Lecture 05: Continuous Distributions and the Distribution of Sample Means

Outline

- Continuous probability distributions: PDF and CDF
- Normal distribution
- Using Normal tables
- Normal approximation for binomial distributions
- Distribution of sample means

Continuous Probability Distributions

■The Probability Density Function (PDF,概率密度函数) of the random variable X is the curve such that the area under the curve between any two points *a* and *b* is equal to the probability that the random variable, *X,* falls between *a* and *b.*

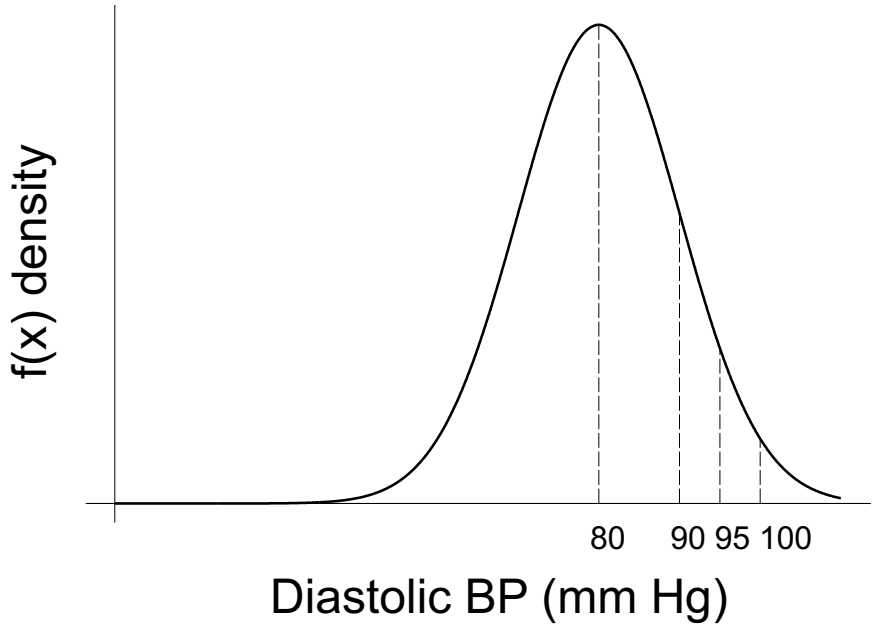
■Thus, the total area under the curve over the possible range of values for the random variable is 1.

■The Normal (or Gaussian 高斯) distribution is the most widely used distribution in statistics.

■It is used extensively in methods of estimation and hypothesis testing (actually almost everything…).

recall PMF 概率质量函数

Example: Distribution of Diastolic Blood Pressure among 35–44-year-old men

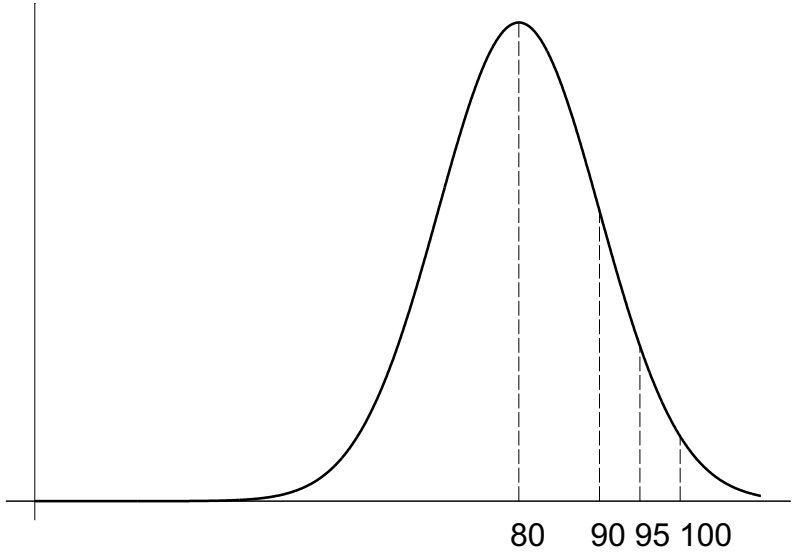


f(x) denotes PDF

$$\int_{-\infty}^{\infty} f(x)dx = 1 \qquad \Pr(a \le x \le b) = \int_{a}^{b} f(x)dx$$

$$\Pr(a) = \int_{a}^{a} f(x)dx = 0$$

Most likely value (?) is 80



Pr(90 < *x* < 95) =

Pr(Borderline Hypertensive)

Pr(95 < *x* < 100) = Pr(Mild Hypertensive)

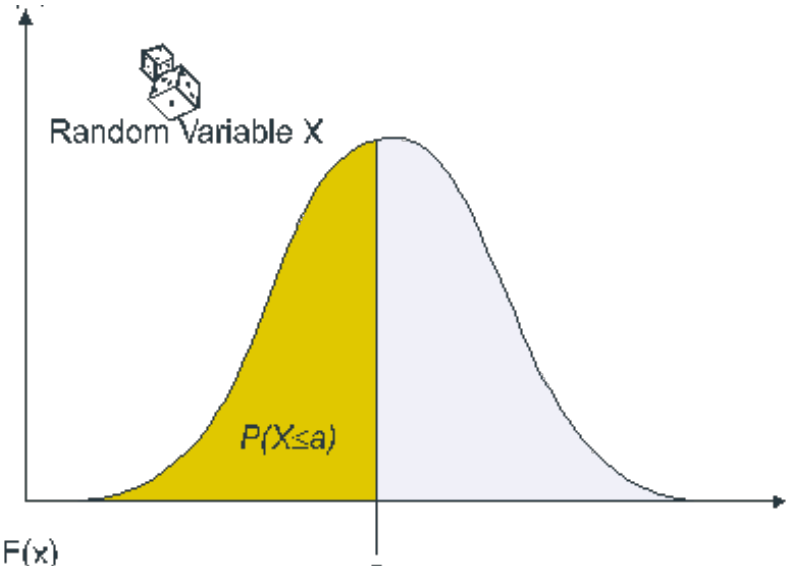
Pr(x> *100*) = Pr(Severe Hypertensive)

The Cumulative Distribution Function (CDF,累积分布函数,*F(a)*) of the random variable X evaluated at the point *a* is defined as the probability that X will take on values ≤ *a.* It is represented by the area under the PDF to the left of *a.*

$$F(a) = \Pr(X \le a) = \int_{-\infty}^{a} f(x)dx$$

CDF vs PDF

PDF

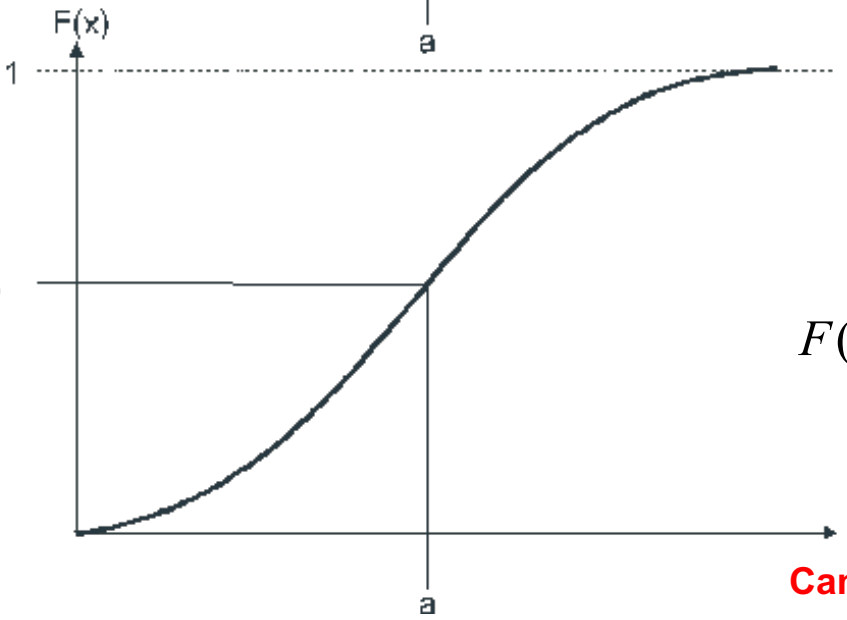


Pr(x)

or,

f(x),p(x)

CDF

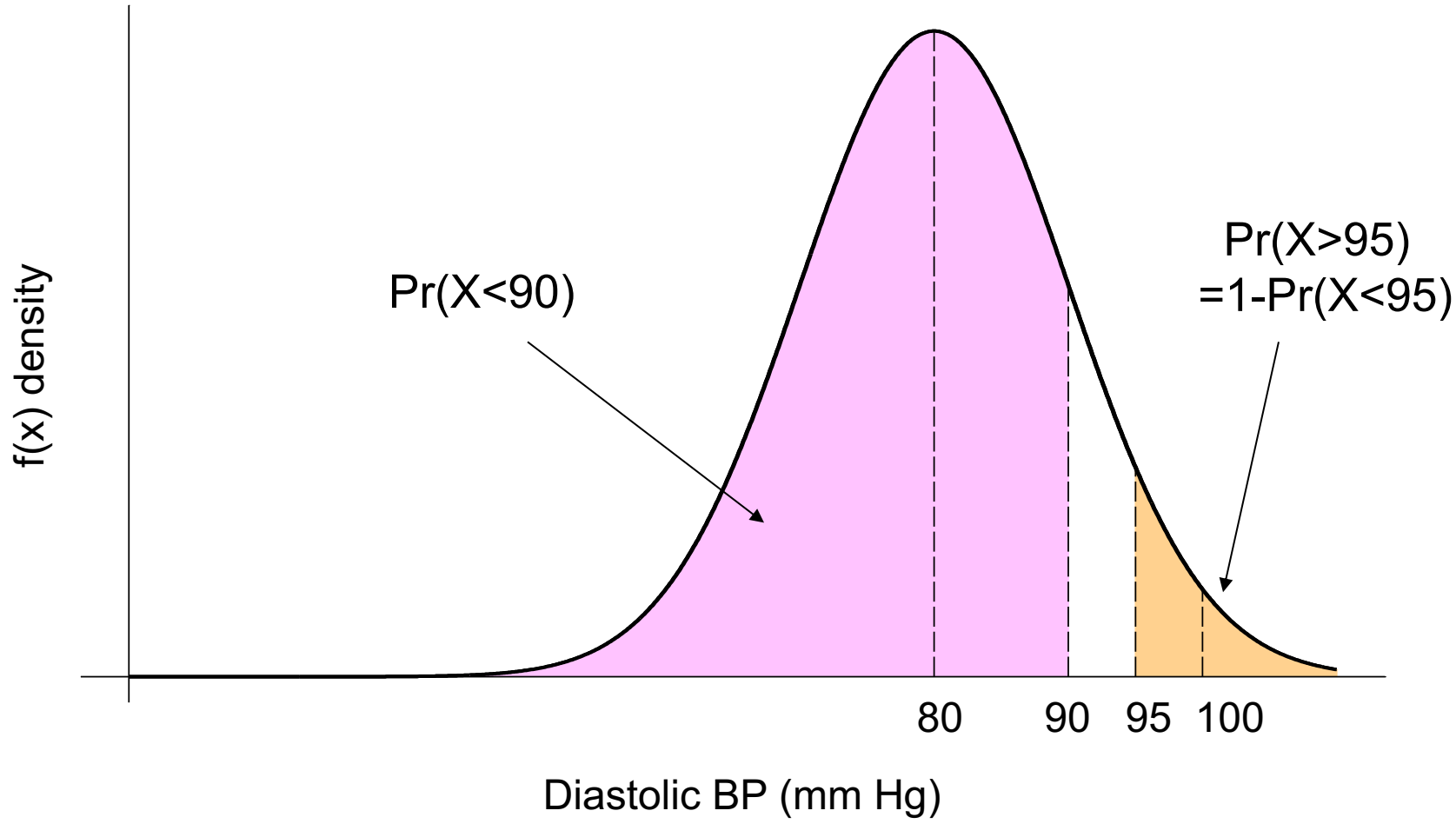


F(x)

$$F(a) = \Pr(X \le a) = \int_{-\infty}^{a} f(x)dx$$

Flashback: CDF for discrete RV!

Can F(x) be greater than 1? And f(x)?



Note: $Pr(X<90) = Pr(X \leq 90) = F(90)$
Since $Pr(X=90)=0$

Expected Value of continuous RV

The Expected Value of a continuous random variable X, denoted E(X) or μ is the average taken on by the random variable:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

The variance of a continuous random variable X, denoted by Var(X) or σ^2, is the average squared deviation of each value of the random variable from its expected value.

