Continuous Random Variables and the Normal Distribution

ANA 500 – Foundations of Data Analytics

McDaniel College

Random variables

- A **continuous random variable** can take on any value in an interval

- The probability of any single value is zero

- A **discrete variable** can take on only certain values along an interval

Random variables

- **Probability density function (PDF)**
  - The probability of a random variable takes on a value between a and b
    - Equal to the area under the PDF between a and b

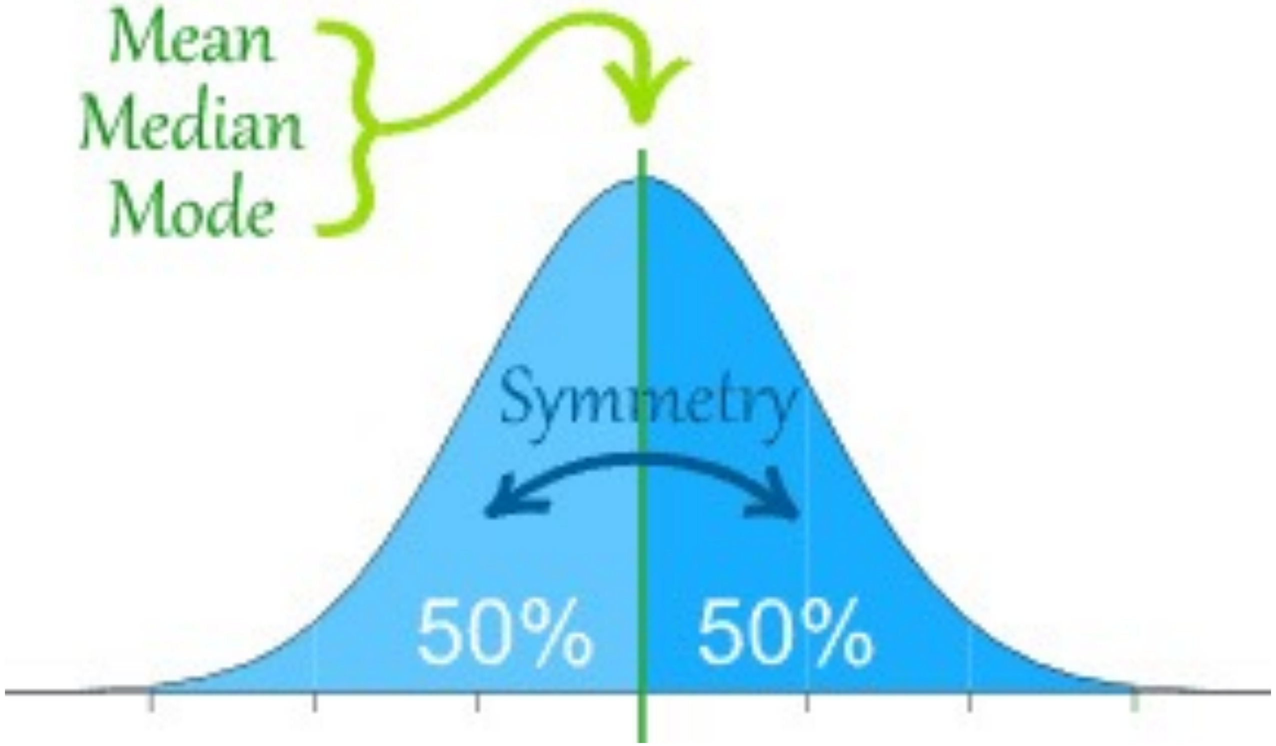
- **Cumulative distribution function (CDF)**
  - The probability a random variable will take on a value equal to or less than some value.
    - Equal to the area under the CDF to the left of the value of interest.

Uniform distribution (continuous)

- All outcomes are equally likely
- $X \sim U(a, b)$
- $f(x) = \frac{1}{b-a}$ for $x \in (a, b)$ $\qquad f(x) = 0 \;\; otherwise$
- Mean = $\frac{1}{2}(\mathrm{a} + \mathrm{b})$
- Variance = $\frac{1}{12}(\mathrm{b} - \mathrm{a})^2$

