Module 5B : A Parametric Test

Z-scores & RNAseq analysis

Agenda:

1. Z-scores
2. RNASeq

The Normal Distribution

What are the best values to describe a normal distribution?

a) Median and variance because they are not influenced by outliers

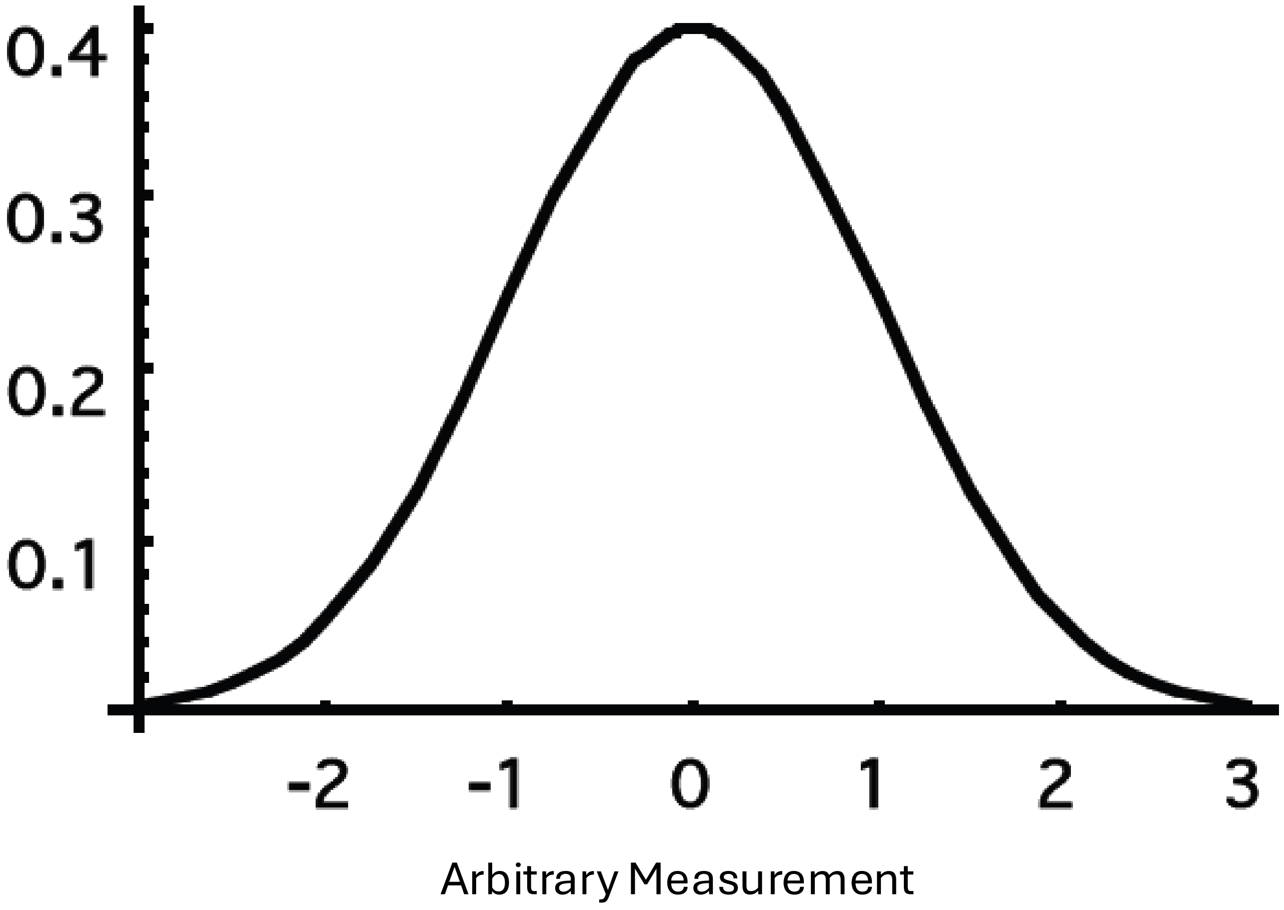
b) Mean and standard deviation when the data is not skewed by outliers

c) Only mean, because the standard deviations are all the same with normal distributions

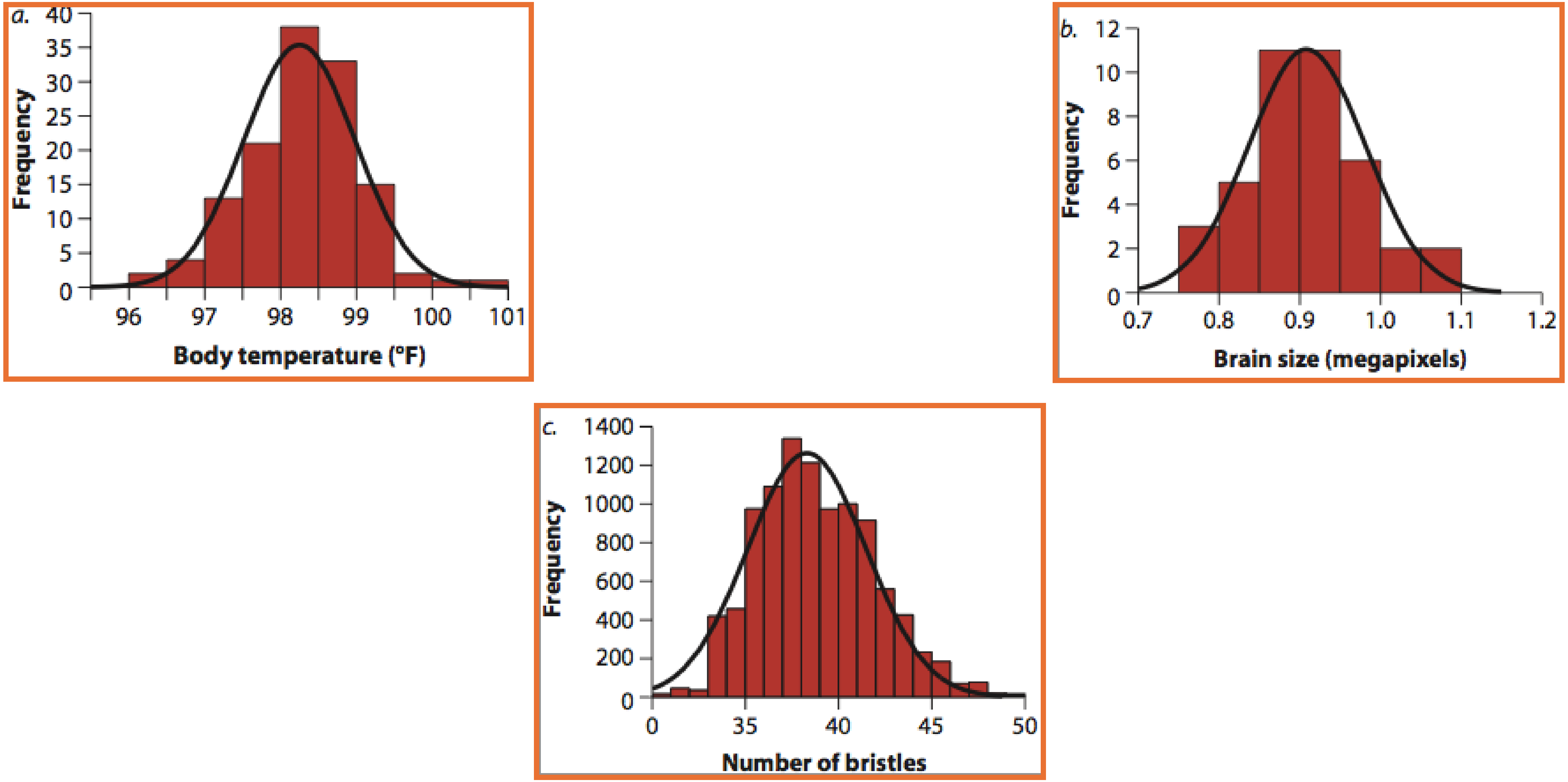
d) Only mode, because that is where the densest part of the curve lies