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

Remember expected value for discrete RV?

$$E(X) = \mu = \sum_{i=1}^{k} x_i \Pr(X = x_i) \qquad Var(X) = \sigma^2 = \sum_{i=1}^{k} (X_i - \mu)^2 \Pr(X = x_i) = E(X - \mu)^2$$

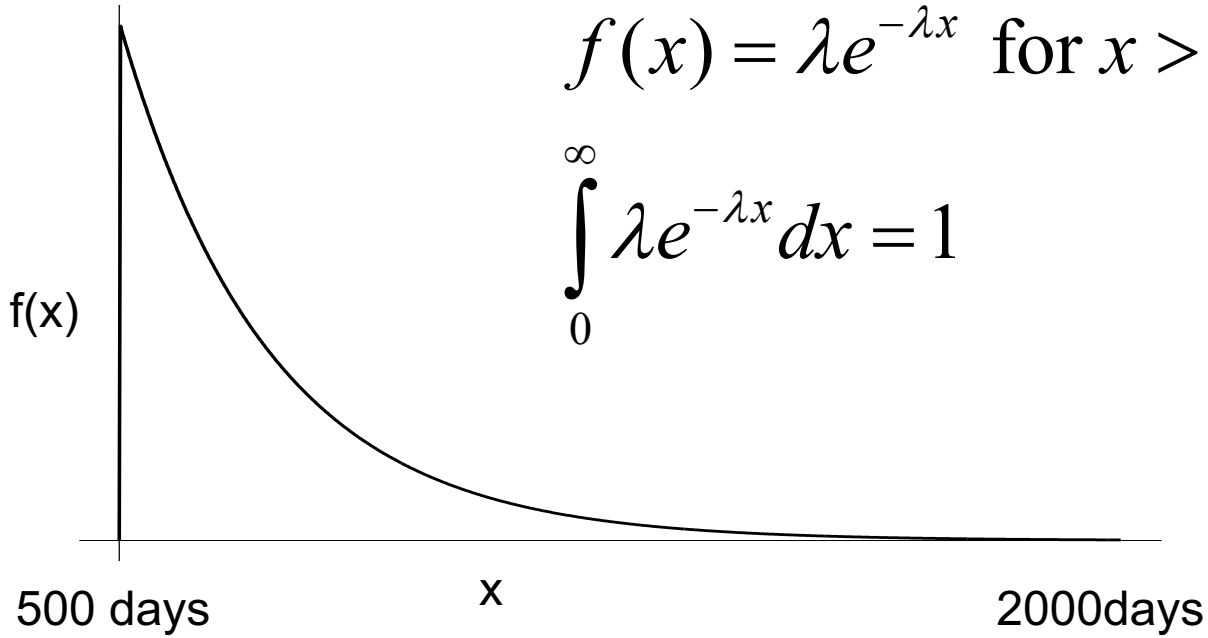
Another example: Exponential Distribution

The lifespan of lightbulbs (or medical equipment) usually follows an Exponential Distribution. The products below a certain threshold are removed from final sale (quality control).

λ is the so-called failure rate
1/**λ** is the expected lifespan

$$f(x) = \lambda e^{-\lambda x} \text{ for } x > 0, \lambda > 0$$

$$\int_0^{\infty} \lambda e^{-\lambda x} dx = 1$$



Exponential distribution is a memoryless distribution, like Bernoulli!

Expected value and variance of Exponential Distribution

$$E(X) = \int_0^\infty x \cdot f(x) dx = \int_0^\infty x \cdot \lambda e^{-\lambda x} dx$$

$$= \lambda \left[-x \frac{1}{\lambda} e^{-\lambda x} + \frac{1}{\lambda} \int e^{-\lambda x} dx \right]_0^\infty$$

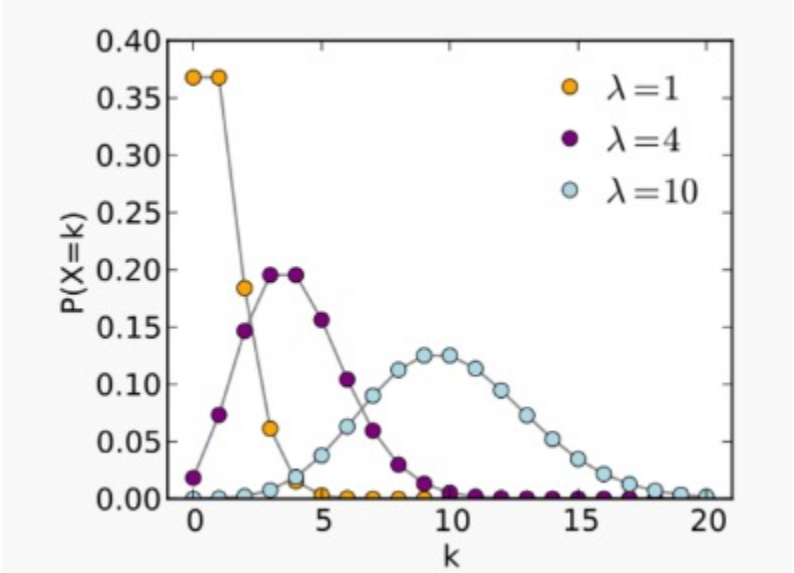
$$= \lambda \left[-x \frac{1}{\lambda} e^{-\lambda x} - \frac{1}{\lambda^2} e^{-\lambda x} \right]_0^\infty = \lambda \left[(0-0) - \left(0 - \frac{1}{\lambda^2} \right) \right] = \frac{1}{\lambda}$$

$$Var(X) = \int_0^\infty \left(x - E(X) \right)^2 f(x) dx = \int_0^\infty \left(x - \frac{1}{\lambda} \right)^2 \lambda e^{-\lambda x} dx = \frac{1}{\lambda^2}$$

Related discrete distribution: Poisson distribution

$$P(k \text{ events in interval}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

The number of accidents（NOT a binary variable about whether an event happens) in a specific interval often follows a Poisson distribution.



The number of neuronal spikes in a specific interval is often assumed to follow a Poisson distribution.

$$E(X) = \lambda$$

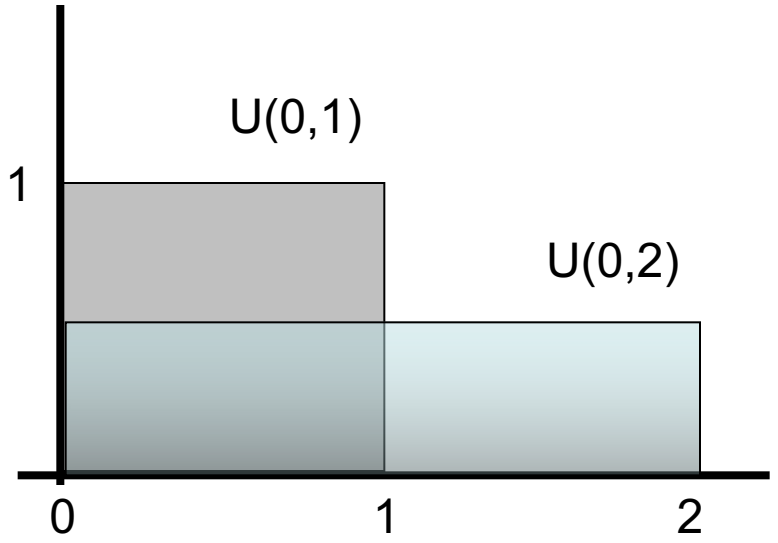
$$Var(X) = \lambda$$

Is it memoryless?

Some other continuous distributions

- Uniform
- Normal
- t
- F
- χ^2
- ….

Examples of uniform distribution



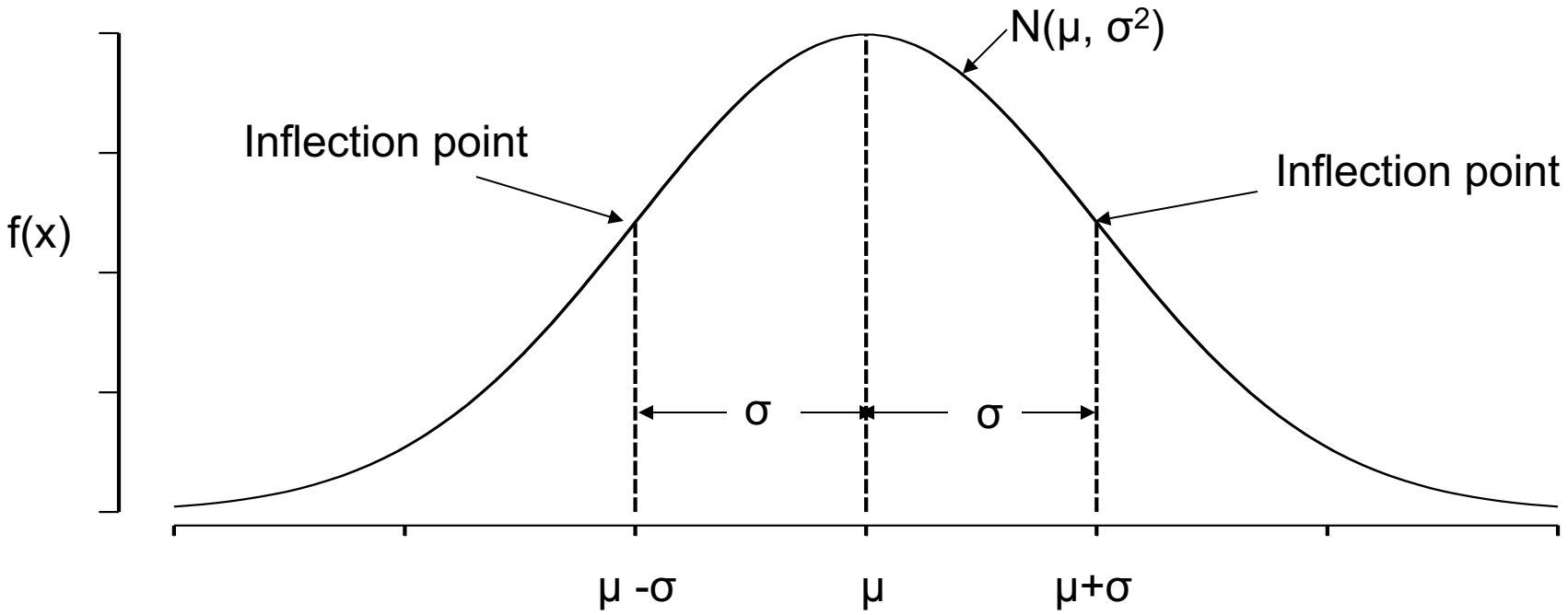
The Normal (Gaussian) Distribution

正态，高斯分布

Probability Density Function

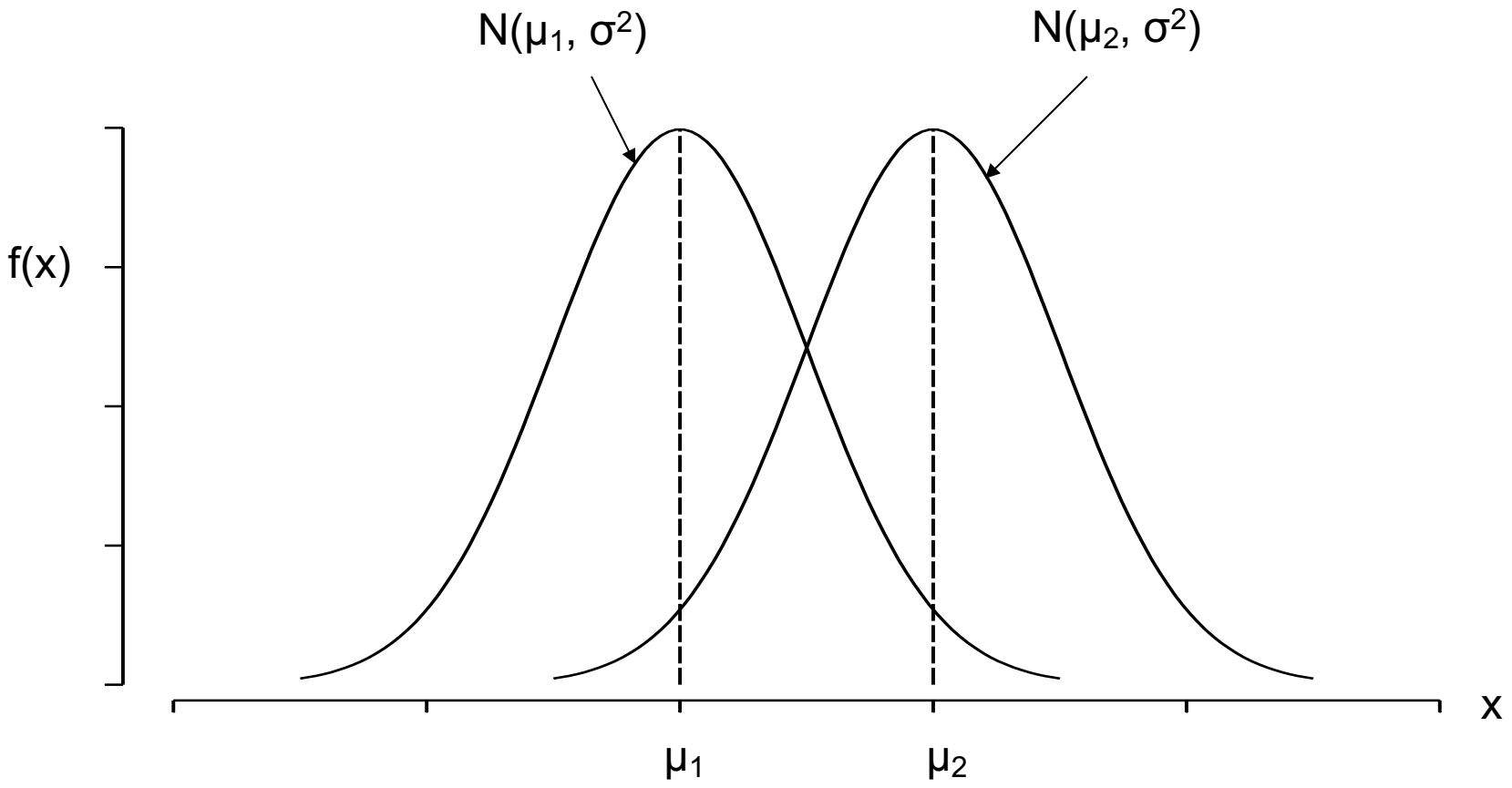
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty$$

Can f(x)>1?