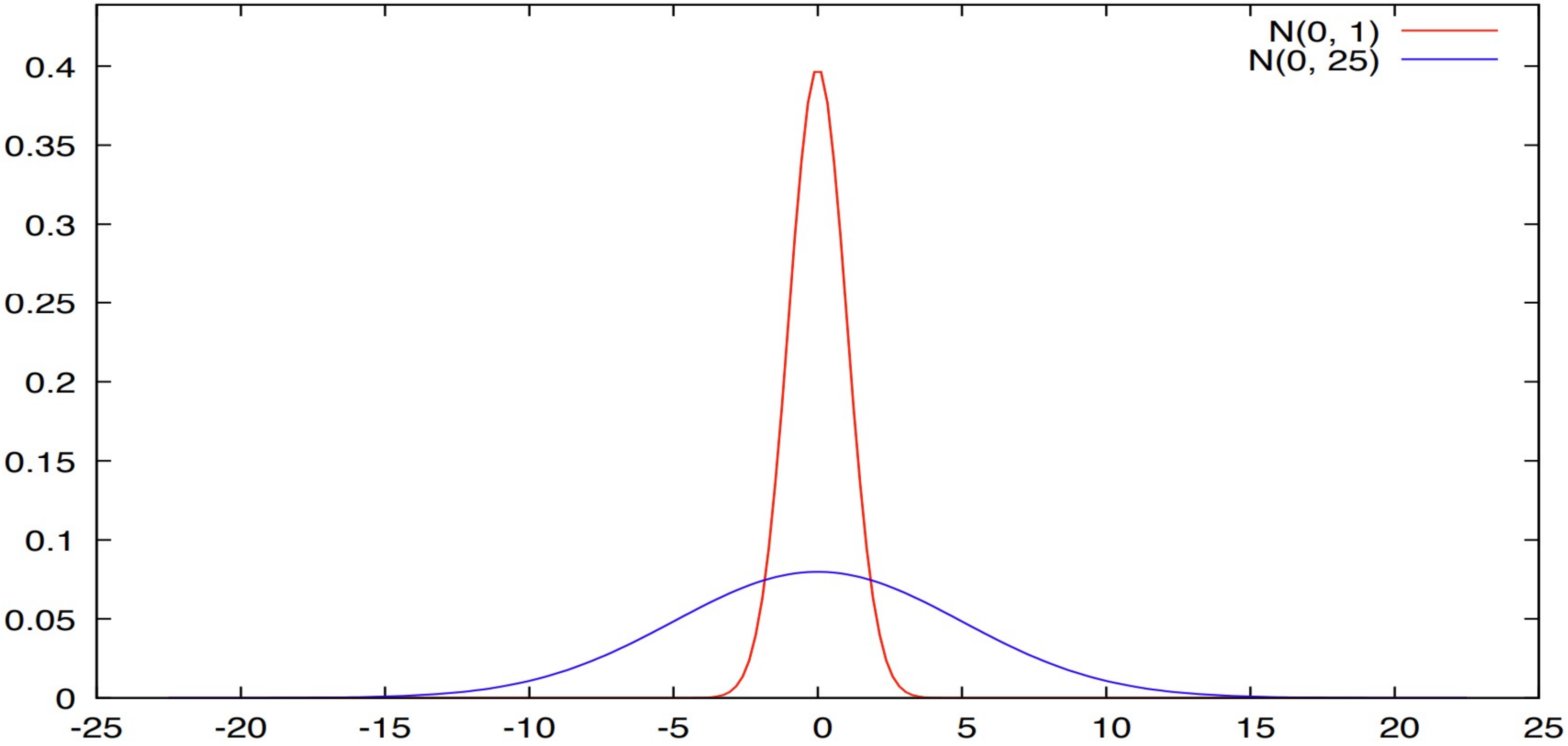
Normal distribution

- By far, the most well-known and widely used probability distribution
- Symmetric
- Mean = median = mode
- $X \sim N(\mu, \sigma^2)$
- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{1}{2}\right)\left(\frac{x-\mu}{\sigma}\right)^2}$
- Mean$= \mu$
- Variance = σ^2



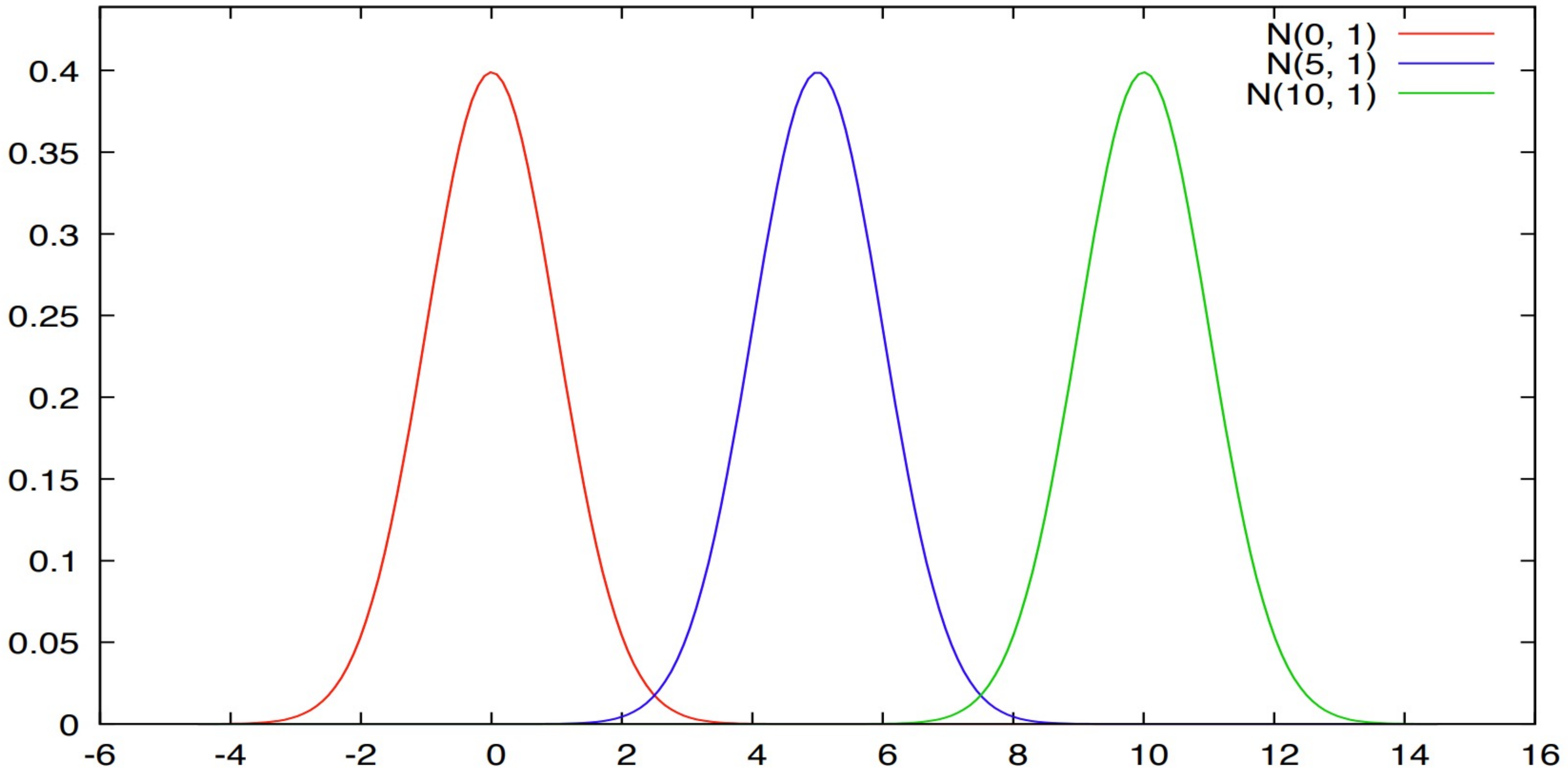
Normal distribution

- Both distributions have a mean of zero
- The values in the blue distribution are far more spread around the mean of zero than the values in the red normal distribution.
- Both distributions are symmetric, and both are centered over their common mean.