The Normal Distribution:

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

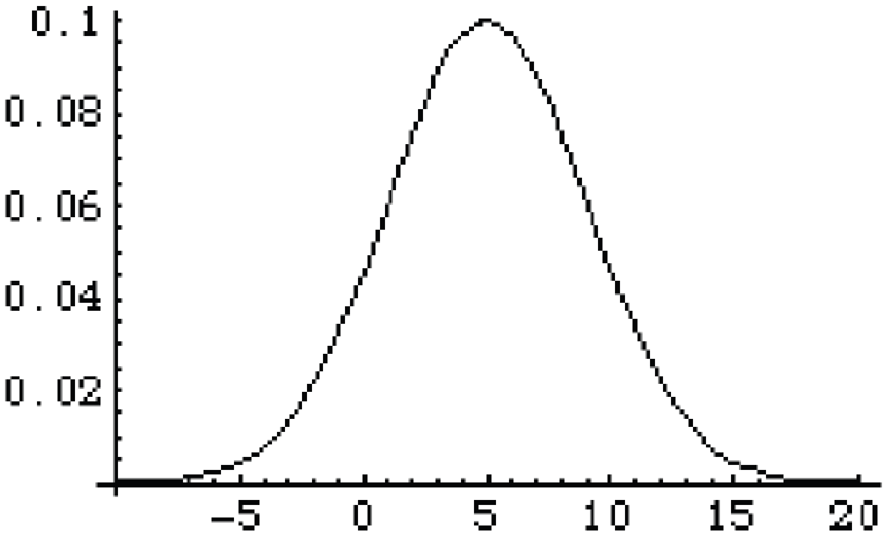


The Normal Distribution is common in nature:

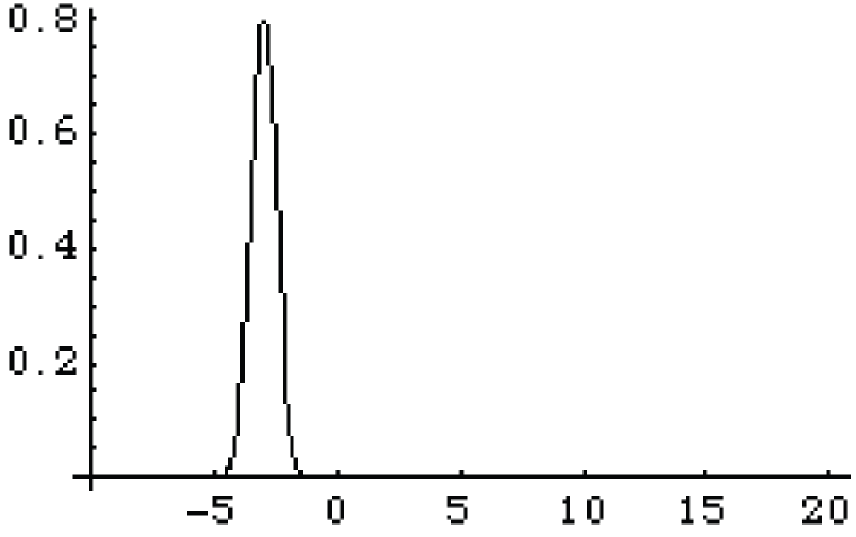


Properties of the Normal Distribution:

