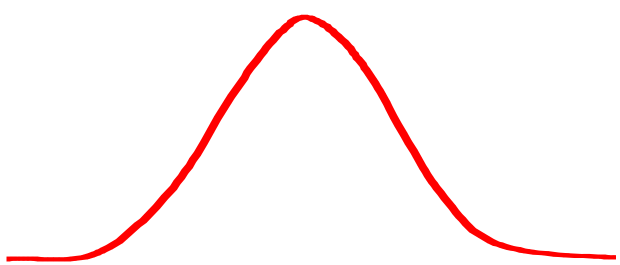
- Variance $$Var = \frac{\sum\left(X - \overline{X}\right)^2}{\left(N - 1\right)}$$
- Standard Deviation (SD) = Variance$^{1/2}$
- Range is approximately = ±3 SDs

➢ For Normal distributions, report the mean and SD

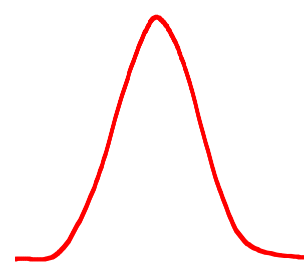
➢ For non-Normal distributions, report the median (50th %tile) and interquartile range (IQR, 25th and 75th %tiles)

Central Limit Theorem

- If a sufficiently large number of random samples of the same size were drawn from an infinitely large population, and the mean was computed for each sample, the distribution formed by these averages would be normal.



Distribution of a single sample



Distribution of multiple sample means. Standard deviation of sample means is called the standard error of the mean (SEM).

Standard Error of the Mean (SEM)

$$SEM = \frac{SD}{\sqrt{n}}$$

- The SEM describes how confident you are that the mean of the sample is the mean of the population

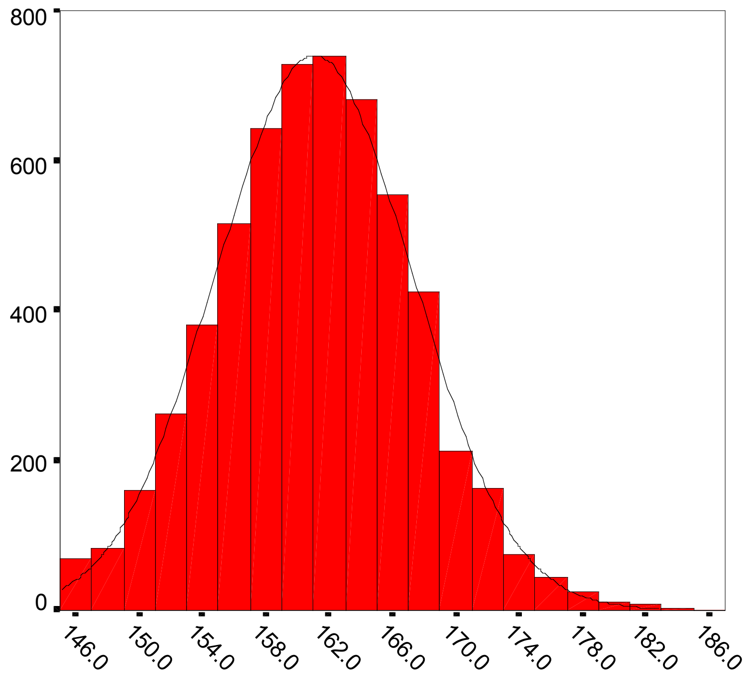
- How does the SEM change as the size of your sample increases?

Standardizing Data

Transform data into standard scores (e.g., Z-scores)