Normal distribution

- Not all normal distributions are exactly the same.



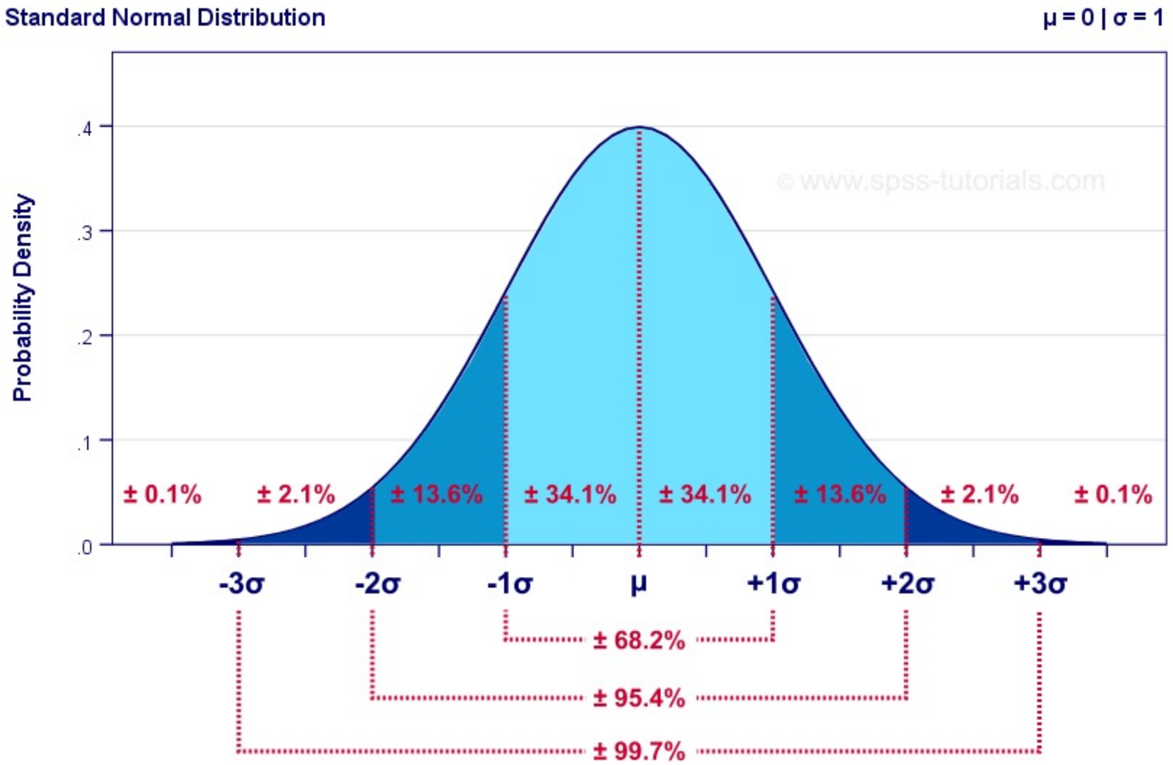
Standard normal distribution

- It is difficult to make comparisons when random variables are measured in different units.
    - E.g., age vs weight
- the standardized versions will have the same units of measurement.
- For a normal distribution, these standardized values are called Z-scores values.
- The standard normal distribution is a normal distribution of standardized.
- A variable is standardized by subtracting the mean from each value and dividing by the standard deviation.

$$X \sim N(\mu, \sigma^2) \qquad z = \frac{x - \mu}{\sigma}$$

Standard normal distribution

- A standard normal distribution will always have a mean of zero and a standard deviation of one.
- Very useful to help make comparisons between variables measured in different units, because the units of measurement are always the same for a standard normal distribution.