1. Fully described by its mean and standard deviation



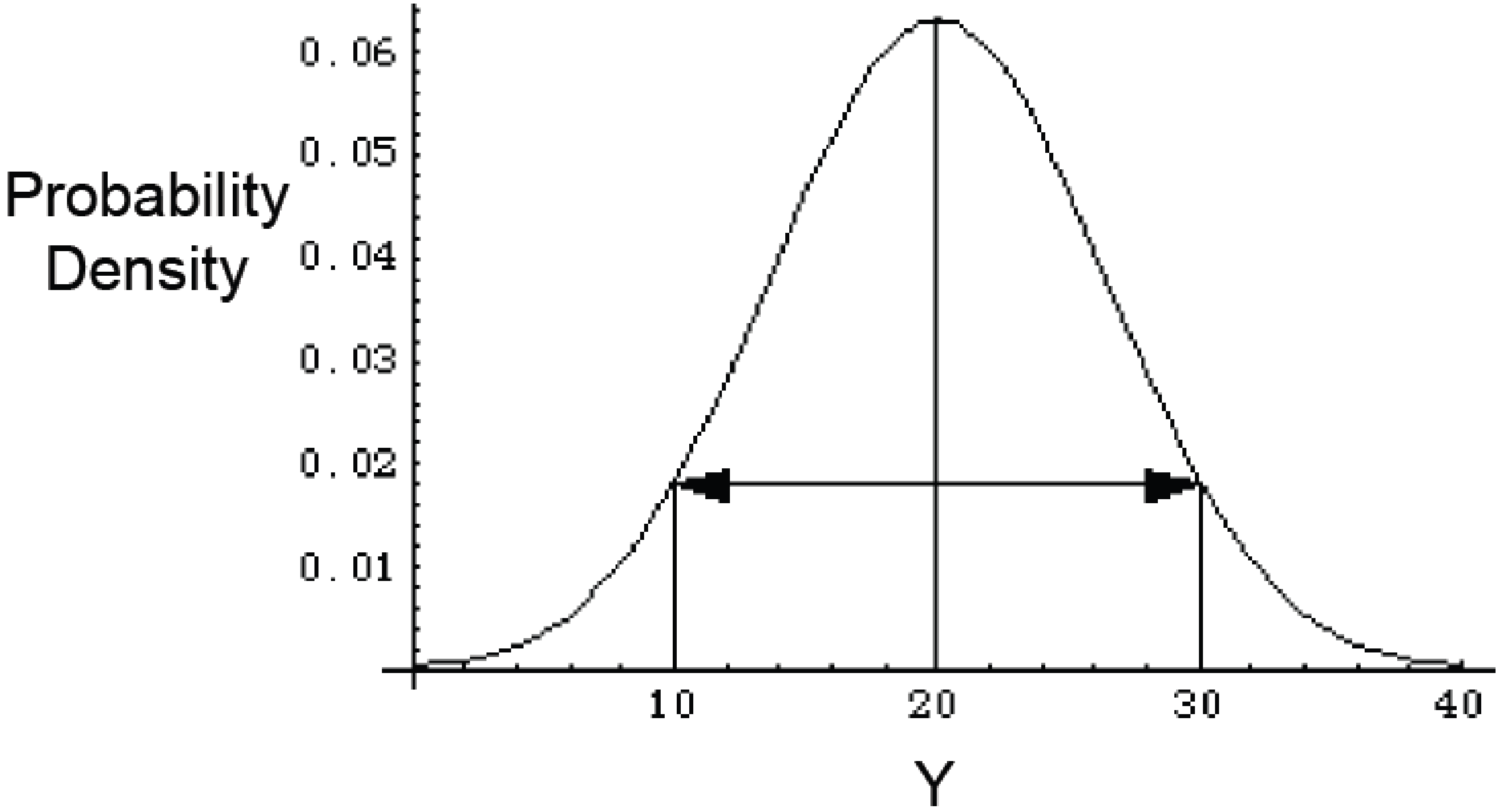
$\mu = 5; \sigma = 4$



$\mu = -3; \sigma = 1/2$

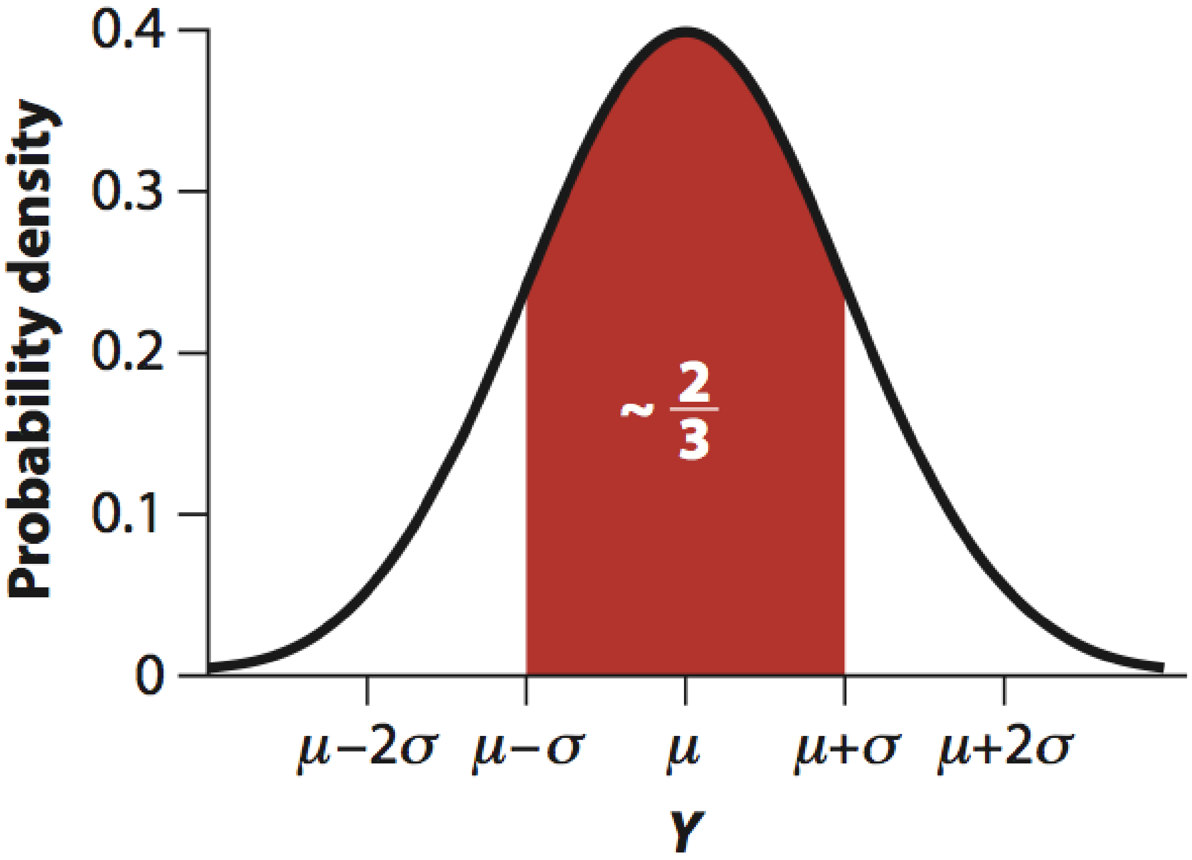
Properties of the Normal Distribution:

1. Fully described by its mean and standard deviation
2. Symmetric around its mean



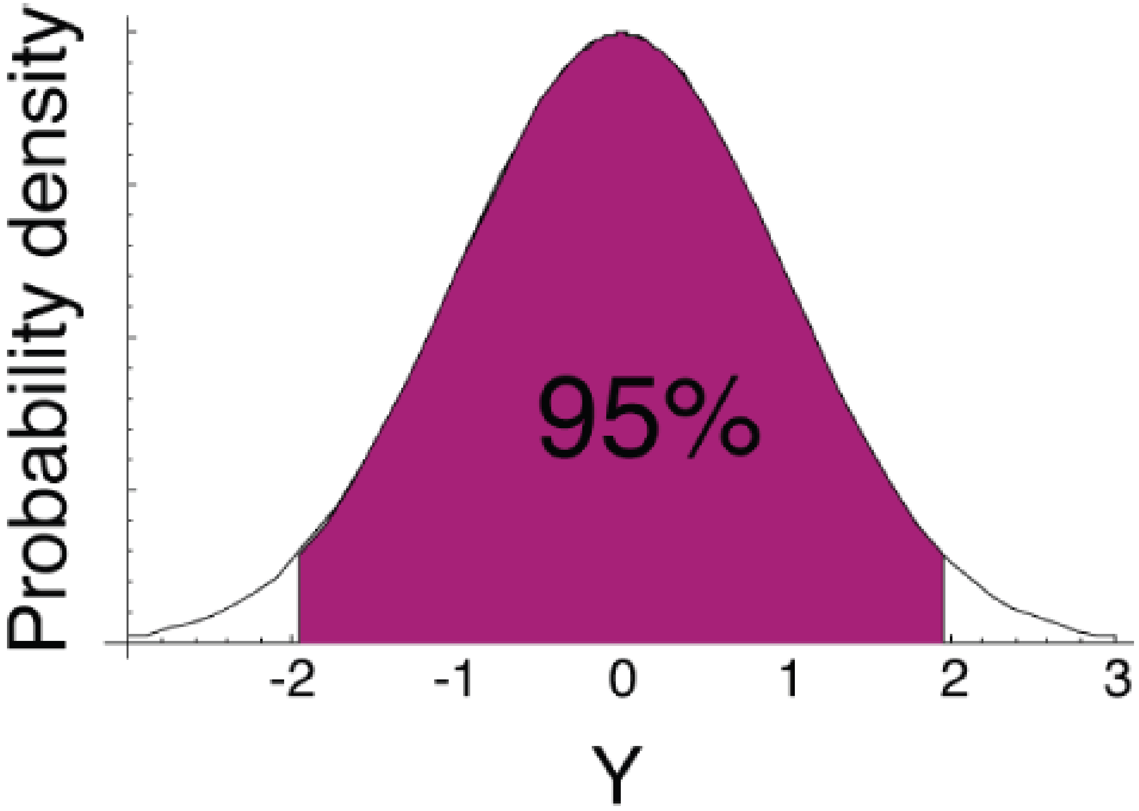
Properties of the Normal Distribution:

1. Fully described by its mean and standard deviation
2. Symmetric around its mean
3. ~ 2/3 of random draws are within one standard deviation of the mean



Properties of the Normal Distribution:

1. Fully described by its mean and standard deviation
2. Symmetric around its mean
3. ~ 2/3 of random draws are within one standard deviation of the mean
4. ~ 95% of random draws are within two standard deviations of the mean (*really, it is 1.96 SD*)



Which is the following is NOT a property of the normal distribution?