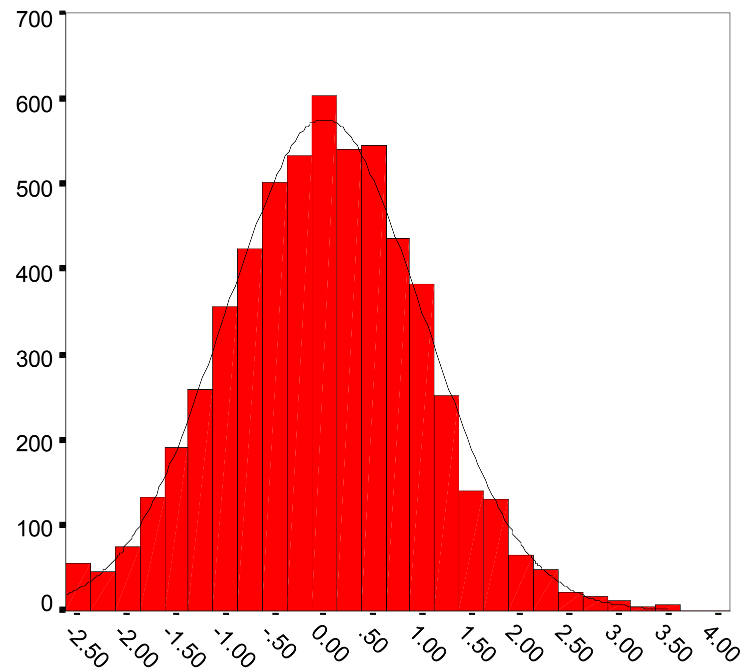
Eliminates units of measurements

Height (cm)



Mean=161.0; SD=6.2; N=5782

Z-Score of Height

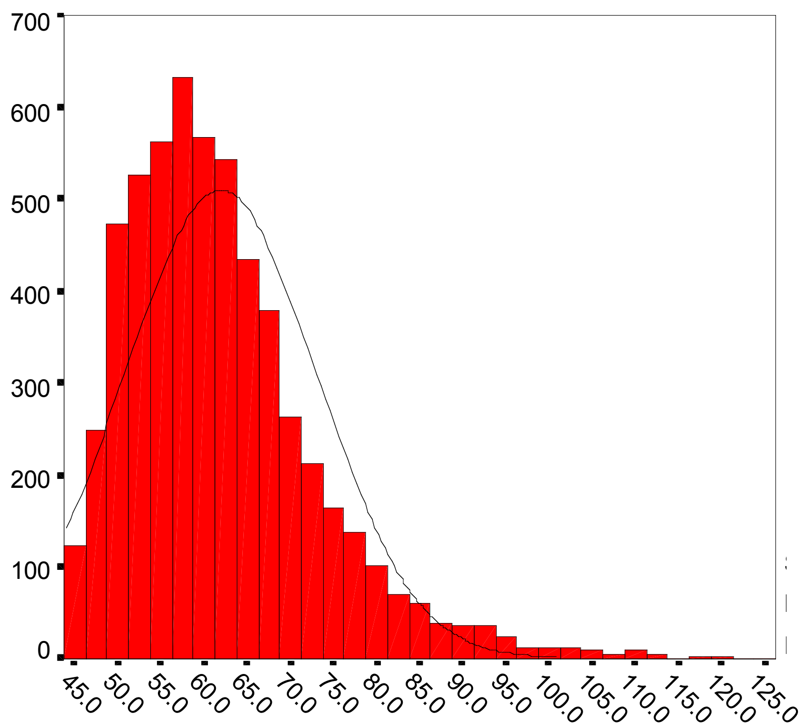


Mean=0.0; SD=1.0; N=5782

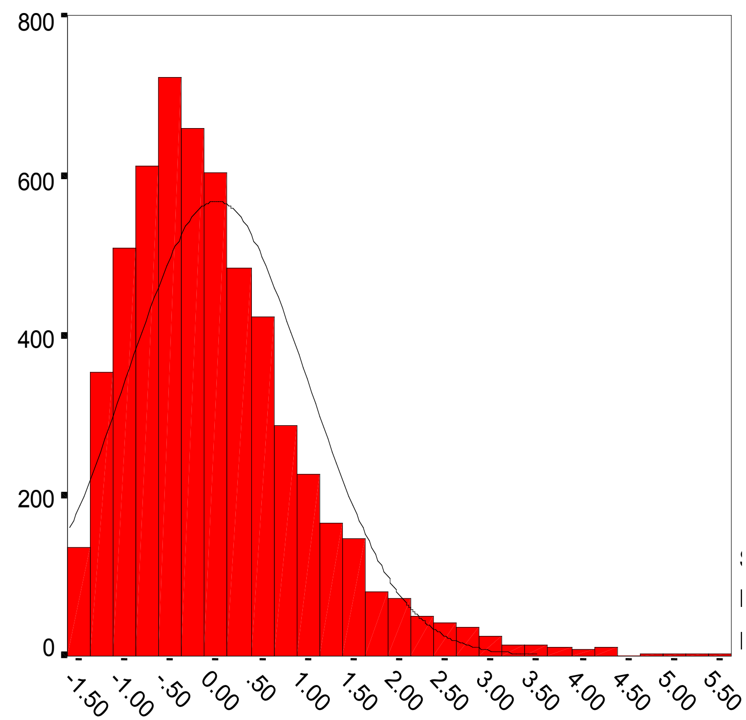
Standardizing Data

Standardizing does **not** change the distribution of the data

Weight (kg)