Standard normal distribution

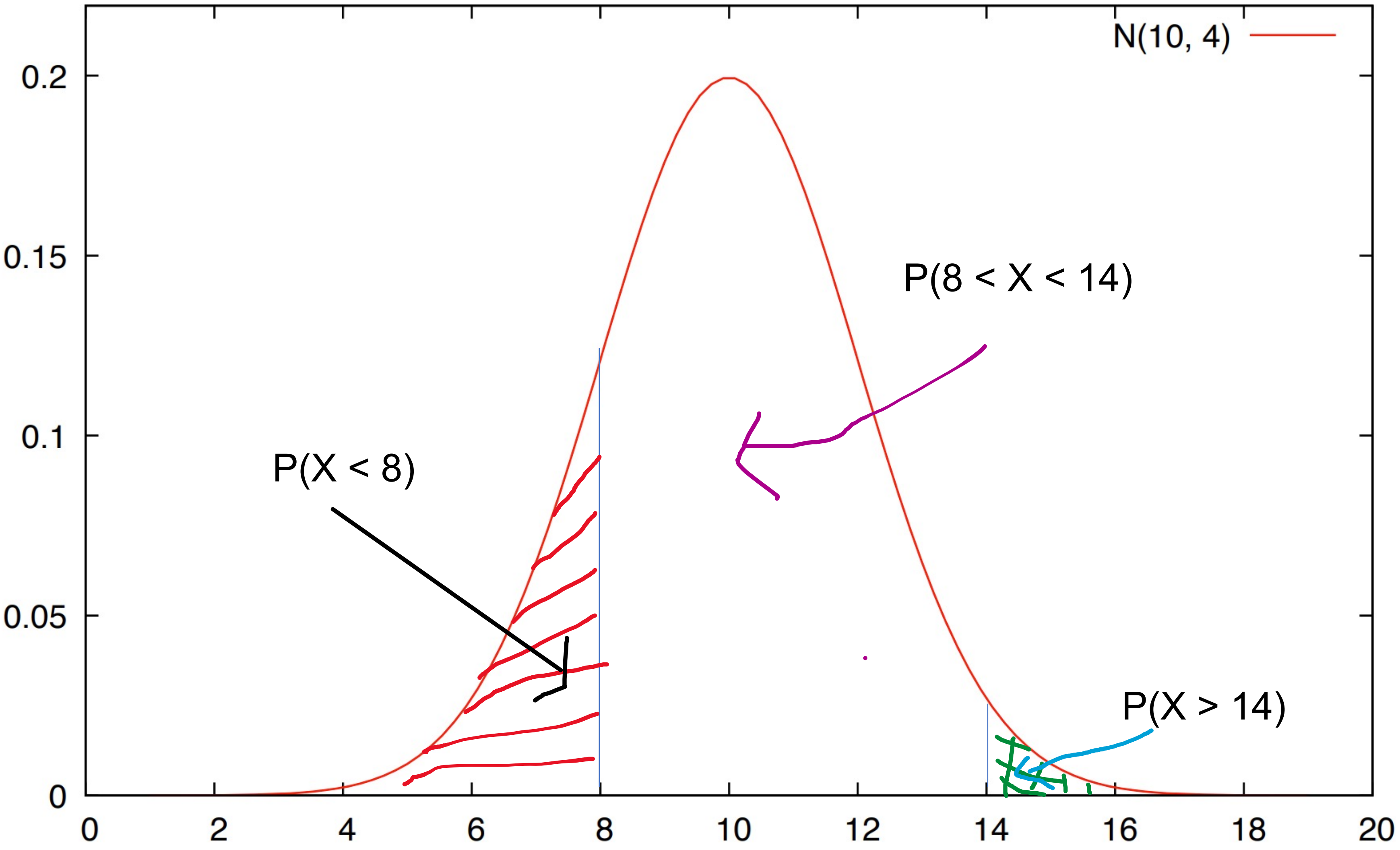
- $Z \sim N(0,1)$
- $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{1}{2}\right)x^2}$
- Mean= 0
- Variance =1

- We can use one distribution, the standard normal distribution, to make probability statements about any normally distributed variable.

Normal distribution

- The nice properties of the normal distribution facilitate calculating probabilities.
- Suppose $X \sim N(10,4)$
- What is the probability that X is greater than 14?
  - P(X > 14)
- What is the probability that X is less than 8?
  - P(X < 8)
- What is the probability that X is in between 8 and 14?
  - P(8 < X < 14)

Normal distribution



Normal distribution

- Convert to **standard** normal distribution (standardize the normal distribution)

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

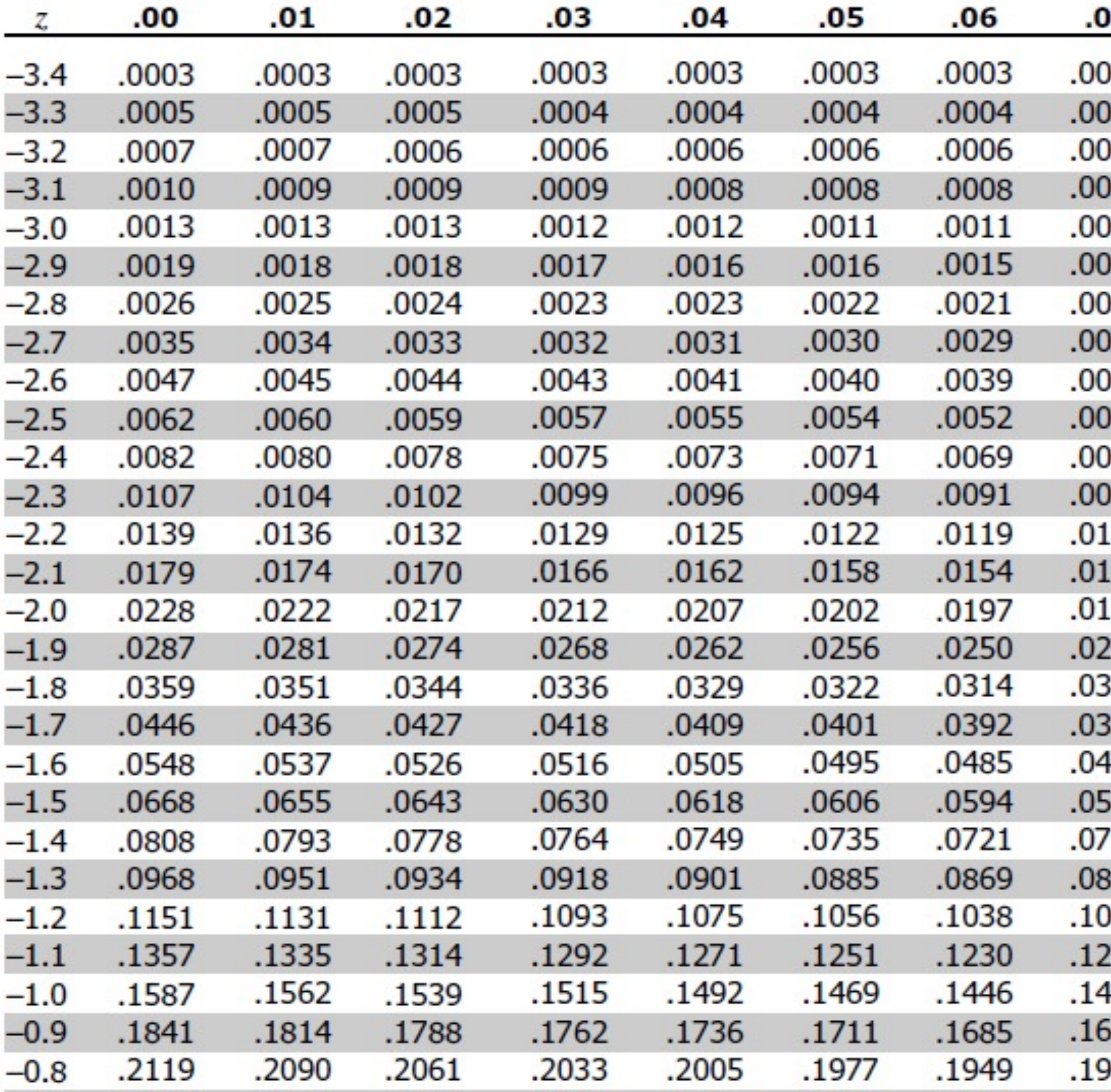
- When X=14: Z=(14-10)/2 = 2
  - 14 is 2 standard deviations above the mean of X