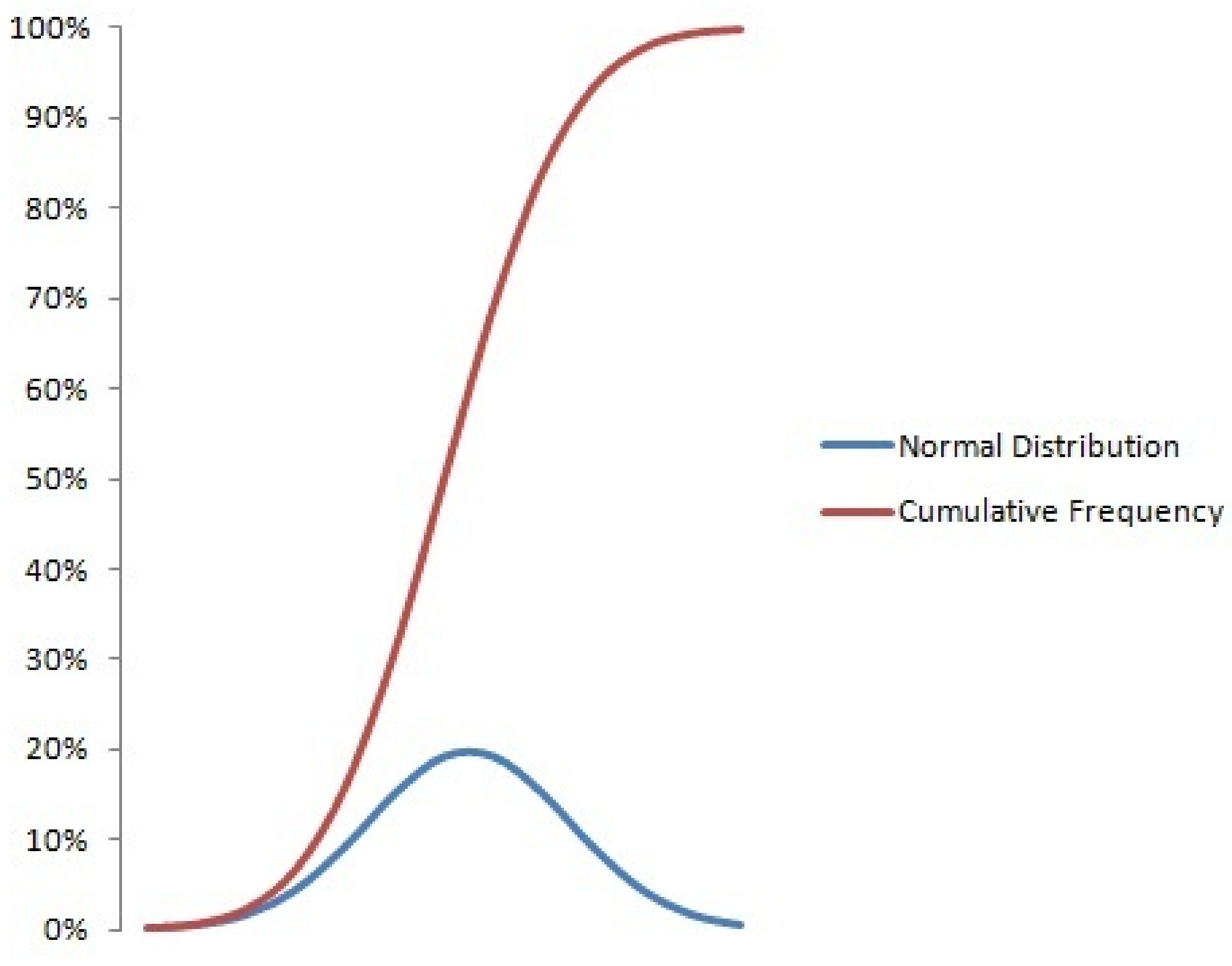
A. The probability density is highest exactly at the mean

B. The mean, mode and median are all equal

C. The normal curve is symmetrical about the mean μ

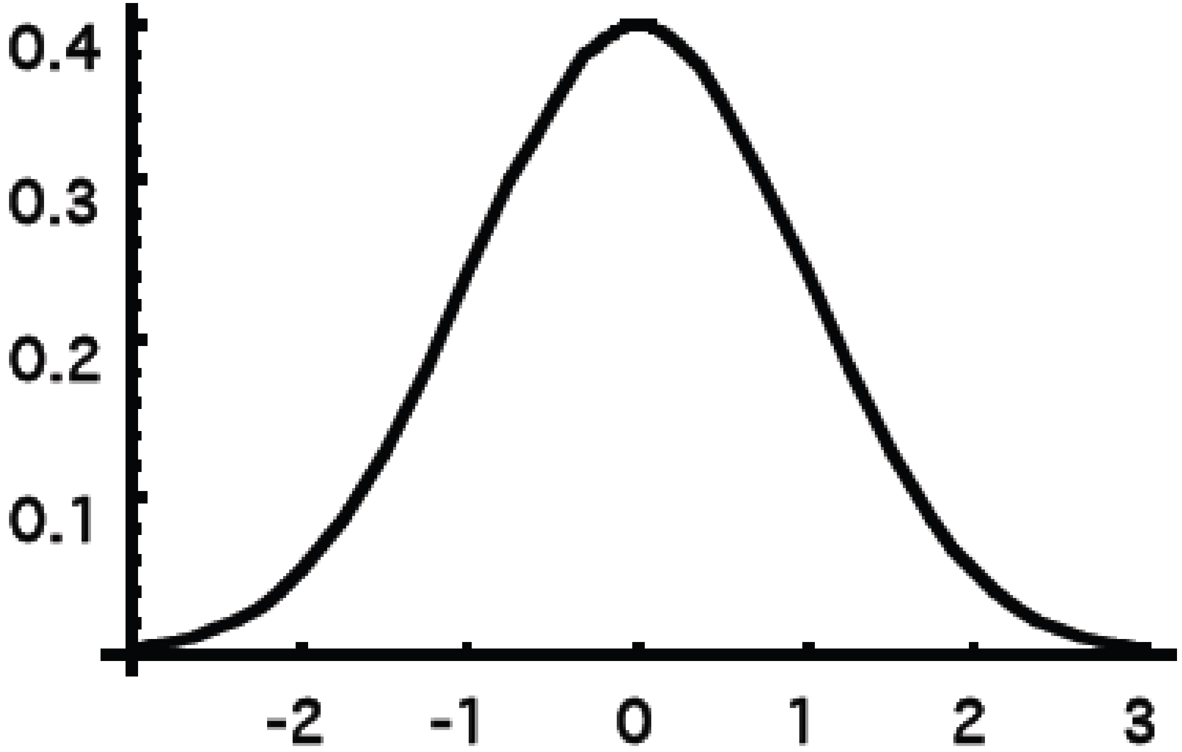
D. The probability that a random data point is within two standard deviation of the mean is approximately 68%

Properties of the Normal Distribution:



The Standard Normal Distribution:

- Mean is zero ($\mu = 0$)
- Standard deviation is 1 ($\sigma = 1$)



Z-scores:

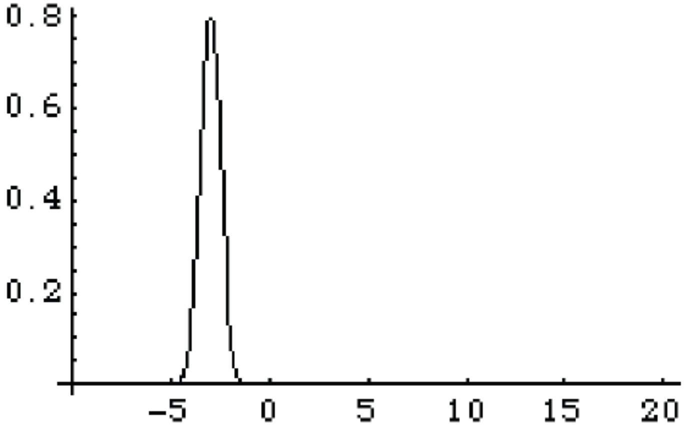
- converts **raw** normally distributed scores into standard deviation units
  - useful for comparing distributions with different scales, for instance.
  - percentiles