Z-Score of Weight



Z- Scores

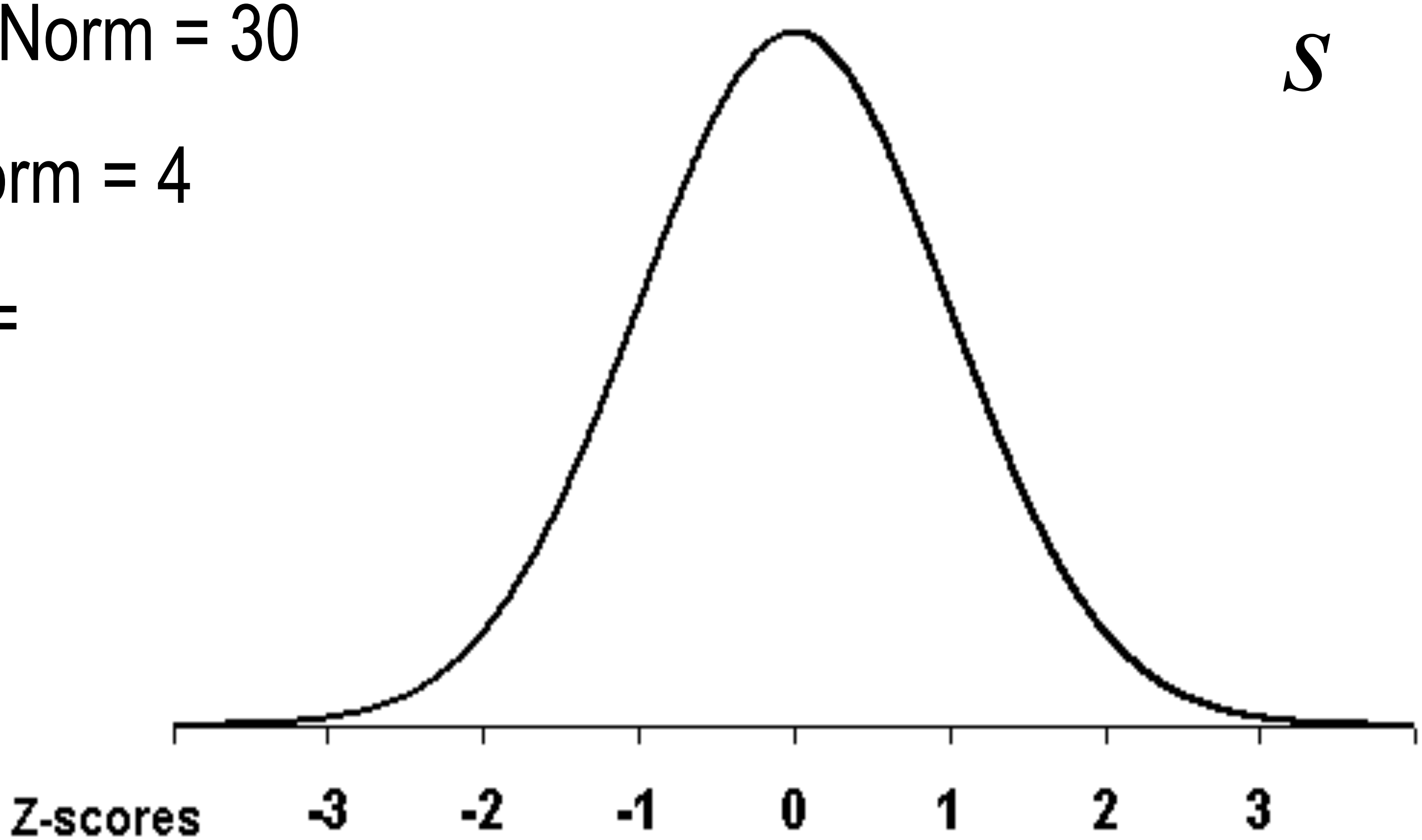
Score = 24

Mean of Norm = 30

SD of Norm = 4

Z-score =

$$Z = \frac{(X - \bar{X})}{s}$$



Internal or External Norm

Internal Norm

A sample of subjects are measured. Z-scores are calculated based upon the mean and SD of the sample. Thus, Z-scores using an internal norm tell you how good each individual is compared to the group they come from. Mean = 0, SD = 1