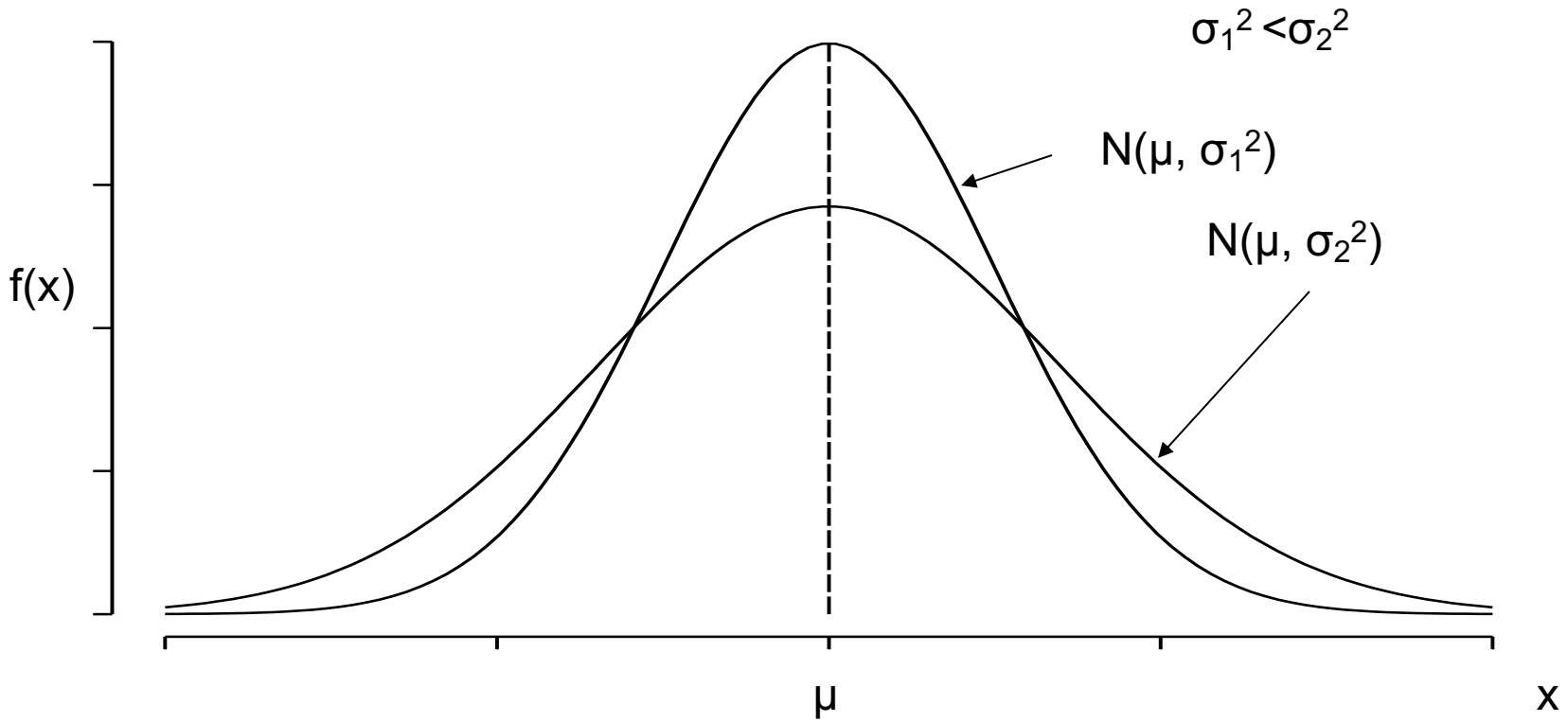


The normal distribution is denoted by $N(\mu, \sigma^2)$. Note: Inflection point is the point where the second derivative changes sign.

Two normal distributions: different means, same variance



Two normal distributions: same mean, different variances

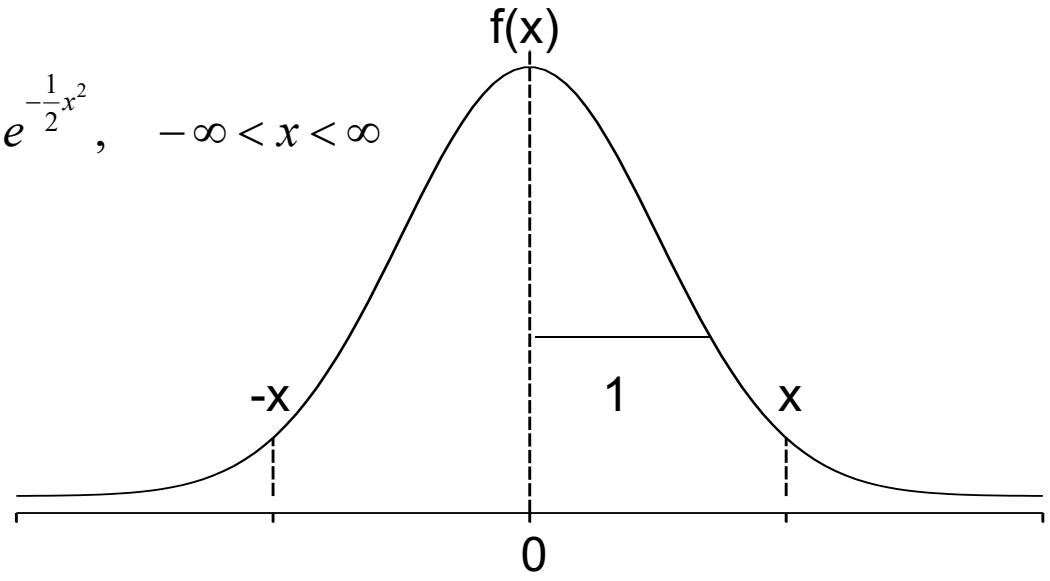


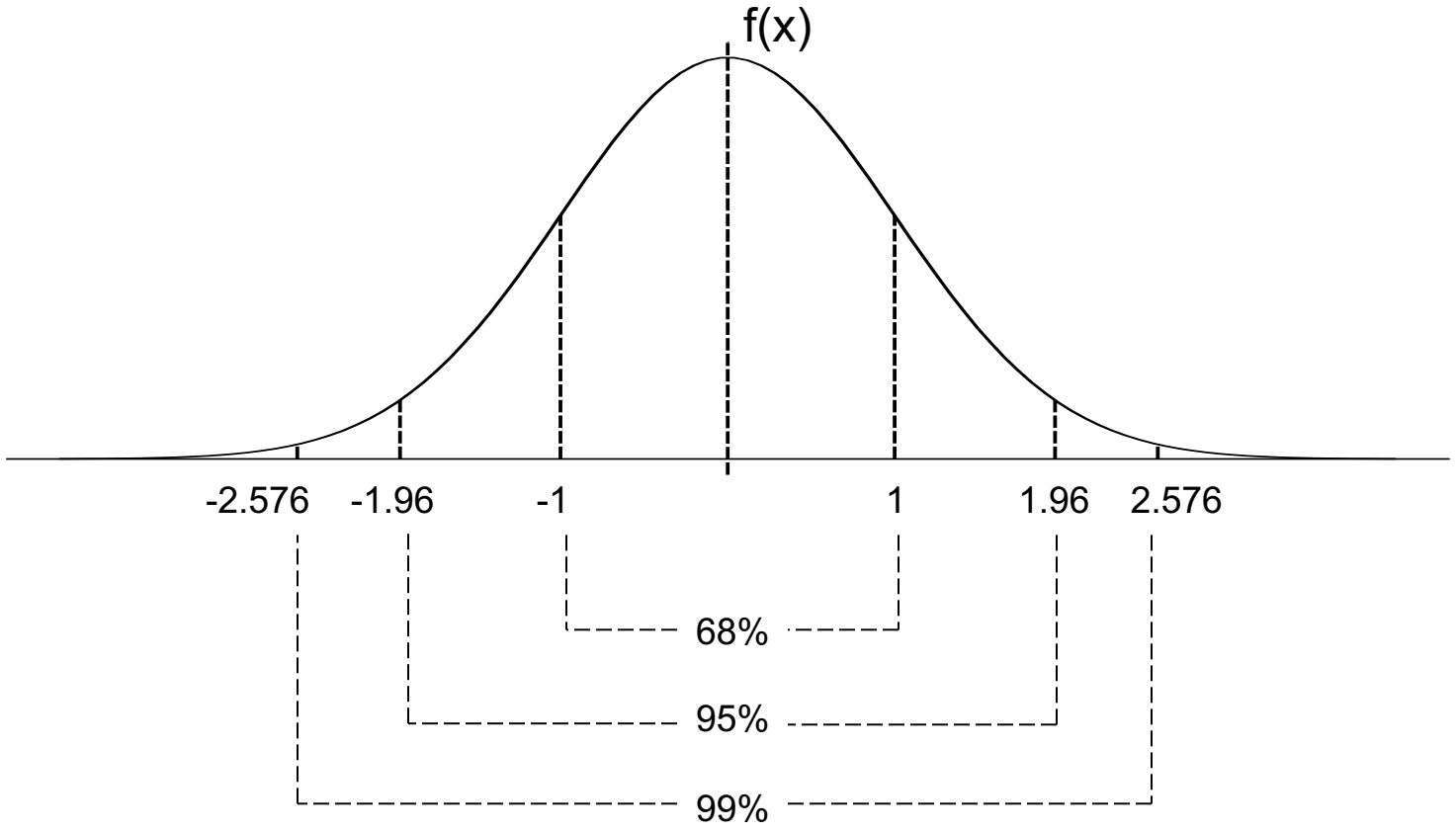
Standard Normal Distribution

- N(0,1) is referred to as the standard normal distribution (标准正态分布，or unit normal, or ***z distribution***).
- Any normal distribution can be transformed into N(0, 1).
- PDF:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2},\quad -\infty<x<\infty,\ \ with\ \ \mu=0,\ \ \sigma^2=1$$