- allows calculation of probability of variable value

- z-score indicates how far above or below the mean a value is in standard deviation units
  - how large/small is individual score **relative** to others in the distribution

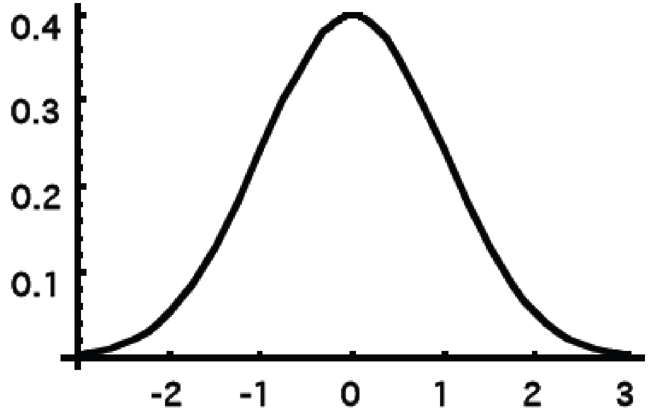
The Standard Normal Distribution:

- Mean is zero ($\mu = 0$)
- Standard deviation is 1 ($\sigma = 1$)



$\mu = -3$; $\sigma = 1/2$

$$\boxed{Z = \frac{X_i - \mu}{\sigma}}$$

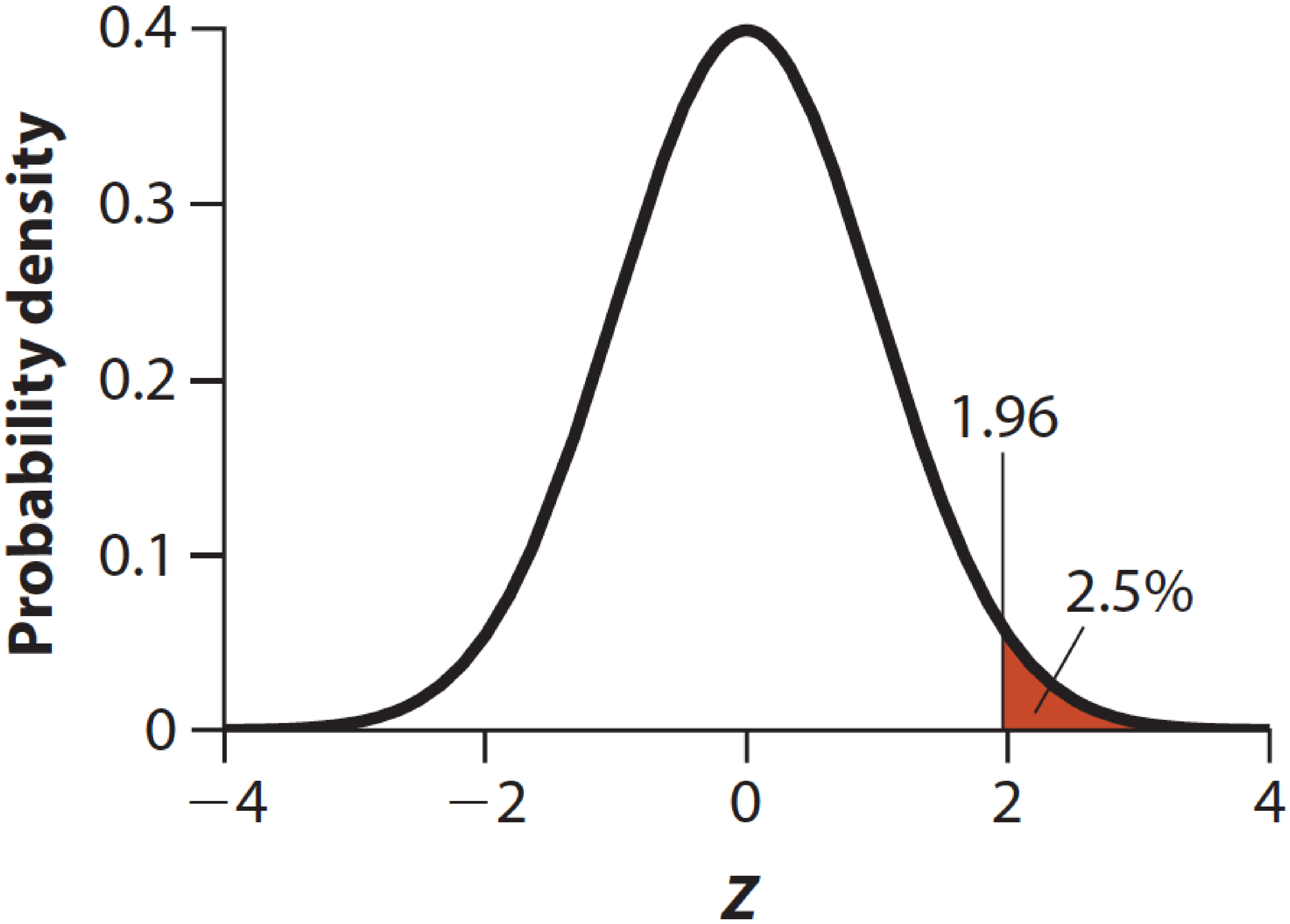


Interpret the following statements:

- Student A gets a *z-score* of -1.5 on an exam

- Student B received a *z-score* of 0.29 on the exam

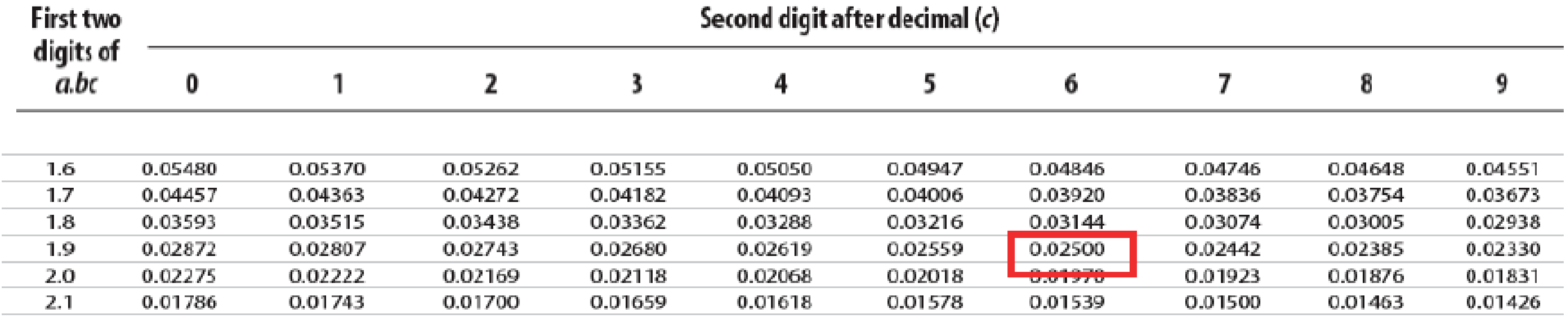
- Does the *z –score* tell you sample size?
- What the mean score on the test was?
- The percentage of answers Student B got right?