External Norm

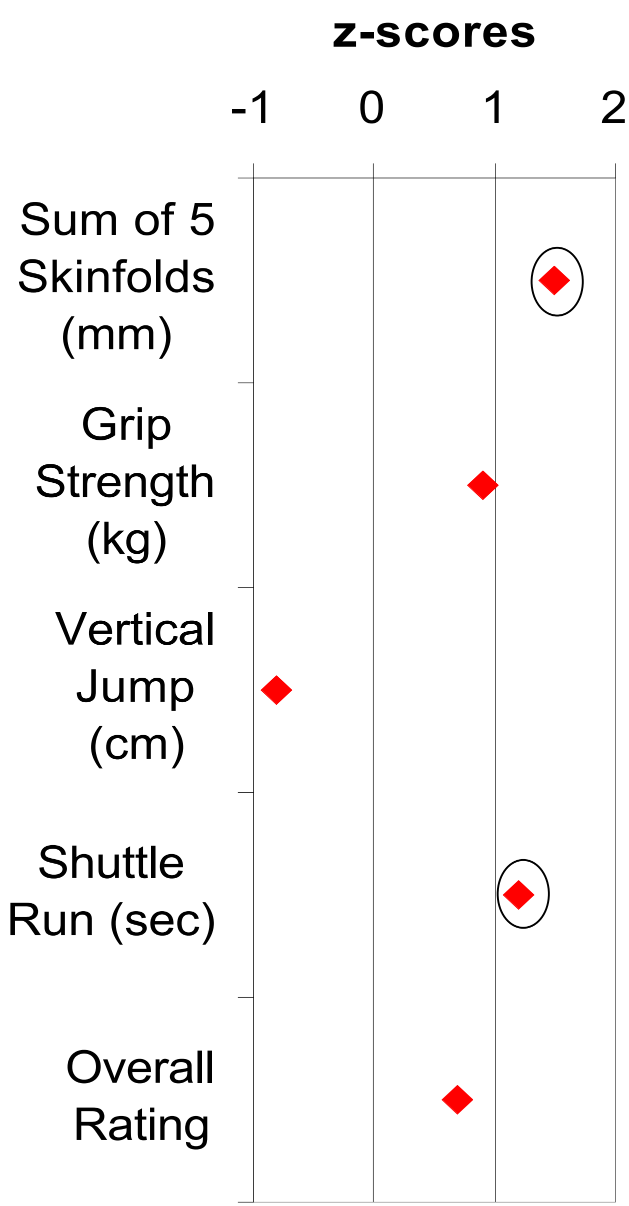
A sample of subjects are measured. Z-scores are calculated based upon mean and SD of an external normative sample (national, sport-specific etc.). Thus, Z-scores using an external norm tell you how good each individual is compared to the external group. Mean = ?, SD = ? (depends upon the external norm)

E.g. You compare aerobic capacity to an external norm and get a lot of negative z-scores? What does that mean?

Z-scores allow measurements from tests with different units to be combined. But beware: higher Z-scores are not necessarily better performances.

Variable	z-scores for profile A	z-scores for profile B
Sum of 5 Skinfolds (mm)	**1.5**	**-1.5***
Grip Strength (kg)	**0.9**	**0.9**
Vertical Jump (cm)	**-0.8**	**-0.8**
Shuttle Run (sec)	**1.2**	**-1.2***
Overall Rating	**0.7**	**-0.65**

*Z-scores are reversed because lower skinfold and shuttle run scores are regarded as better performances