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2},\quad -\infty<x<\infty$$



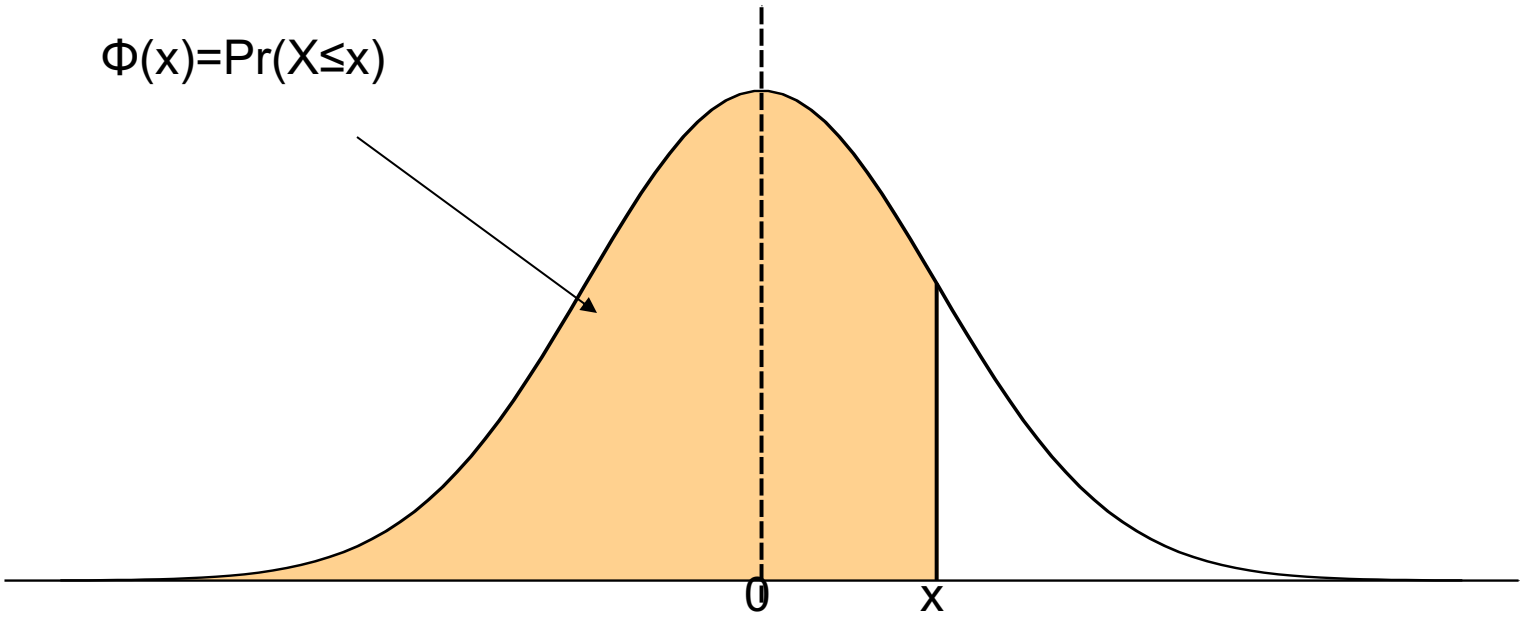


68% of area lies within 1 sd of mean (-1,1)
95% of area lies within 1.96 sd of mean (-1.96,1.96)
99% of area lies within 2.576 sd of mean (-2.576,2.576)

CDF of N(0,1)

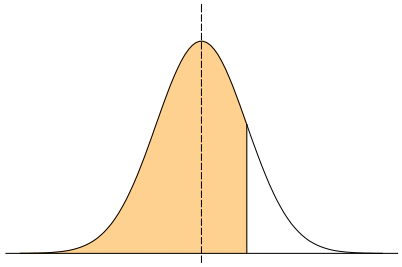
The cumulative distribution function for a standard normal distribution N(0,1) is denoted as

$$\Phi(x) = \Pr(X \le x) = F(x)$$



x = 0.13, what is F(x)?

Normal Table from wiki



Second decimal place of z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857

Normal Table from the textbook

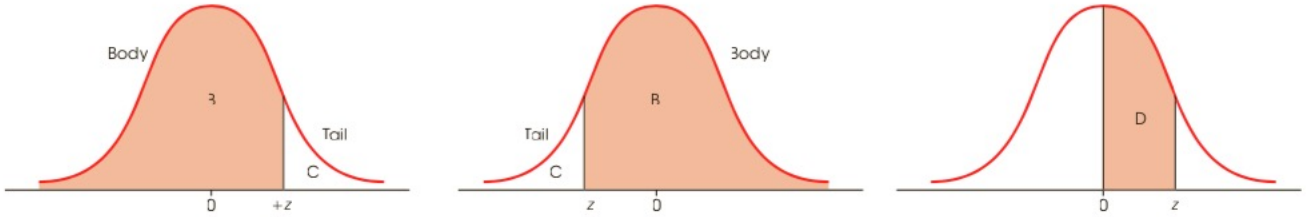
APPENDIX

B

Statistical Tables

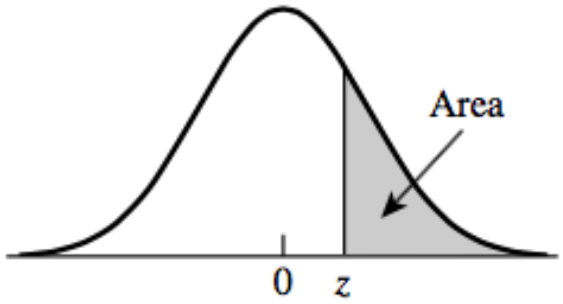
TABLE B.1 The Unit Normal Table*

*Column A lists z-score values. A vertical line drawn through a normal distribution at a z-score location divides the distribution into two sections.
Column B identifies the proportion in the larger section, called the *body*.
Column C identifies the proportion in the smaller section, called the *tail*.
Column D identifies the proportion between the mean and the z-score.
Note: Because the normal distribution is symmetrical, the proportions for negative z-scores are the same as those for positive z-scores.



(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and z	(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and z
0.00	.5000	.5000	.0000	0.25	.5987	.4013	.0987
0.01	.5040	.4960	.0040	0.26	.6026	.3974	.1026
0.02	.5080	.4920	.0080	0.27	.6064	.3936	.1064
0.03	.5120	.4880	.0120	0.28	.6103	.3897	.1103
0.04	.5160	.4840	.0160	0.29	.6141	.3859	.1141
0.05	.5199	.4801	.0199	0.30	.6179	.3821	.1179
0.06	.5239	.4761	.0239	0.31	.6217	.3783	.1217
0.07	.5279	.4721	.0279	0.32	.6255	.3745	.1255
0.08	.5319	.4681	.0319	0.33	.6293	.3707	.1293
0.09	.5359	.4641	.0359	0.34	.6331	.3669	.1331
0.10	.5398	.4602	.0398	0.35	.6368	.3632	.1368

Normal table from Wackerly et al. 2008 (Table 4 in Appendix 3)



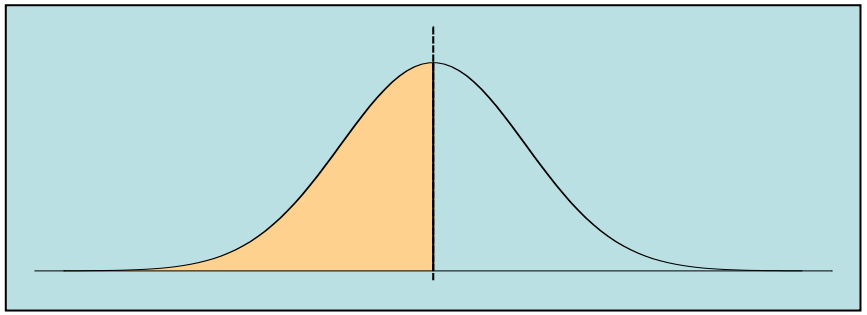
	Second decimal place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.00135									
3.5	.000 233									
4.0	.000 031 7									
4.5	.000 003 40									
5.0	.000 000 287									

From R. E. Walpole, *Introduction to Statistics* (New York: Macmillan, 1968).