The probability of getting a random draw from a standard normal distribution greater than a given value which is the area under the curve.

Mechanics of Z tables:

The table works for P[Z>a.bc]



First two digits of *a.bc*	Second digit after decimal (*c*)									
	0	1	2	3	4	5	6	7	8	9
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426

For Z = 1.96 → P[Z>1.96]=0.025

https://www.z-table.com/

Since the standard normal is symmetric:

$$P[Z > x] = P[Z < -x]$$

$$P[Z < x] = 1 - P[Z > x]$$

remember: instead of α, you have $\alpha/2$ at each tail

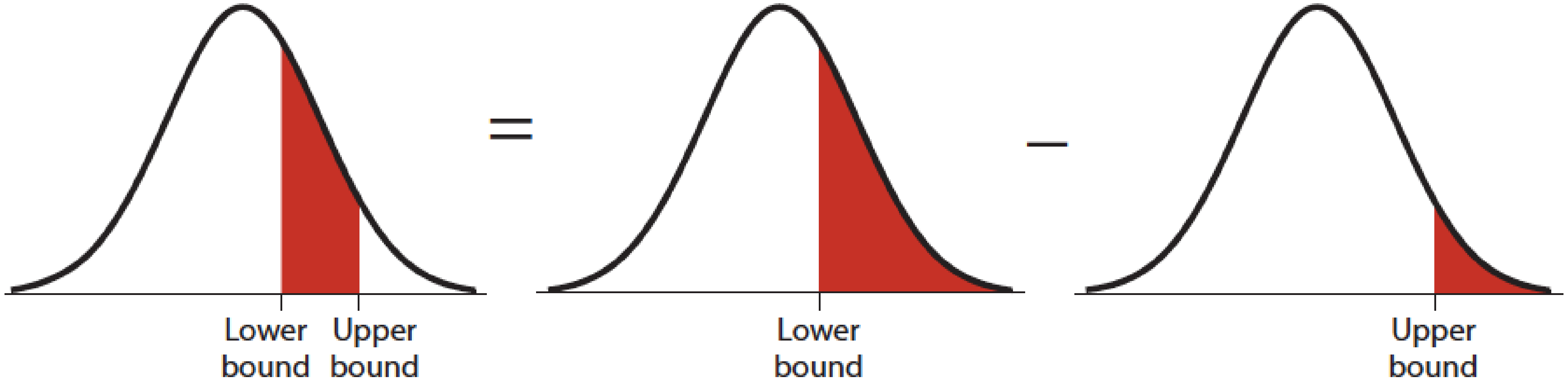
Example:

$$P[Z < -1.96] = P[Z > 1.96]$$

Example:

P*[lower bound < Z < Upper bound]* = P[Z > *lower bound*] - P[Z > *Upper bound*]

$$P(a < Z < b) = F\left(\frac{b-\mu}{\sigma}\right) - F\left(\frac{a-\mu}{\sigma}\right)$$



Sara and Jerry took a math exam. Sara's percentile score on the exam was 85; Jerry's percentile score on the same test was 70. We know that

A. Sara scored better than 85 of her classmates.

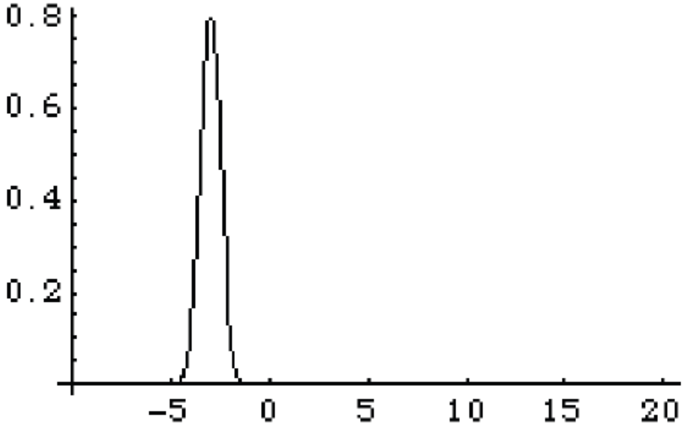
B. Sara correctly answered 15 more questions Jerry did.

C. They both scored better than average on the math exam.

D. Sarah correctly answered more items than Jerry did.

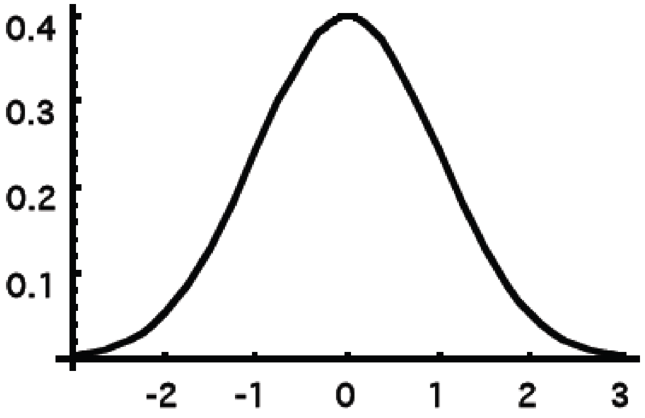
The Standard Normal Distribution:

- Mean is zero (μ = 0)
- Standard deviation is 1 (σ = 1)



μ = -3; σ = 1/2

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$



3 major motivations:

- Z score tells us how many standard deviations our normally distributed variable is from the mean

$$Z = \frac{\text{Raw score - Mean}}{\text{Standard deviation}}$$

- Confidence interval: $(a,b) = \overline{x} \pm z_{\alpha/2}(\sigma / \sqrt{2})$

- Determine proportion of scores that fall between two raw scores

- Allows us to use standard normal table

Example: British Spies. MI5 says that a man must be shorter than 180.3 cm tall to be a spy. The mean height of British men is 177.0 cm, with standard deviation 7.1 cm, and with a normal distribution.

What proportion of British men are excluded from a career as a spy by this height criteria?

Example: British Spies. MI5 says that a man must be shorter than 180.3 cm tall to be a spy. The mean height of British men is 177.0 cm, with standard deviation 7.1 cm, and with a normal distribution.

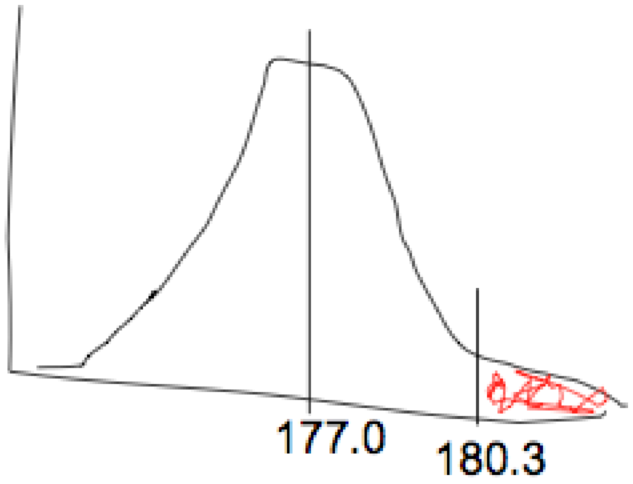
What proportion of British men are excluded from a career as a spy by this height criteria?

Step 1: Draw out question.

Example: British Spies. MI5 says that a man must be shorter than 180.3 cm tall to be a spy. The mean height of British men is 177.0 cm, with standard deviation 7.1 cm, and with a normal distribution.