- When X=8 : Z=(8-10)/2 = -1
  - 8 is 1 standard deviation below the mean of X

Normal distribution

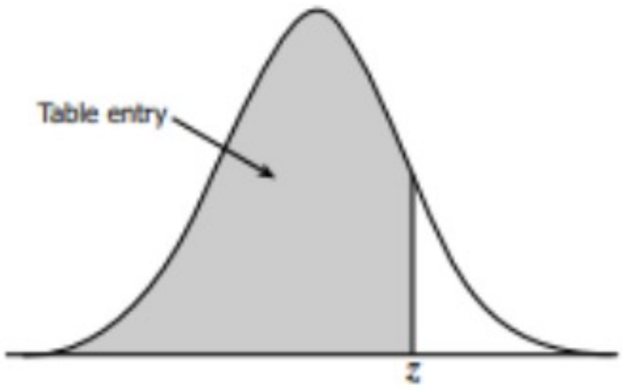
- A z-score table shows the percentage of values (usually a decimal figure) to the **left** of a given z-score on a standard normal distribution.

- Using a statistical table for the standard normal distribution (or a computer) we find:
- The area to the left of
- Z = -1.00 (i.e., left of x=8) = 0.1587

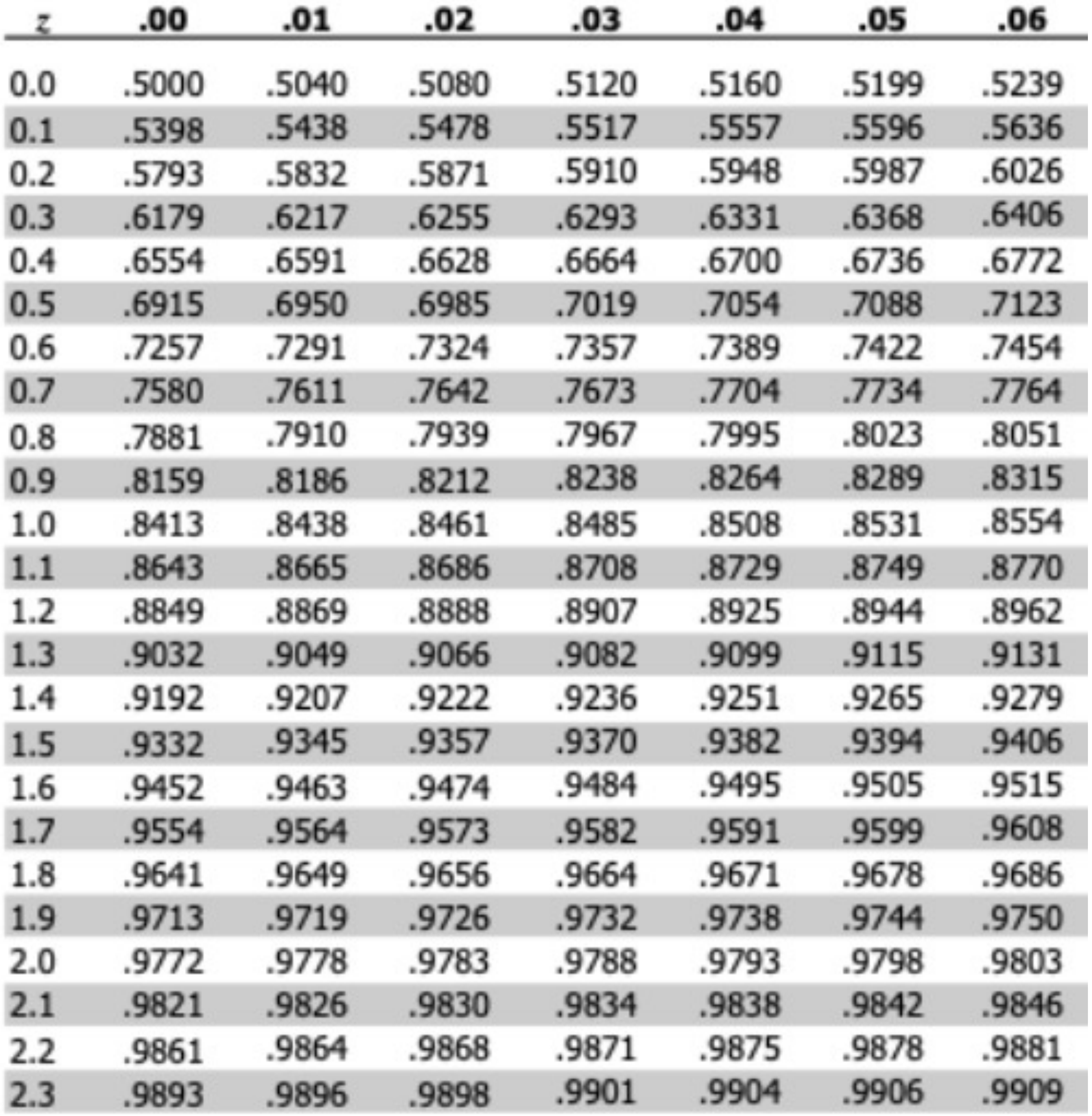


z	.00	.01	.02	.03	.04	.05	.06	.0
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.00
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.00
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.00
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.00
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.00
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.00
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.00
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.00
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.00
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.00
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.00
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.00
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.01