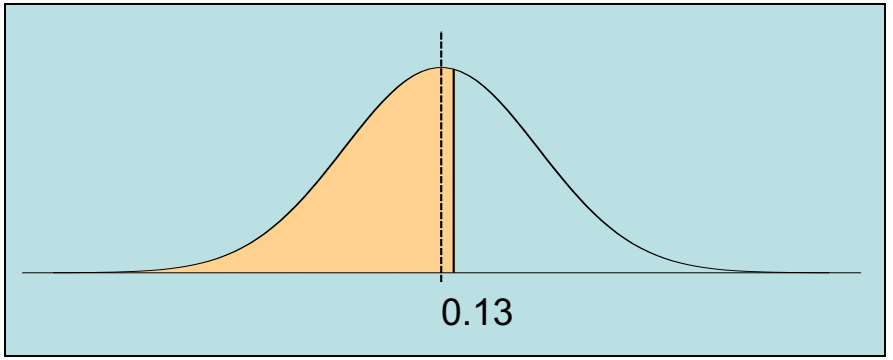
Using Normal Tables

From the table

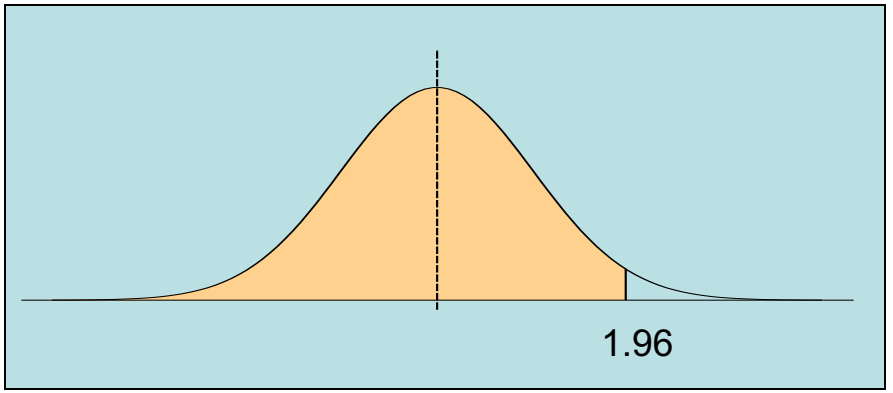
$\Pr(X \leq 0) = 0.5$



$\Pr(X \leq 0.13) = 0.5517$

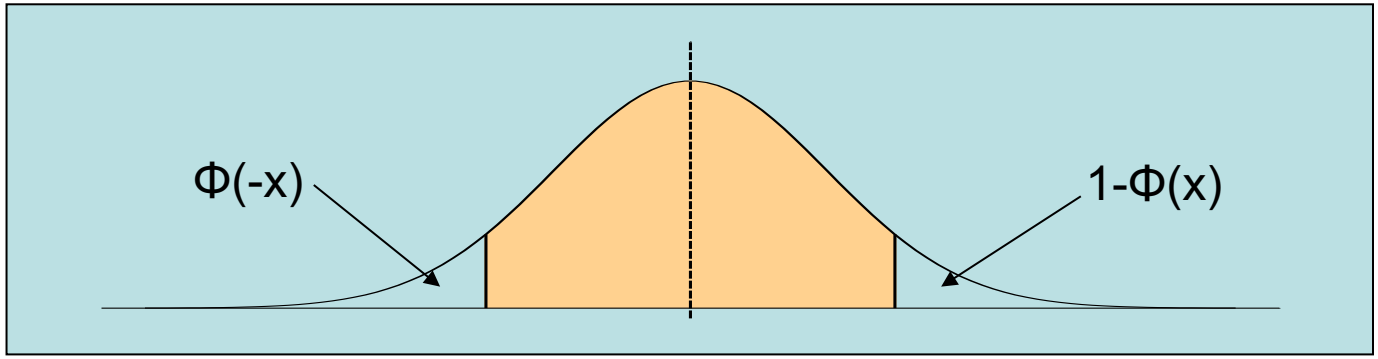


$\Pr(X \leq 1.96) = 0.9750$

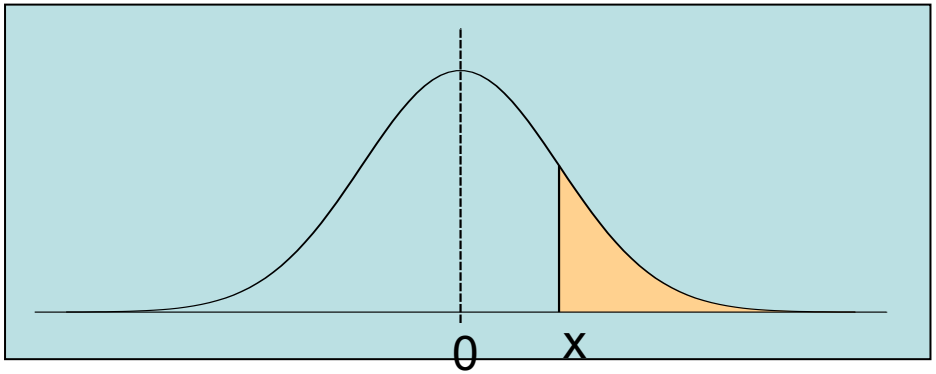


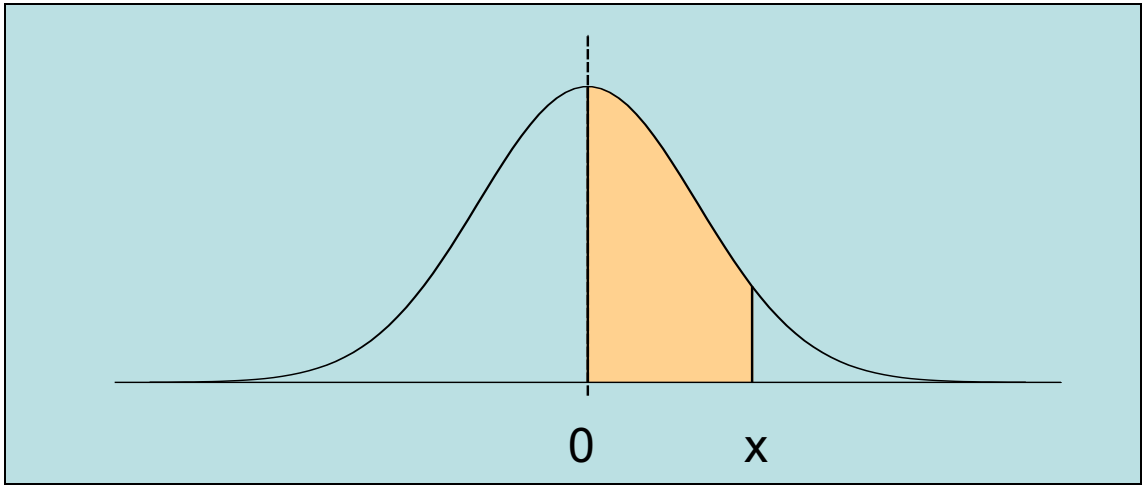
Symmetry of Φ and Z score

$$\Phi(-x) = \Pr(X \le -x) = \Pr(X \ge x) = 1 - \Pr(X \le x) = 1 - \Phi(x)$$

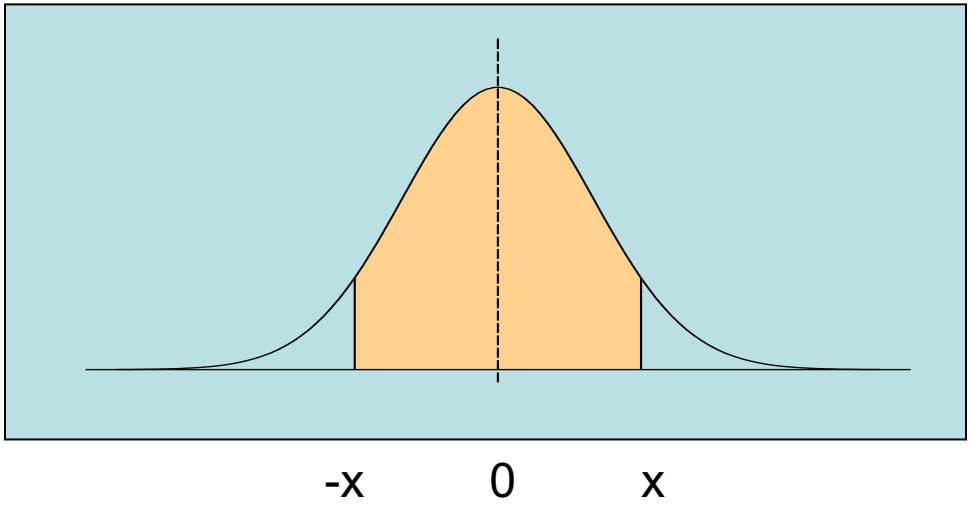


Since Pr(X≤x) + Pr(X≥x)=1, Pr(X≥x)=1–Pr(X≤x)





X = 1.2; What's the area of the shaded area?

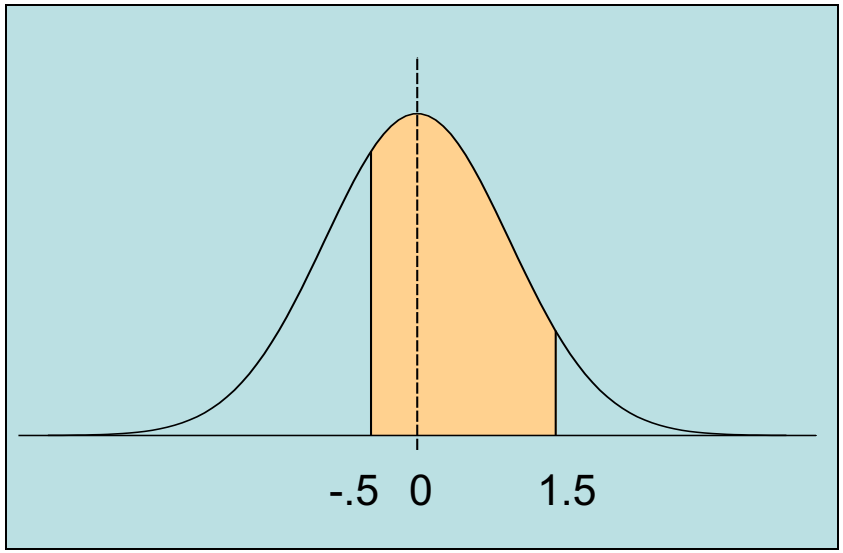


X = 1.2 again, what's the area of the shaded area?

Practice looking up values in N(0,1) table

Let X~N(0,1)

- $\Pr(X \le 0)=0.5$
- $\Pr(X \le -1.96)=0.025$
- $\Pr(X \le 1.96)=0.975$
- $\Pr(-1.96 \le X \le 1.96)=0.95$



- $\Pr(-0.5 \le X \le 1.5) = \Pr(X \le 1.5) - \Pr(X \le -0.5)$
  $= 0.9332 - 0.3085$
  $= 0.6247$
- $\Pr(|X| \ge 1.96) = 1 - \Pr(-1.96 \le X \le 1.96)$
  $= 1 - 0.95$
  $= 0.05$

Percentiles and Percentile Ranks

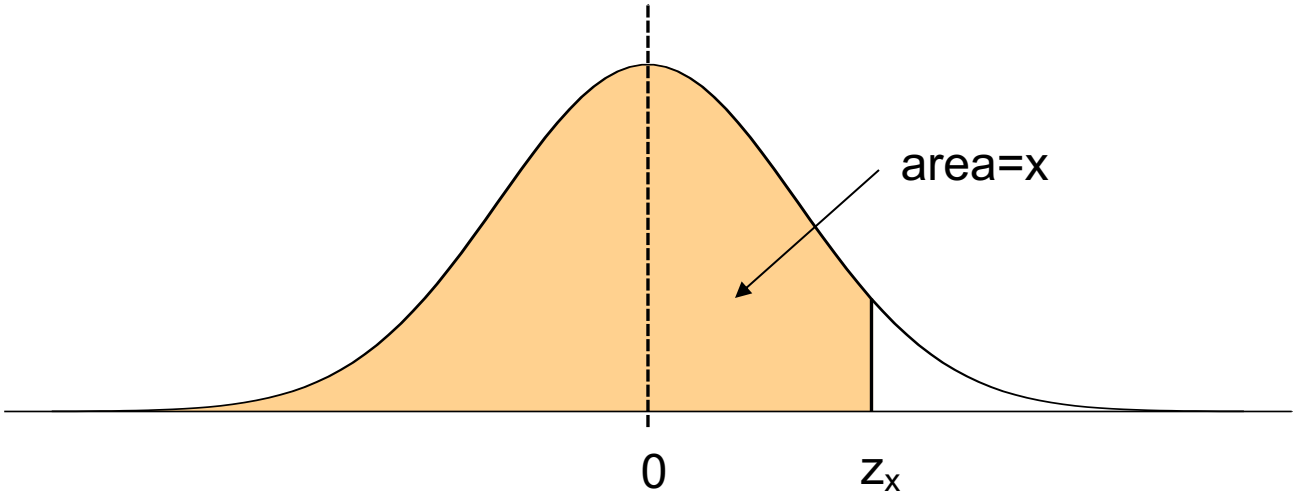
- The relative location of individual scores within a distribution can be described by percentiles and percentile ranks.
- The **percentile rank** for a specific X value is the percentage of individuals with scores at or below that value.
- When a score is referred to by its rank, the score is called a **percentile**. The percentile rank for a score in a normal distribution is simply the proportion to the left of the score.
- A grade of 85 is a percentile, corresponding to a percentile rank of 30% (P30; or 40% if allowed).

Z score and CDF

The x^{th} percentile of a standard normal distribution is denoted by z_x (z分数). It is defined by the relationship.

$$\Phi(z_x) = F(z_x) = \Pr(X < z_x) = x$$

where X~N(0,1).



Three very commonly used percentiles

- Find point such that $\Pr(X<z_{.975})=.975$

$$\Phi(1.96) = 0.975 \Rightarrow z_{.975} = 1.96$$

- Find point such that $\Pr(X<z_{.95})=.95$

$$\Phi(1.645) = 0.95 \Rightarrow z_{.95} = 1.645$$

- Find point such that $\Pr(X<z_{.025})=.025$

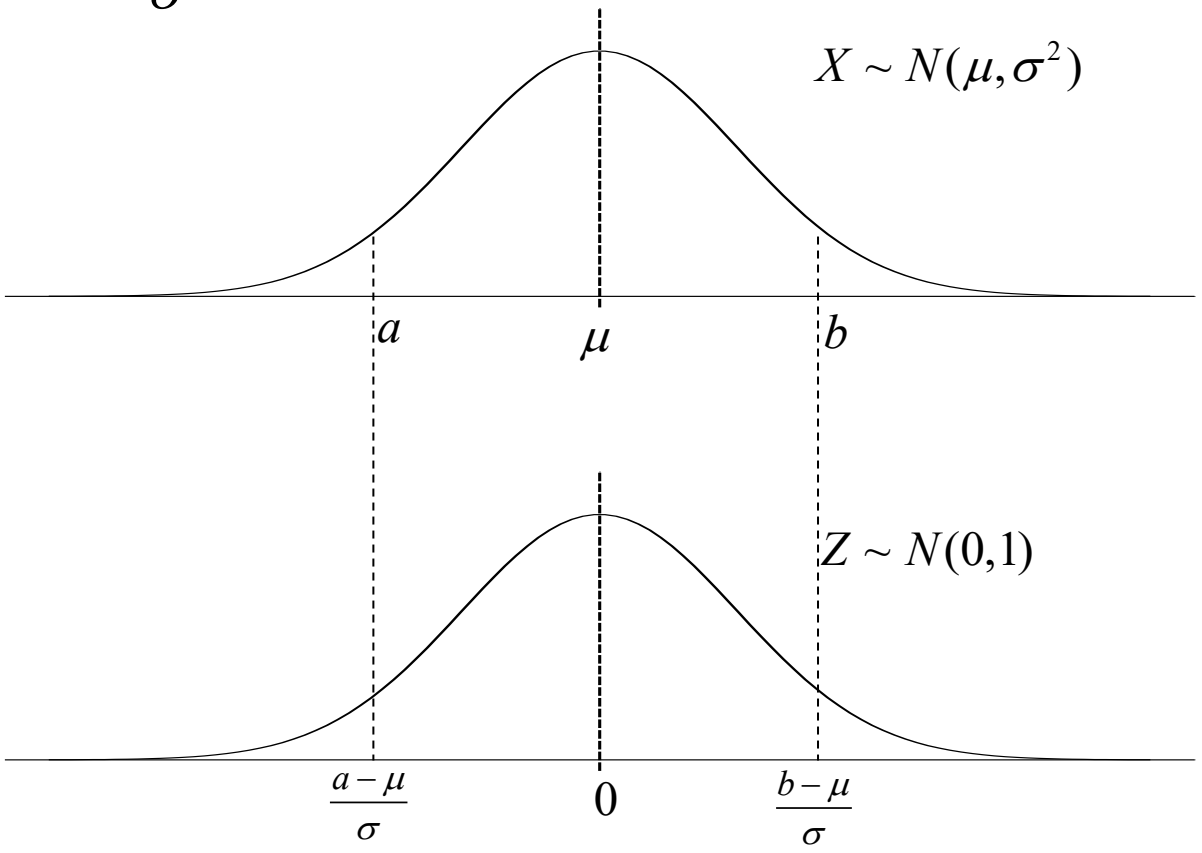
$$\Phi(-1.96) = 0.025 \Rightarrow z_{.025} = -1.96$$

$$\left(\Phi^{-1}(0.025) = -1.96\right)$$

From N(μ, σ²) to N (0, 1)

Suppose X~N(μ, **σ**^{2}). We want Pr(a≤X≤b) for general a,b, a≤b.

consider $Z = \frac{X - \mu}{\sigma}$ → Z~N(0,1).