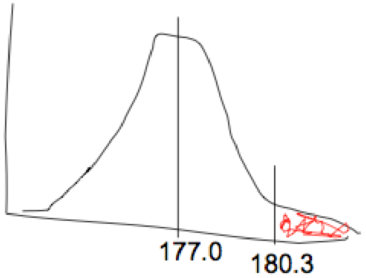
What proportion of British men are excluded from a career as a spy by this height criteria?

Step 1: Draw out question.



Example: British Spies. MI5 says that a man must be shorter than 180.3 cm tall to be a spy. The mean height of British men is 177.0 cm, with standard deviation 7.1 cm, and with a normal distribution. ***What proportion of British men are excluded from a career as a spy by this height criteria?***

Step 1: Draw out question.



Step 2: Transform into Standard Normal

$$\mu = 177.0cm$$
$$\sigma = 7.1cm$$
$$X = 180.3cm$$
$$P[height > 180.3]$$

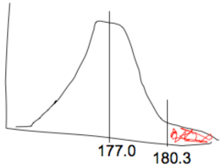
$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{180.3 - 177.0}{7.1}$$

$$Z = 0.46$$

Example: British Spies. MI5 says that a man must be shorter than 180.3 cm tall to be a spy. The mean height of British men is 177.0 cm, with standard deviation 7.1 cm, and with a normal distribution. ***What proportion of British men are excluded from a career as a spy by this height criteria?***

Step 1: Draw out question.



Step 2: Transform into Standard Normal

$$\boxed{\begin{array}{l}\mu = 177.0cm \\ \sigma = 7.1cm \\ X = 180.3cm \\ P[height > 180.3]\end{array}}$$

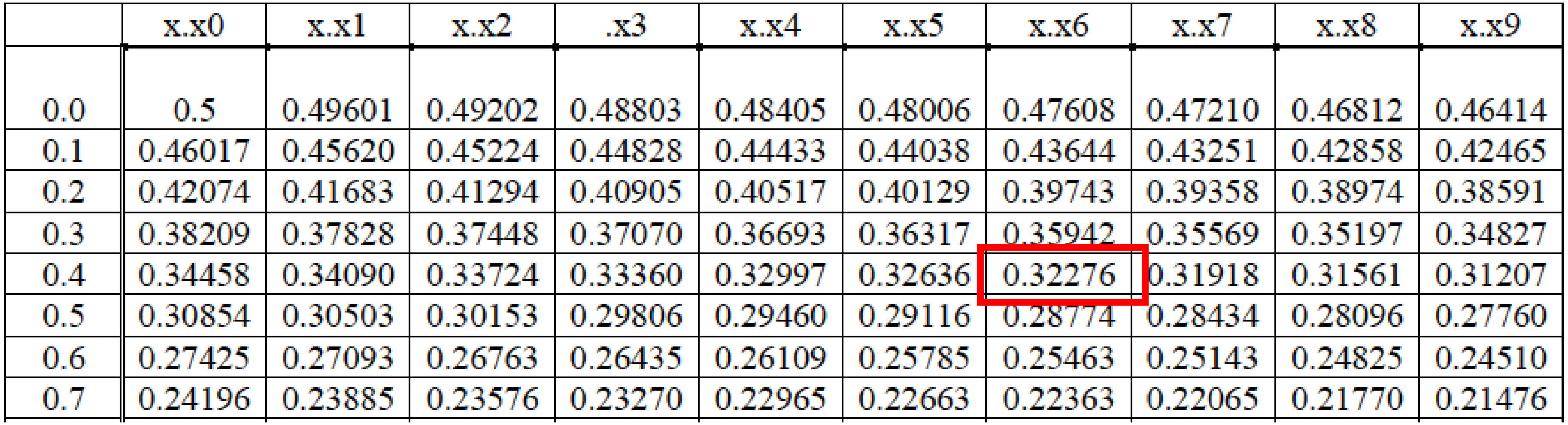
$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{180.3 - 177.0}{7.1}$$

$$Z = 0.46$$

Step 3: Look up probability online

https://www.z-table.com/

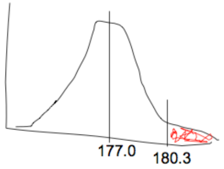


	x.x0	x.x1	x.x2	.x3	x.x4	x.x5	x.x6	x.x7	x.x8	x.x9
0.0	0.5	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476

Example: British Spies. MI5 says that a man has to be shorter than 180.3 cm tall to be a spy. The mean height of British men is 177.0 cm, with standard deviation 7.1 cm, and with a normal distribution. ***What proportion of British men are excluded from a career as a spy by this height criteria?***

Step 1: Draw out question.

Step 2: Transform into Standard Normal



$$\boxed{\begin{aligned}&\mu = 177.0cm\\&\sigma = 7.1cm\\&X = 180.3cm\\&P[height > 180.3]\end{aligned}}$$

$$Z = \frac{X-\mu}{\sigma}$$

$$Z = \frac{180.3 - 177.0}{7.1}$$

$$Z = 0.46$$

Step 3: Look up probability in Z table

P[Z > 0.46] = 0.32276

So, P[height > 180.3] = 0.32276

The fraction of British males who are too tall to be spies is approx. 1/3.

Example: For a particular year, the average SAT-math scores was 517 (out of 800) with a standard deviation of 100. What score marks the **90%** percentile?