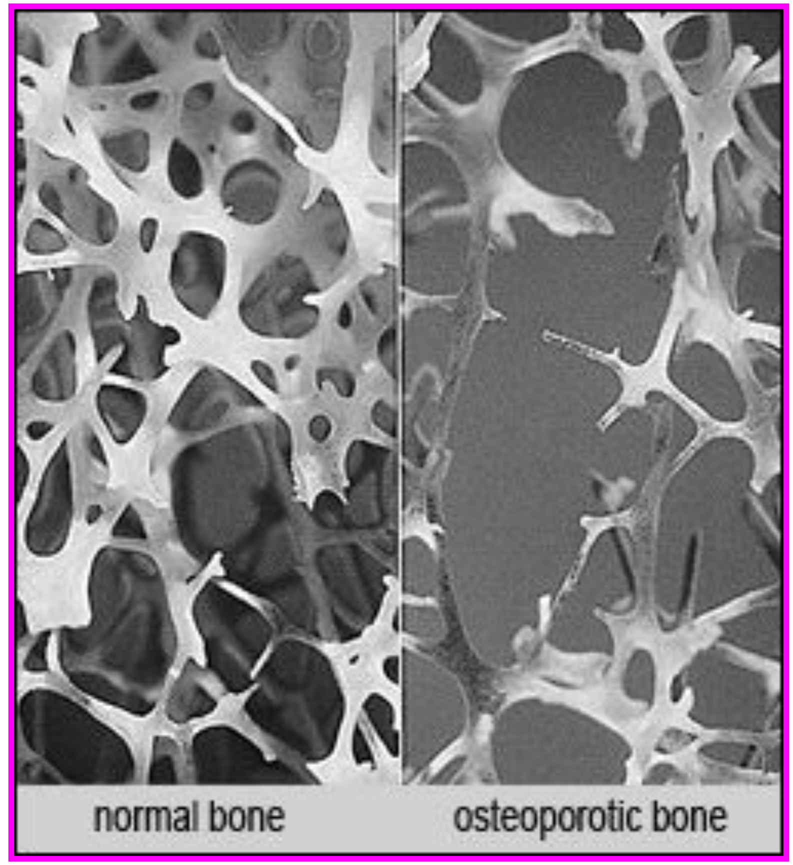
Test Profile A

Test Profile B

Clinical Example: T-scores and Osteoporosis

- To diagnose osteoporosis, clinicians measure a patient's bone mineral density (BMD)

- They express the patient's BMD in terms of standard deviations above or below the mean BMD for a "young normal" person of the same sex and ethnicity



BMD T-scores and Osteoporosis

$$T - score = \frac{\left(BMD_{\text{patient}} - \overline{BMD}_{\text{young normal}}\right)}{SD_{\text{young normal}}}$$