Proof

First note that for *X* normal, the transformation to *Z* preserves normality.

$$E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}E(X-\mu) = \frac{1}{\sigma}(E(X)-\mu) = 0$$

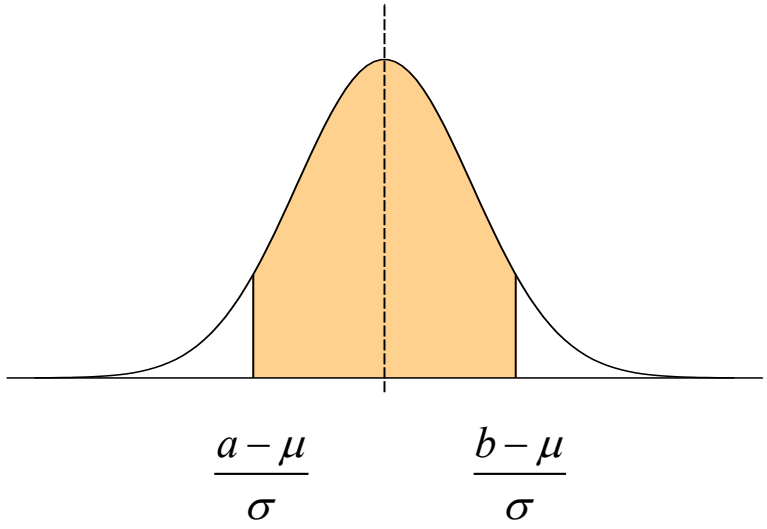
$$Var(Z) = Var\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}E(X-\mu)^2 = \frac{1}{\sigma^2}\sigma^2 = 1$$

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

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

How to compute Pr(a ≤ X ≤ b) for $X \sim N(\mu, \sigma^2)$, b≥a

$$
\begin{aligned}
a \le \quad & X \quad \le b \\
a-\mu \le\ & X-\mu \le b-\mu \\
\frac{a-\mu}{\sigma} \le\ & \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma} \\
\frac{a-\mu}{\sigma} \le\ & \quad Z \quad \le \frac{b-\mu}{\sigma}
\end{aligned}
$$



Thus

$$\Pr(a \le X \le b) = \Pr\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Let X be a normal, mean =10, SD=2, Var=4 : X~N(10,4). Find Pr(11≤X≤13.6)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857

Example

Let X be a normal, mean =10, SD=2, Var=4 : X~N(10,4)

Find Pr(11≤X≤13.6)

$$\Pr(11 \le X \le 13.6) = \Pr\left(\frac{11-10}{2} \le \frac{X-10}{2} \le \frac{13.6-10}{2}\right)$$

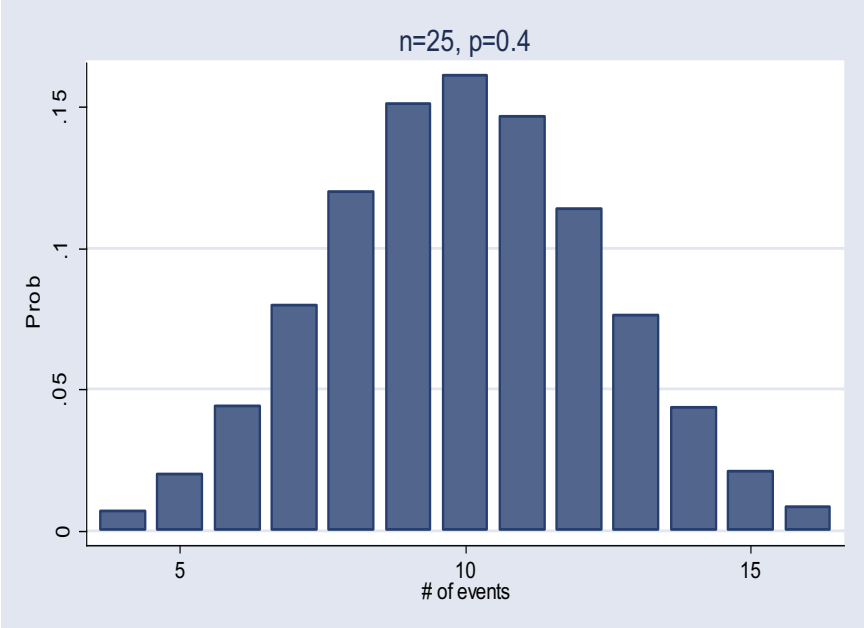
$$= \Pr(0.5 \le Z \le 1.8) = \Pr(Z \le 1.8) - \Pr(Z \le 0.5)$$

$$= 0.9641 - 0.6915$$

$$= 0.2726$$

Normal approximation to binomial distributions

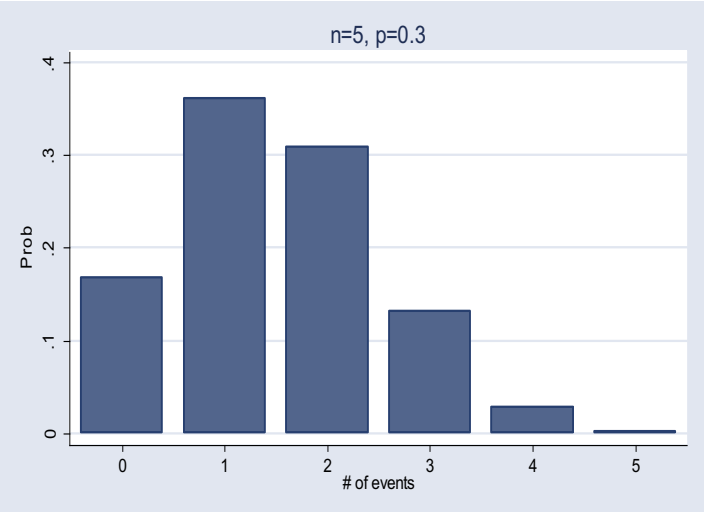
When *n* is large and *p* is not near 0 or 1, the binomial distribution tends to be symmetric and is well approximated by a normal distribution with mean $\mu=np$, variance $\sigma^2=npq$.

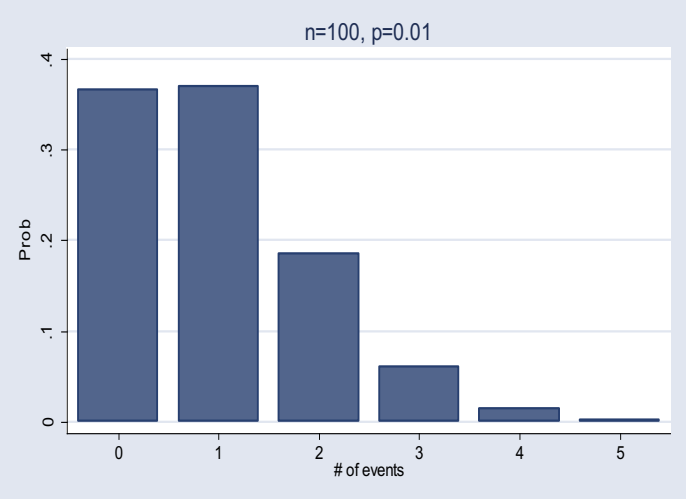


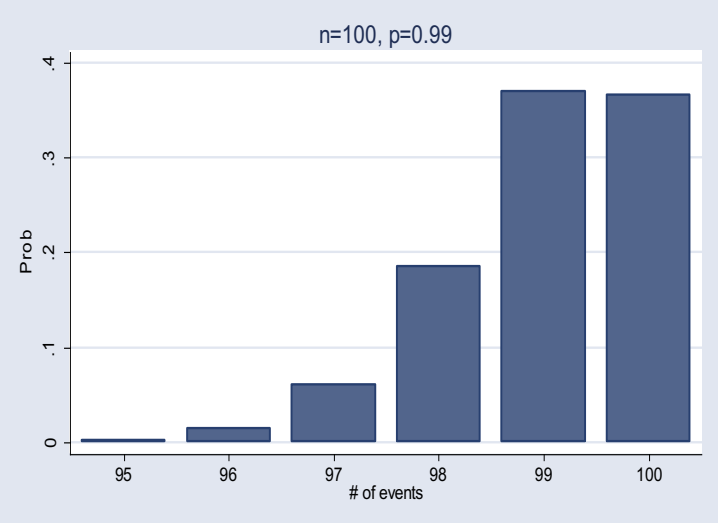
E(X) for binomial distributions = *np*
Var(X) for binomial distributions = *npq,* where *q = 1–p.*
Use N(*np, npq*) to approximate it

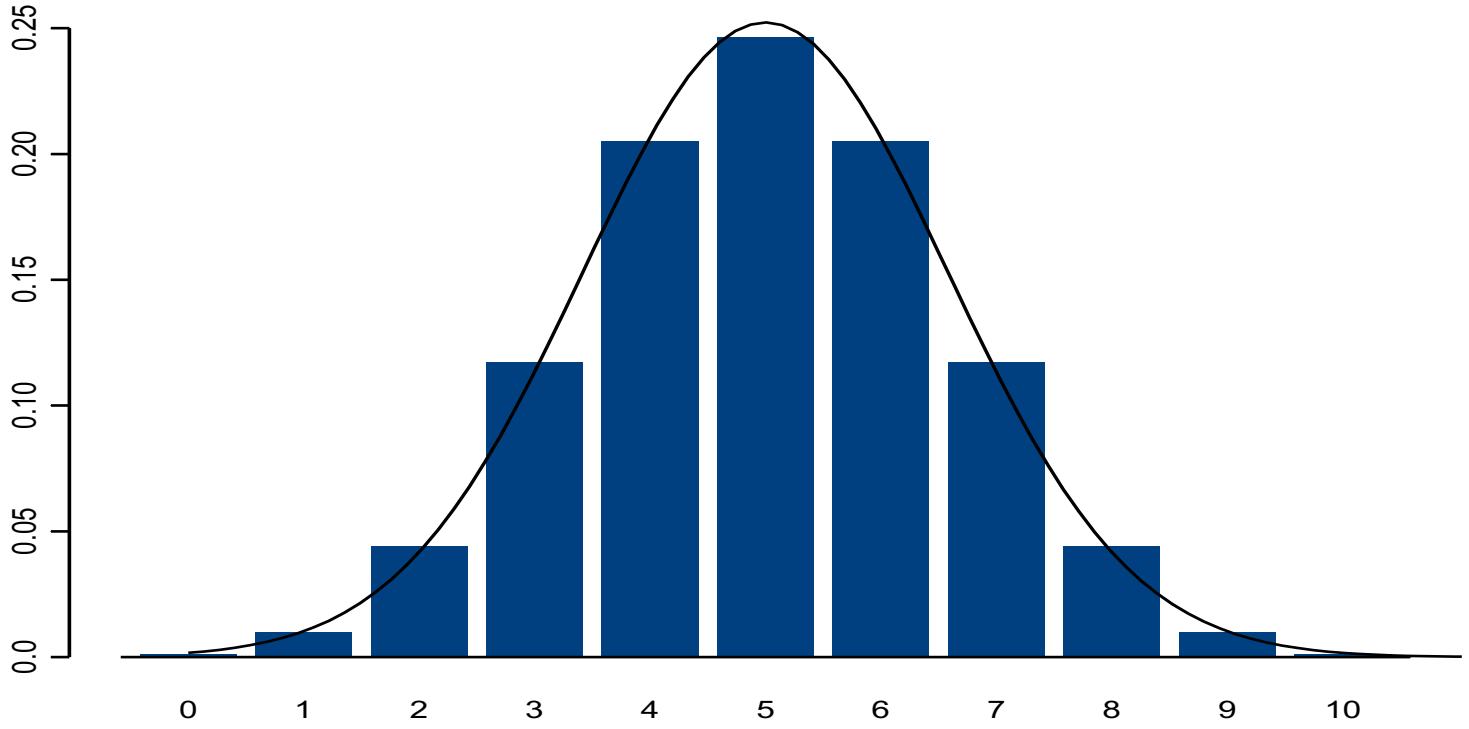
Why approximation at all?