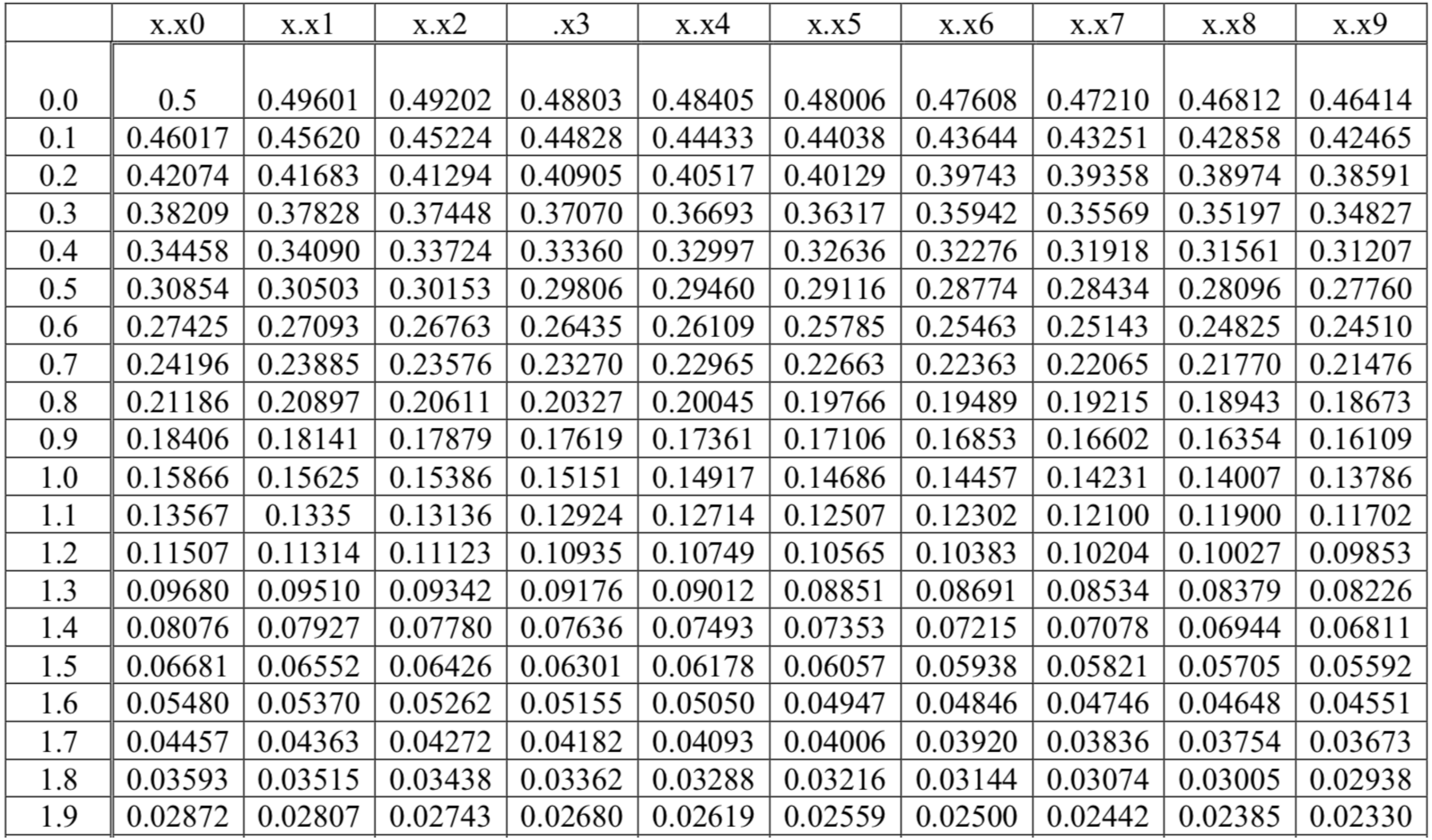
a. 681.5

b. 645.5

c. 527

d. 568.7

https://www.z-table.com/



	x.x0	x.x1	x.x2	.x3	x.x4	x.x5	x.x6	x.x7	x.x8	x.x9
0.0	0.5	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
1.1	0.13567	0.1335	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330

Example: What is your friend casually mentioned to you that they had scored a 750 while you had scored 425 on the math section of the SAT? What percentage of scores is between you?

Step 1: Convert raw score into *z-score*:

$$\boxed{Z = \frac{X_i - \mu}{\sigma}} \qquad \boxed{Z = \frac{425 - 517}{100} = -0.92} \qquad \boxed{Z = \frac{750 - 517}{100} = 2.33}$$

Step 2: Find the proportion of the normal distribution that falls below a score of:

1. -0.92
2. 2.33.

Sample means are normally distributed:

If a variable itself is normally distributed then the distribution of sample means, $\overline{Y}$, is also normally distributed

Sampling distribution for $\overline{Y}$:

The range of different values for $\overline{Y}$ *that could have been obtained by sampling, and their associated probabilities, constitute the sampling distribution for* $\overline{Y}$.

Sample means are normally distributed:

If a variable itself is normally distributed then the distribution of sample means, $\overline{Y}$, is also normally distributed

- The mean of the sample means is μ
- The standard deviation of the sampling distribution for $\overline{Y}$ is called the ***Standard error:***

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}} \xrightarrow{\text{approx}} SE_{\overline{Y}} = \frac{s}{\sqrt{n}}$$

Standard error **is the spread/variation statistic for a "collection" of means (or proportions)**.

It can be thought of as the theoretical variation present when you repeat a study many times.

This contrasts with standard deviation which is used to describe the natural variation of something that you can measure

Standard Normal applied to sampling means:

- Sample means distributed normally with mean equal to μ and standard error then:

$$Z = \frac{\overline{Y} - \mu}{\sigma_{\overline{Y}}}$$

- The means of samples taken from a normal distribution are themselves normally distributed but….

- ….the sampling distribution of sample means is *approximately normal* **even when the distribution of Y is not normal** *if the sample size is large enough (depends on the shape of the data)*

Central Limit Theorem:

The sum (or mean) of a large number of measurements randomly sampled from any population is approximately normally distributed.

Practice Problem: Singleton babies born have a mean weight of 3.339 kg and a standard deviation of 0.573 kg.

a. What is the probability a newborn weighs more than 5 kg?

b. What is the probability a newborn weighs between 3 kg and 4 kg?

c. What fraction of newborns will be > 1.5 sd from the mean?

d. What fraction of newborns will be > 1.5 kg from the mean?

e. A random sample of 10 newborns is taken, what is the probability that their mean weight would be > 3.5 kg?

Which of the follow statements about distributions is not true?

a. The normal distribution can be used as an approximation to the binomial distribution

b. The Poisson distribution is used under conditions where n approaches 0 while p approaches infinity

c. The binomial distribution expresses the probability of getting X successes out of n trials

d. As degrees of freedom increase, the X^2 distribution approaches a normal distribution

Normal Approximation to the Binomial Distribution*:

- Remember the binomial distribution?
  - discrete
  - number of successes in n independent trials n
- Number of successes is a sum
- mean = np
- stand dev = $\sqrt{np(1-p)}$

Standard Normal Approximation to the Binomial Distribution:

1. State H_0 and H_A
2. Test Statistic
3. P-Value or Critical value/Compare to critical value (remember to double it!)

$$\boxed{P[NumSuccesses \square X] = P[Z > \frac{X - np}{\sqrt{np(1-p)}}]}$$

4. State a conclusion

RNAseq:

https://pluto.bio/resources/learning-series/overview-of-z-scores-in-rna-seq-experiments

In RNA-seq analysis, Z-scores are used to compare expression levels between samples. The Z-score of a gene is calculated by comparing its expression level in a given sample to the expression level of that gene across all samples. A Z-score of zero indicates that the gene's expression level is the same as the mean expression level across all samples, while a positive Z-score indicates that the gene is expressed at a higher level than the mean, and a negative Z-score indicates that the gene is expressed at a lower level than the mean.

Once Z-scores are calculated, they can be used to identify differentially expressed genes. For example, genes with a Z-score greater than a certain threshold (such as 2 or 3) can be considered as differentially expressed. Additionally, Z-scores can be used to create a heatmap or volcano plot, which can be a valuable way to visualize the data and identify patterns of expression

Simple example of using z-scores for RNA-seq data:

Gene	Control 1	Control 2	Treat 1	Treat 2
G1	5	6	9	10
G2	200	210	220	230
G3	1.2	1.1	1.3	1.4

This data set has two controls, two treatments for three genes. Each number is log-scaled or normalized expression value.

For each gene, ask:

“How far above or below that gene’s own average is each sample?”

$$z = \frac{x - \overline{x}}{s}$$

x = value for one sample
$\bar{x}$ = *mean for that gene*
s = *Standard deviation for that gene*

Simple example of using z-scores for RNA-seq data:

Gene	Control 1	Control 2	Treat 1	Treat 2
G1	5	6	9	10
G2	200	210	220	230
G3	1.2	1.1	1.3	1.4

$$z = \frac{x - \overline{x}}{s}$$

For Gene 1:

Sample	Value	Mean ($\overline{x}$)	Std Dev (s)	z-score
Control 1	5	7.5	≈ 2.38	(5 – 7.5)/2.38 = -1.05
Control 2	6	7.5	≈ 2.38	-0.63
Treat 1	9	7.5	≈ 2.38	0.63
Treat 2	10	7.5	≈ 2.38	1.05

-0.84

+0.84

Controls have negative z's (below average); treatments have positive z's (above average) → up-regulated with treatment. Typically, you would have many examples of treatment and control at a gene and then you would take their mean (we only have two from each category)