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.01
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.01
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.02
–1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.03
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.03
–1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.04
–1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.05
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.07
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.08
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.10
–1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.12
–1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.14
–0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.16
–0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.19

Normal distribution



- Using a statistical table for the standard normal distribution. (or a computer) we find:
- The area to the right of Z=2 (i.e., right of x=14) = 1-0.9772 = 0.0228



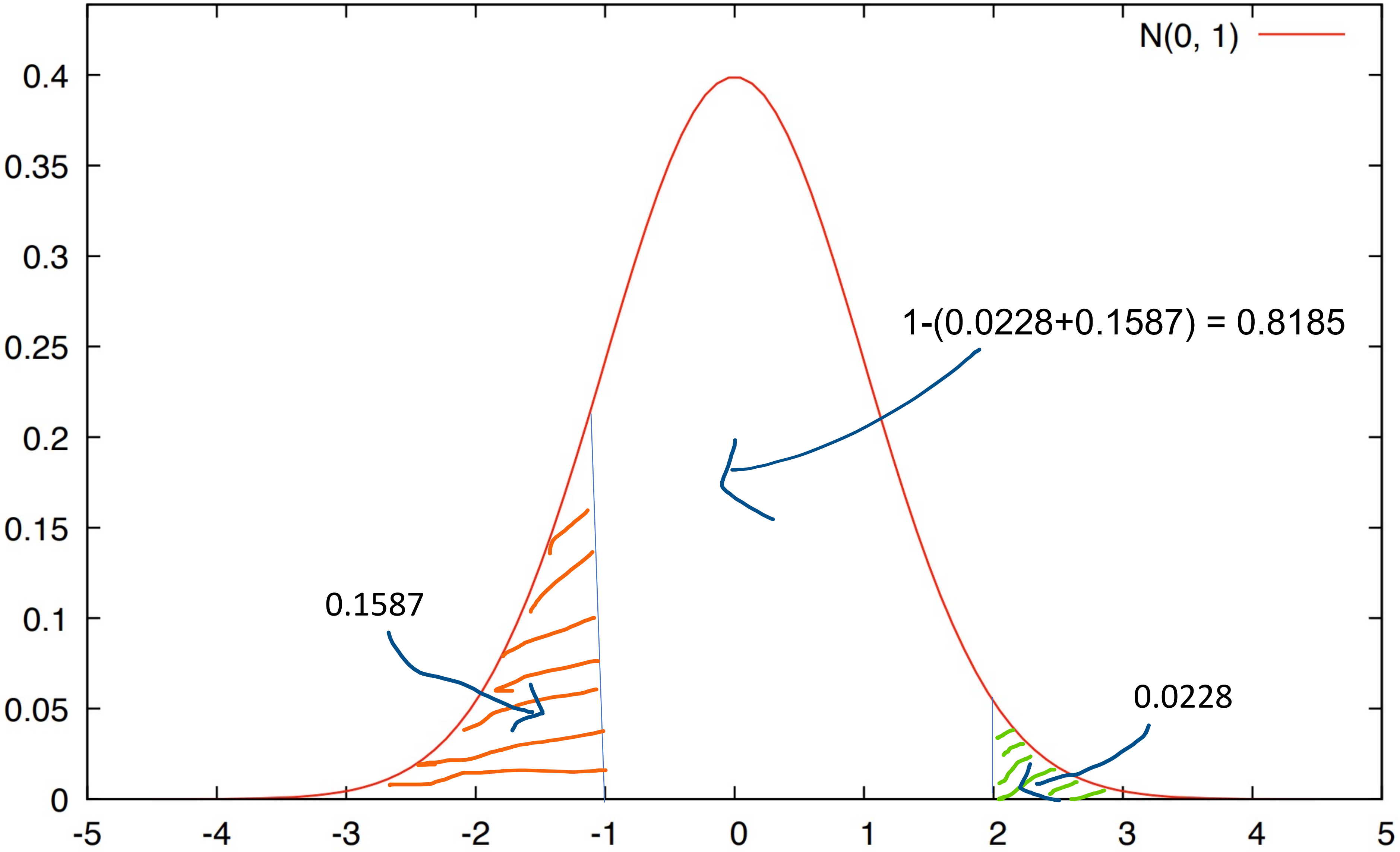
z	.00	.01	.02	.03	.04	.05	.06
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909

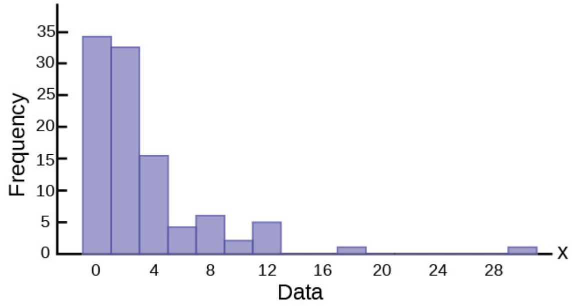
Normal distribution

- Using a statistical table for the standard normal distribution (or a computer) we find:

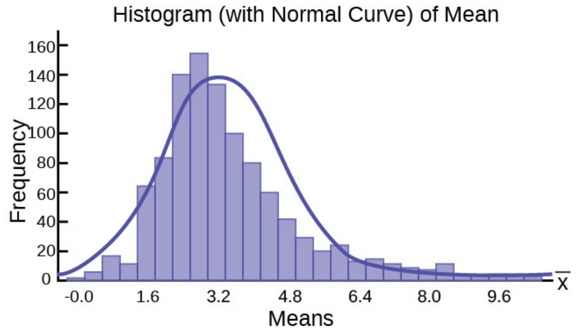
- The area to the right of Z=2 (i.e., right of x=14) = 0.0228
- The area to the left of Z=-1 (i.e., left of x=8) = 0.1587
- The area in between Z-scores of -1 and 2 = 1-(0.0228+0.1587)
  = 0.8185

Normal distribution

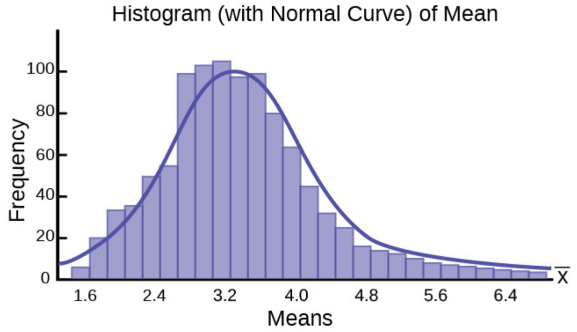




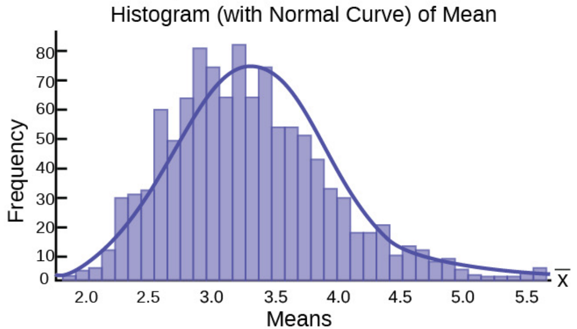
Distribution of Sample means with *n* = 10



Distribution of Sample means with *n* = 25



Distribution of Sample means with *n* = 50



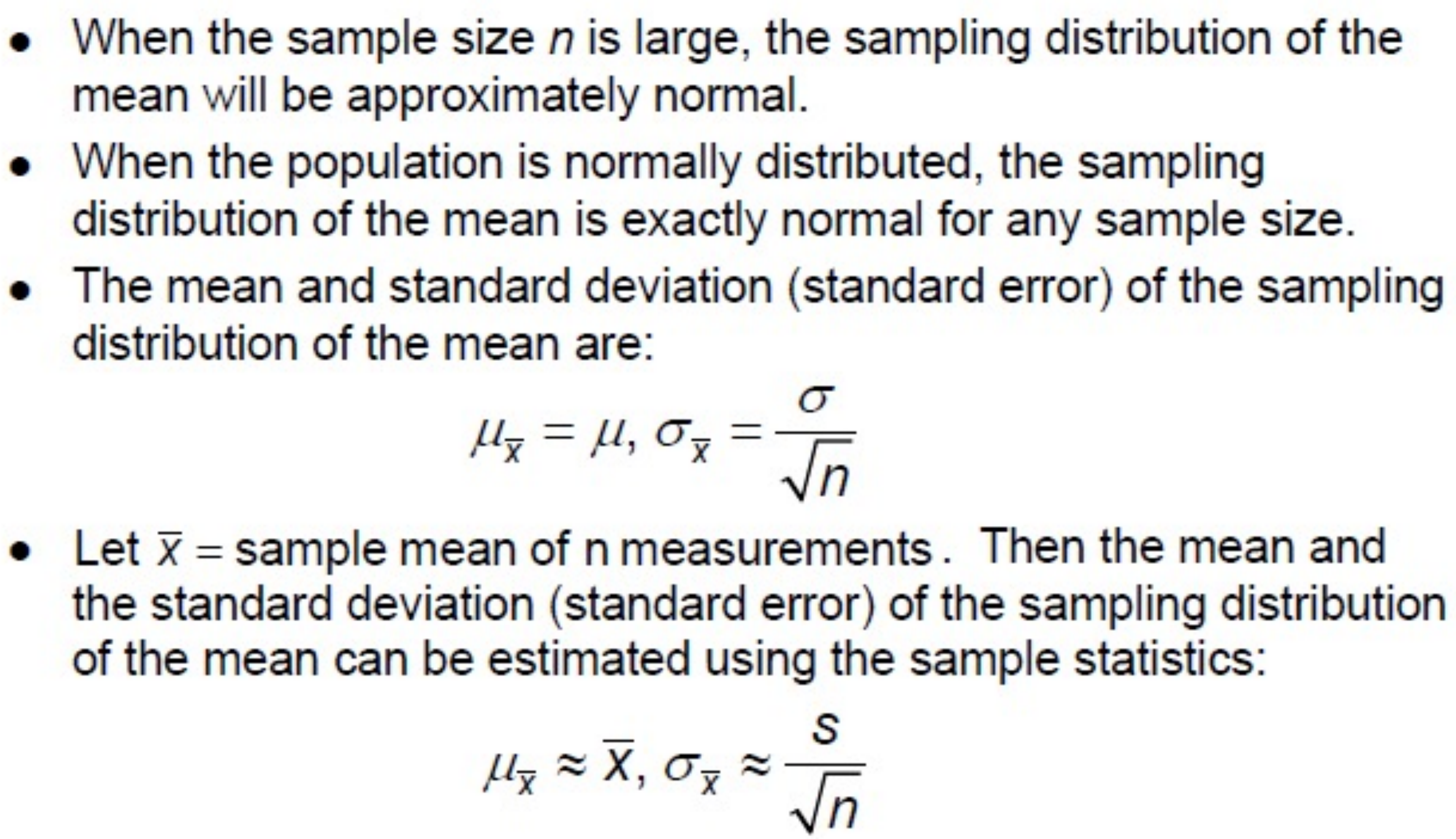
Central limit theorem (CLT)