When the approximation goes bad…



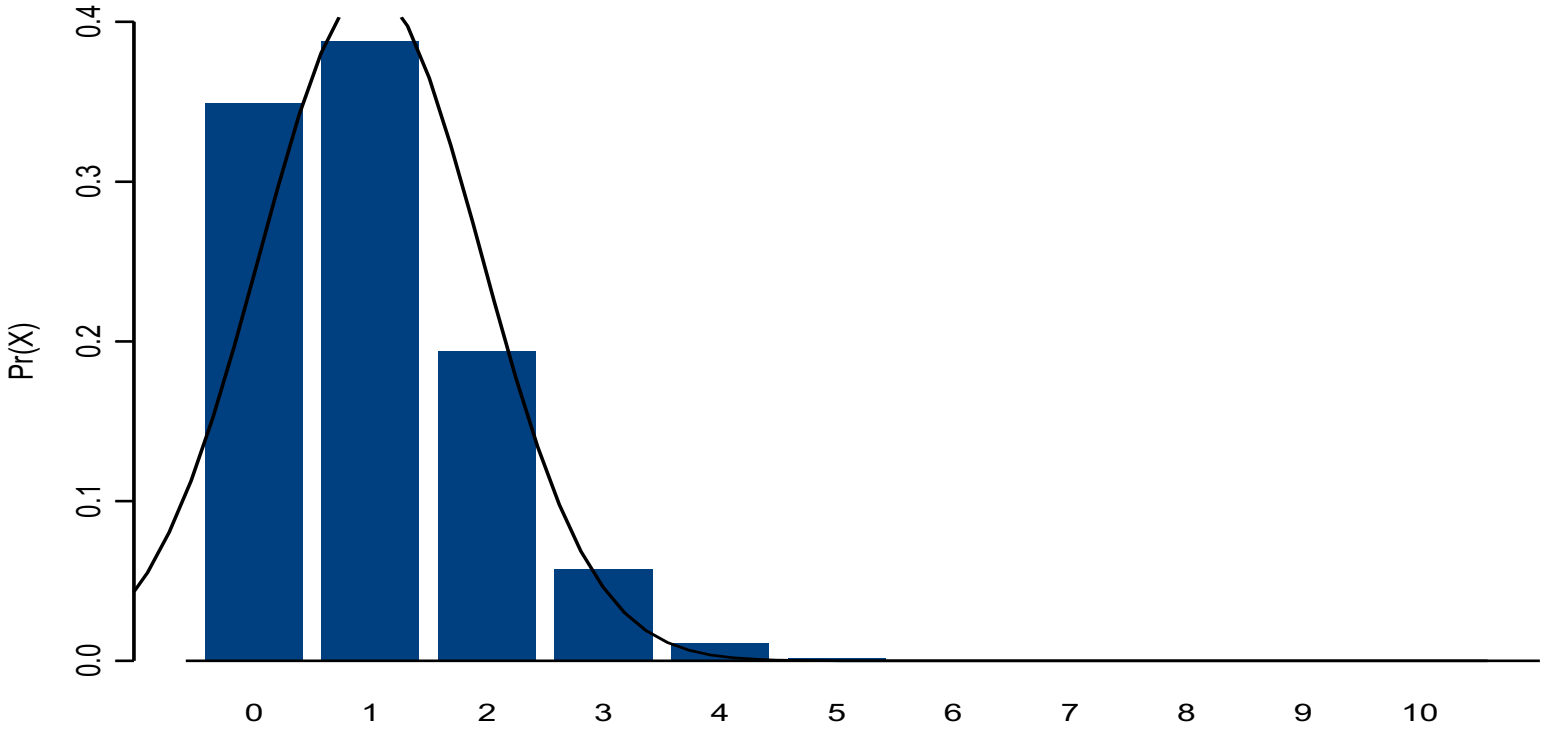






Comparison of a Binomial distribution and approximating normal distribution for n=10, p=0.5, μ=np=5, and

$$\sigma = \sqrt{npq} = 1.58$$



Comparison of a Binomial distribution and approximating normal distribution for n=10, p=0.1, μ=np=1, and $\sigma = \sqrt{npq} = 0.95$

When the approximation is good

- Rule of thumb for when approximation of Binomial by normal is “reasonably good”:

  $npq \geq 5$

  or n moderate and p away from 0 and 1.

- From the text book, the rule of thumb:

  $np > 10$ and $nq > 10$

Example:

A company produces smoke filters and knows that, on average, 10% are defective and will not pass inspection. What is the probability that *at least* 15% of a random sample of 100 filters are defective?

Let X = # defective filters in a random sample of 100.

If on average 10% of filters are defective and each filter is selected independently, then X can be assumed to follow Binomial distribution with n=100, p=0.10.

Using normal distribution approximation

- E(*X*)= *np* = 10, Var(*X*)= *npq* = 9, SD(*X*) = 3
- Expected # defectives is 10, *npq*=9
- Want probability of ***at least*** 15 defectives in 100 samples.

Use 15.0-0.5=14.5 defectives to compute probability Pr(>=15):

Find Pr(*X*≥14.5) , rather than Pr(*X*≥15)

$$\Pr(X \geq 15) \approx \Pr(X \geq 14.5) = \Pr(Z \geq \frac{14.5-10}{3}) \sim N(0,1)$$

$$\Pr(Z \geq 1.5) = 1 - \Pr(Z < 1.5) = 0.0668$$

Thus, the probability that the number of defective filters is approximately 0.067 or 6.7%.

连续性校正**(continuity correction)**

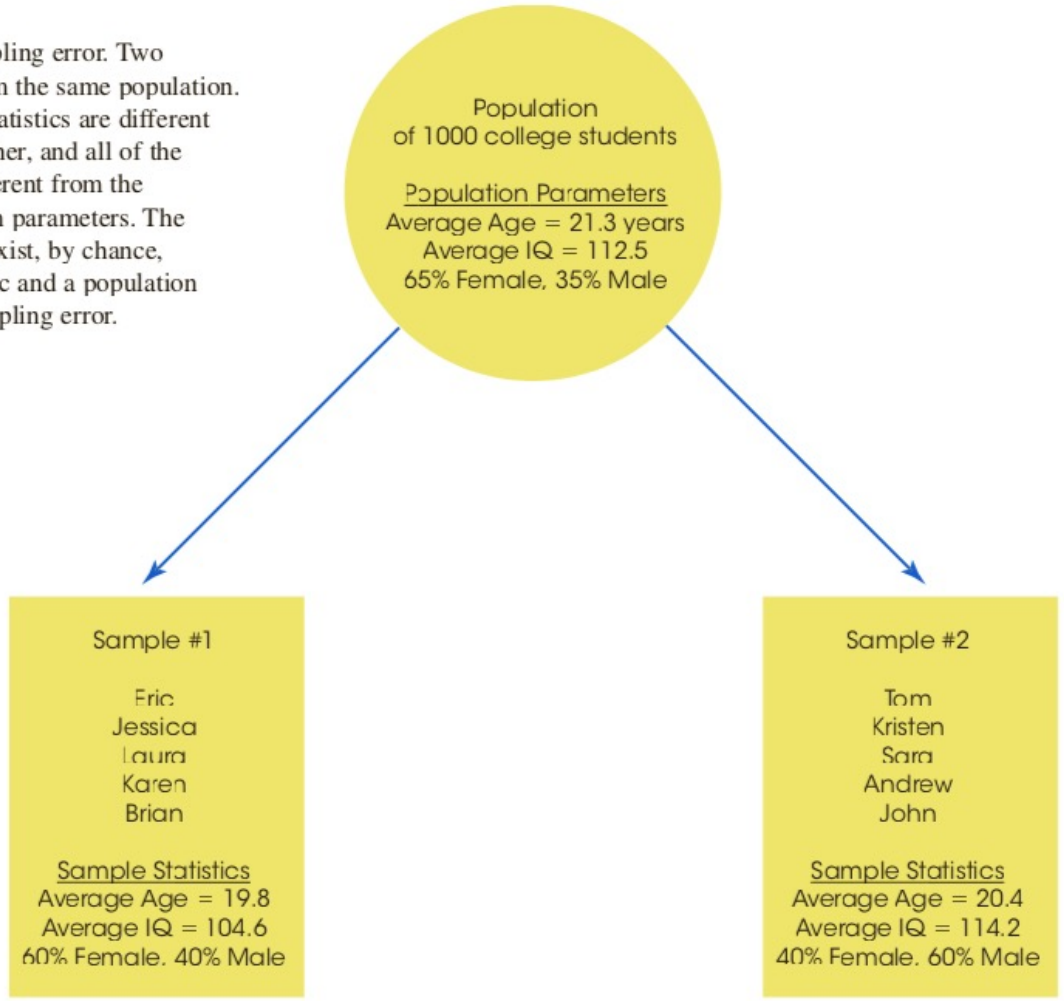
If X~ Binomial(n,p), then Pr(a=<X=<b) is approximated by the area under an N(np, npq) curve from (a-0.5) to (b+0.5).

The Distribution of Sample Means

Recap

- **Sampling error** is the naturally occurring discrepancy, or error, that exists between a sample statistic and the corresponding population parameter.
(a fundamental problem for inferential statistics)

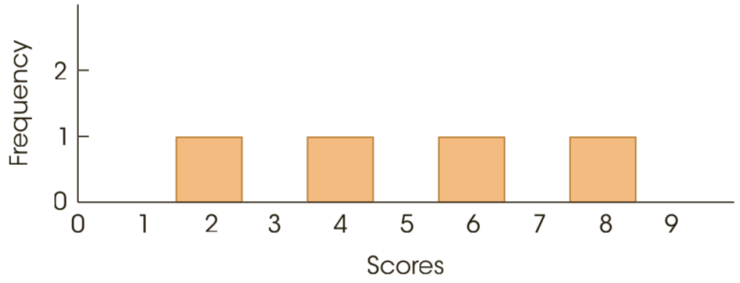
FIGURE 1.2
A demonstration of sampling error. Two samples are selected from the same population. Notice that the sample statistics are different from one sample to another, and all of the sample statistics are different from the corresponding population parameters. The natural differences that exist, by chance, between a sample statistic and a population parameter are called sampling error.



The Distribution of Sample Means

- The **distribution of sample means** is defined as the set of means from *all the possible random samples of a specific size* (*n*) selected from a specific population.
- A **sampling distribution** is a distribution of statistics obtained by selecting *all the possible samples of a specific size (n)* from a population.
- The distribution of sample means is an example of a sampling distribution, often called the sampling distribution of *M*.

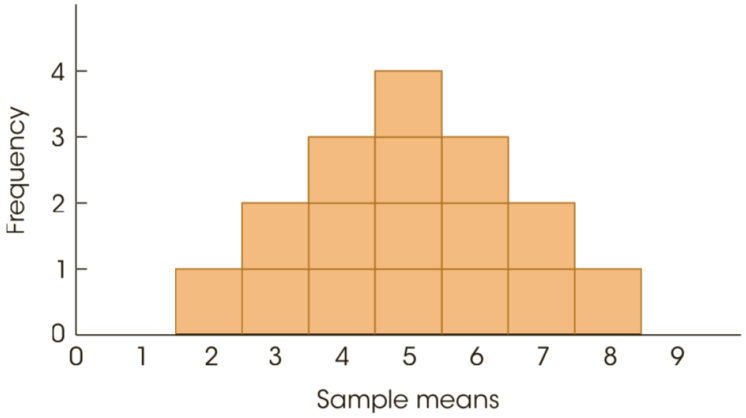
Example: Consider a population that consists of only four scores: 2, 4, 6, 8.



The complete set of possible samples of $n = 2$ scores that can be obtained from the population:

Sample	Scores		Sample Mean (M)
	First	Second	
1	2	2	2
2	2	4	3
3	2	6	4
4	2	8	5
5	4	2	3
6	4	4	4
7	4	6	5
8	4	8	6
9	6	2	4
10	6	4	5
11	6	6	6
12	6	8	7
13	8	2	5
14	8	4	6
15	8	6	7
16	8	8	8

The distribution of sample means for $n = 2$:



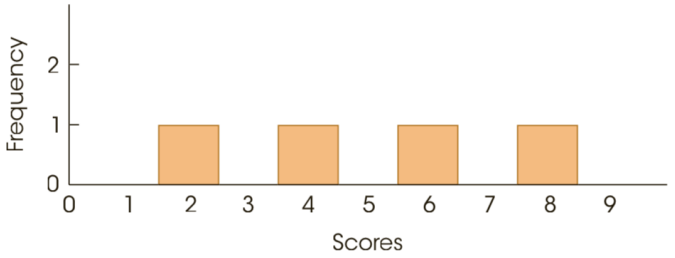
Three Characteristics of the Distribution of Sample Means

1. The sample means should pile up around the population mean.

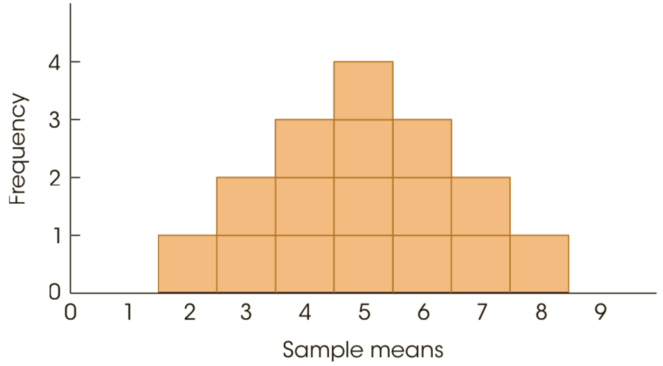
2. The pile of sample means should tend to form a normal-shaped distribution. See the **Central Limit Theorem**.

3. In general, the larger the sample size *n*, the closer the sample means should be to the population mean, μ. See the **Law of Large Numbers**.