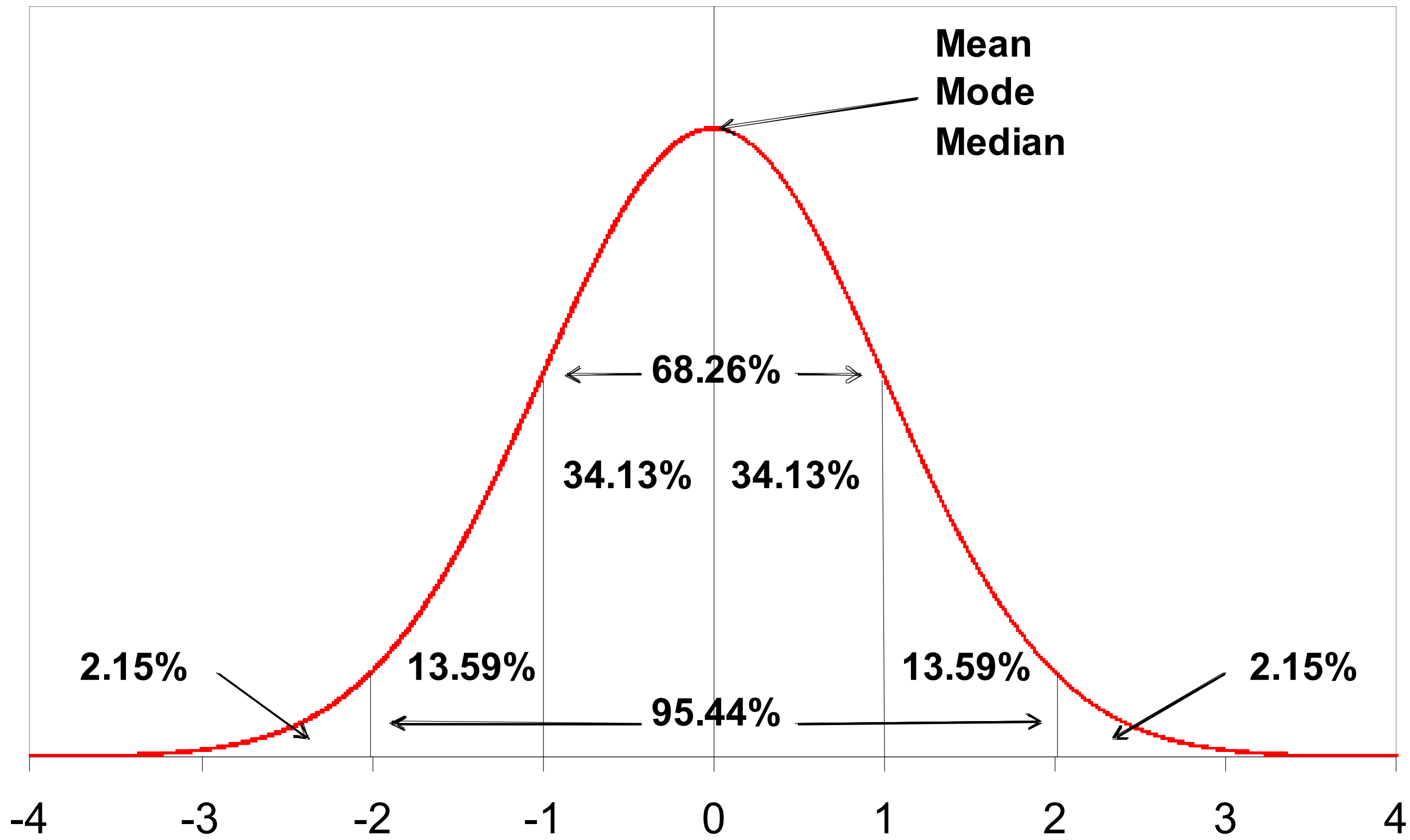
Although physician's call this standardized score a T-score, it is really just a Z-score where the reference mean and standard deviation come from an external population (i.e., young normal adults of a given sex and ethnicity).

Classification using BMD T-scores

- Osteoporosis T-scores are used to classify a patient's BMD into one of three categories:
  - T-scores of ≥ -1.0 indicate **normal** bone density
  - T-scores between -1.0 and -2.5 indicate **low bone mass** ("osteopenia")
  - T-scores ≤-2.5 indicate **osteoporosis**
- Decisions to treat patients with osteoporosis medication are based, in part, on T-scores.
- http://www.nof.org/sites/default/files/pdfs/NOF_ClinicianGuide2009_v7.pdf

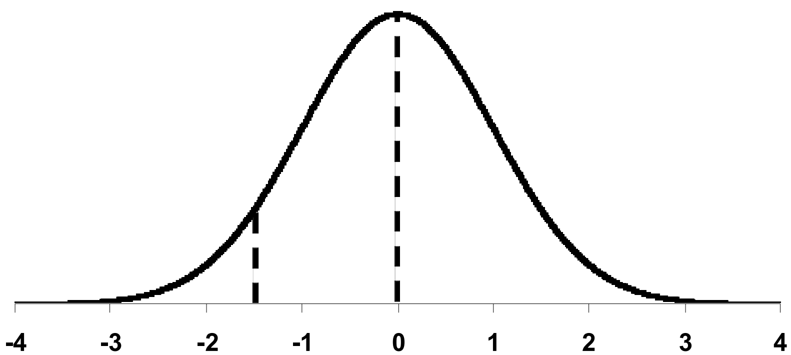
Percentiles

Percentile: The percentage of the population that lies at or below that score



Area under the Standard Normal Curve

What percentage of the population is above or below a given z-score or between two given z-scores?



Percentage between 0 and -1.5

43.32%

Percentage above -1.5

50 + 43.32% = 93.32%

TABLE A 3

CUMULATIVE NORMAL FREQUENCY DISTRIBUTION

(Area under the standard normal curve from 0 to Z)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000									

Predicting Percentiles from Mean and SD assuming a Normal Distribution

Percentile	Z-score for Percentile	Predicted Percentile value based upon Mean = 170, SD = 10
5	-1.645	153.55
25	-0.675	163.25
50	0	170
75	+0.675	176.75
95	+1.645	186.45

Predicted percentile value = Mean + (Z-score x SD)