- The central limit theorem (CLT) is one reason why the normal distribution is so important in statistical methods.

- The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed. This will hold true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually n ≥ 30).

- This means that we can use the normal probability model to quantify uncertainty when making inferences about a population mean based on the sample mean.

Central limit theorem (CLT)



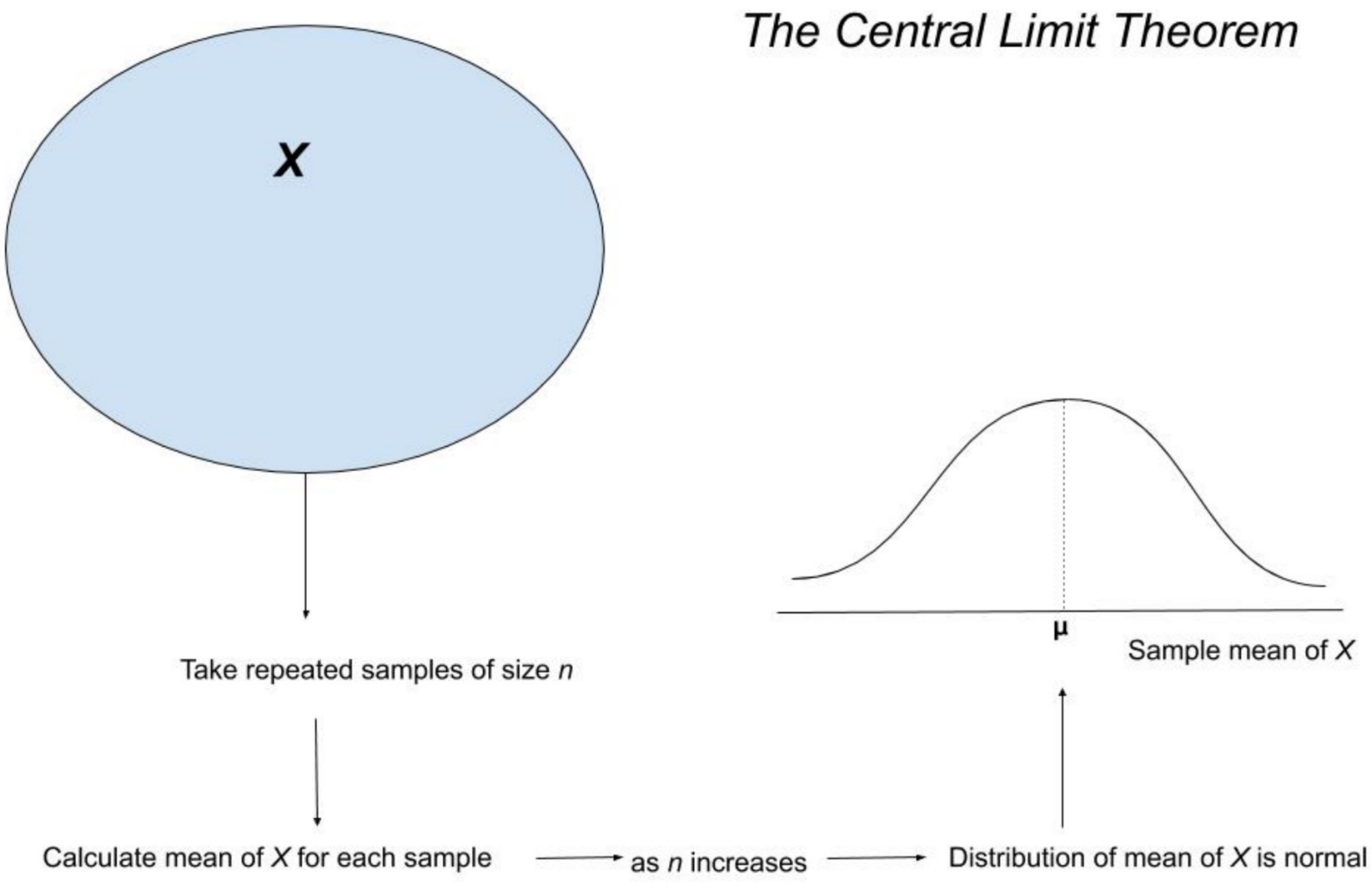
- When the sample size *n* is large, the sampling distribution of the mean will be approximately normal.
- When the population is normally distributed, the sampling distribution of the mean is exactly normal for any sample size.
- The mean and standard deviation (standard error) of the sampling distribution of the mean are:

$$\mu_{\bar{x}} = \mu,\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- Let $\bar{x}$ = sample mean of n measurements. Then the mean and the standard deviation (standard error) of the sampling distribution of the mean can be estimated using the sample statistics:

$$\mu_{\bar{x}} \approx \bar{X},\ \sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$$

Central limit theorem (CLT)



The Central Limit Theorem

Take repeated samples of size *n*

Calculate mean of *X* for each sample → as *n* increases → Distribution of mean of *X* is normal