The population distribution:



The distribution of sample means for $n = 2$:

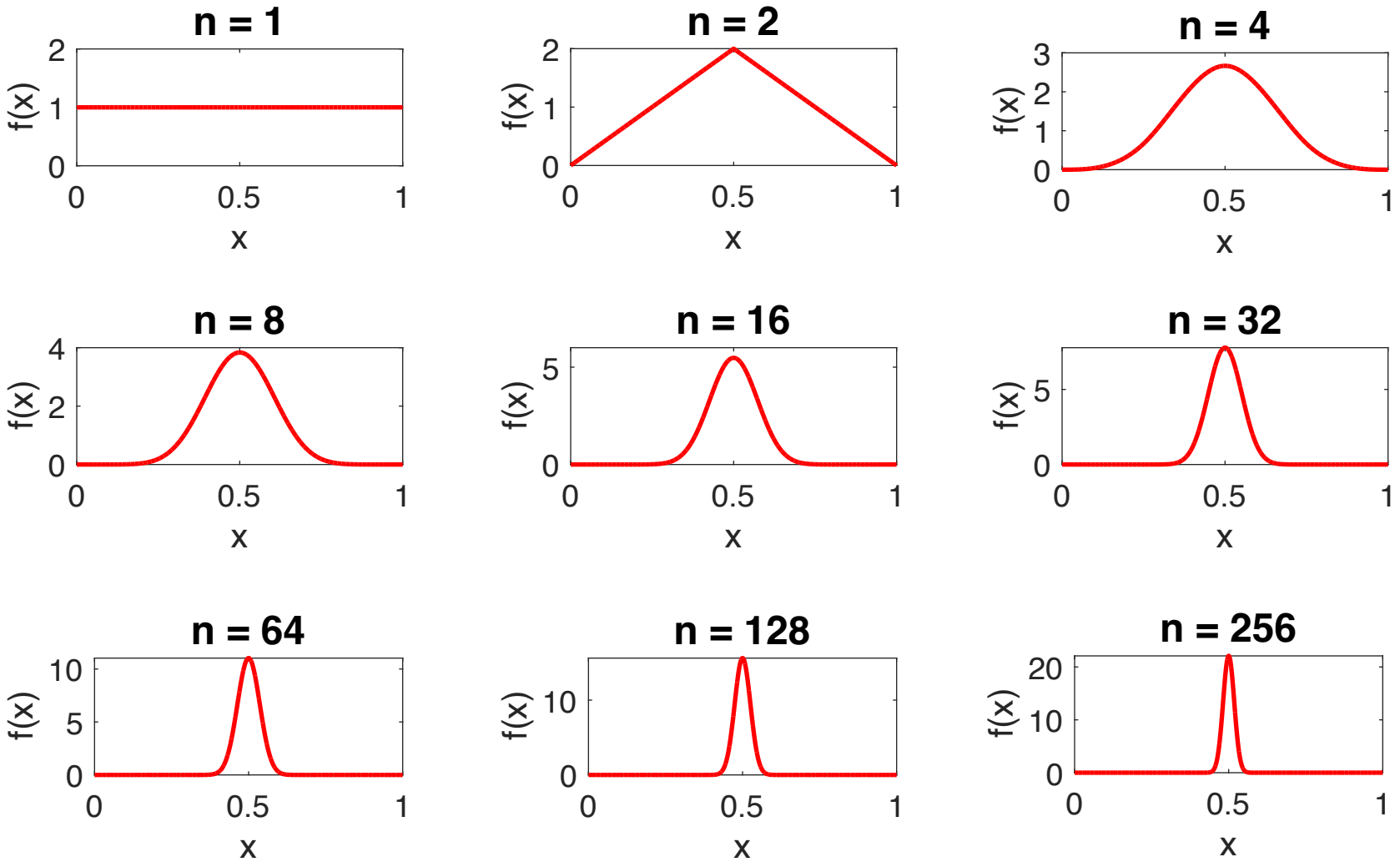


The Central Limit Theorem

(a cornerstone of inferential statistics)

- **Central limit theorem**: For any population with mean μ and standard deviation σ, the distribution of sample means for sample size n will have a mean of μ and a standard deviation of $\sigma/\sqrt{n}$ and will approach a normal distribution as n approaches infinity.

Illustration of the **central limit theorem**: the distribution of sample means for a uniform distribution



The distribution of sample means is perfectly or almost perfectly normal when

Either of the following two conditions is satisfied:

1. The population from which the samples are selected is a normal distribution.

2. The number of scores (*n*) in each sample is relatively large, around 30 or more.

The Mean of the Distribution of Sample Means: The Expected Value of *M*

- The mean of the distribution of sample means is called the **Expected Value of *M***, and is equal to the population mean μ.

Suppose $E(X_i) = \mu, \quad i = 1,2,...,n.$

$$E(M) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n\mu}{n} = \mu$$

The Standard Error of *M*

- The standard deviation of the distribution of sample means, σ_M, is called the **standard error of *M***.
- The standard error provides a measure of how much distance is expected on average between a sample mean (*M*) and the population mean (μ).

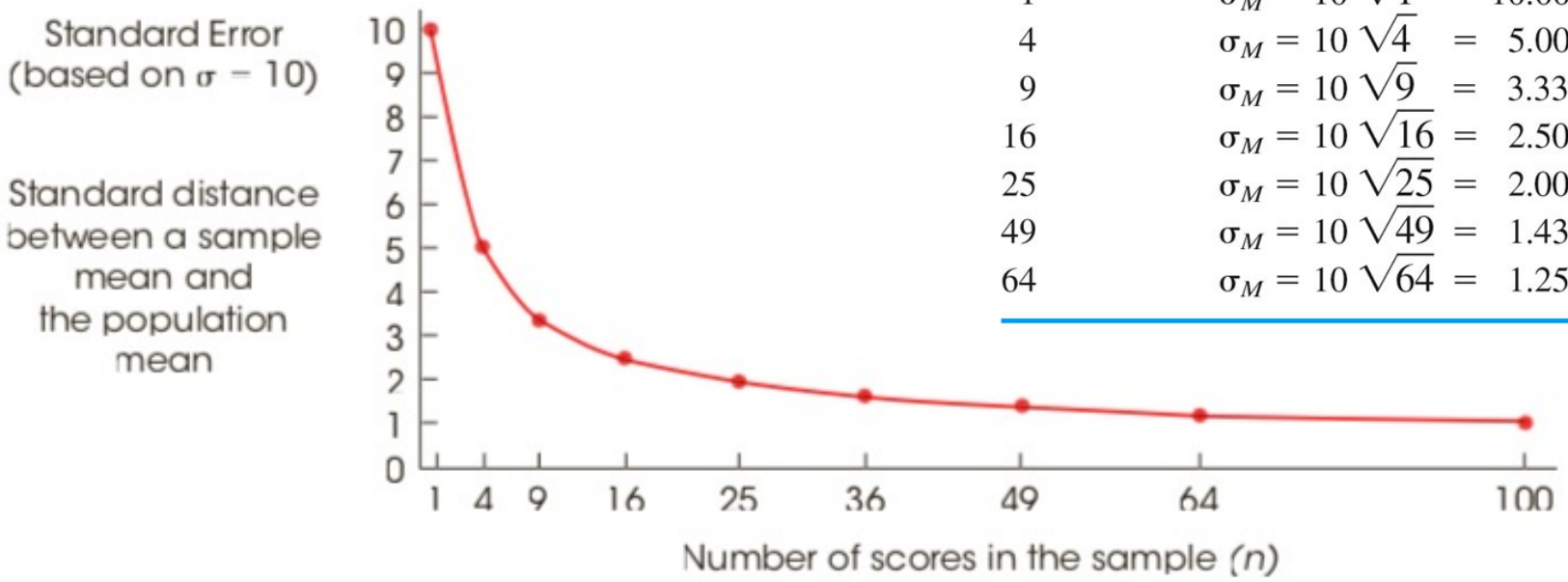
Suppose $$Var(X_i) = \sigma, \quad i = 1,2,...,n.$$

$$\sigma_M^2 = Var(M) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

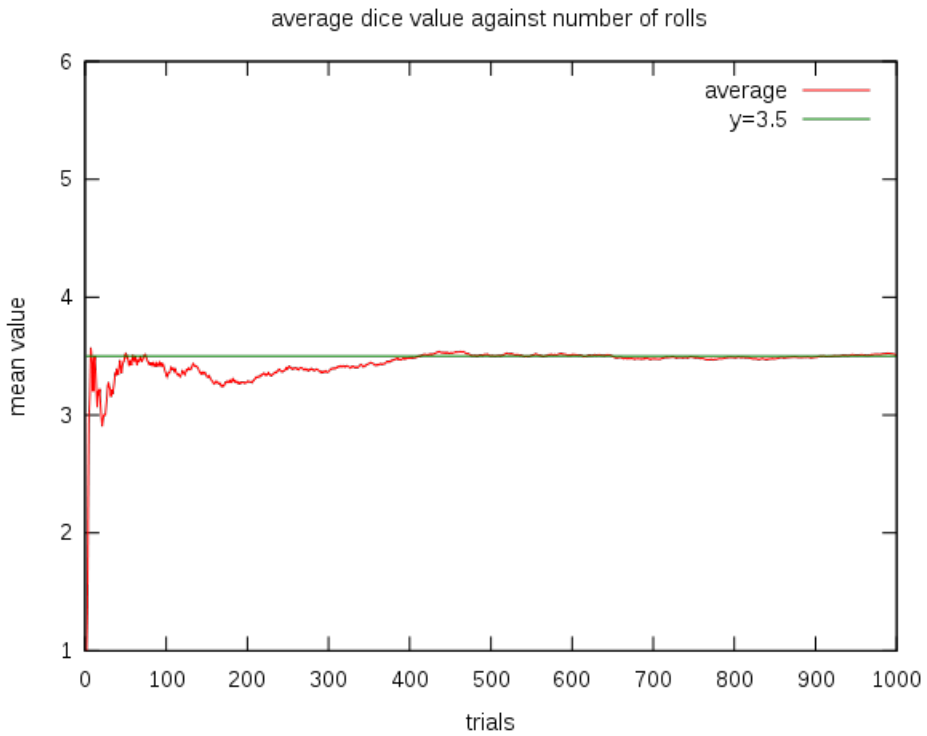
According to $\sigma_M = \sigma / \sqrt{n}$, the magnitude of the standard error is determined by two factors:
(1) the size of the sample (*n*) and (2) the standard deviation of the population from which the sample is selected (σ).

Sample size (*n*)	Standard error
1	$\sigma_M = 10/\sqrt{1} = 10.00$
4	$\sigma_M = 10/\sqrt{4} = 5.00$
9	$\sigma_M = 10/\sqrt{9} = 3.33$
16	$\sigma_M = 10/\sqrt{16} = 2.50$
25	$\sigma_M = 10/\sqrt{25} = 2.00$
49	$\sigma_M = 10/\sqrt{49} = 1.43$
64	$\sigma_M = 10/\sqrt{64} = 1.25$



The Law of Large Numbers (LLN)

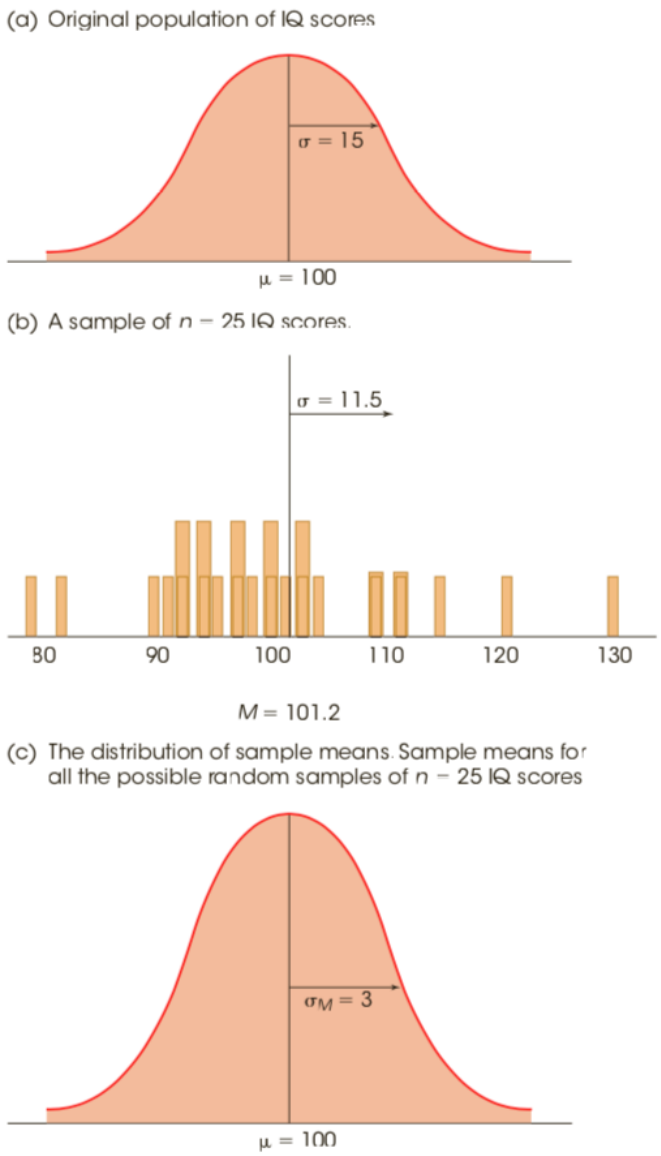
- The **law of large numbers** states that the larger the sample size (*n*), the more probable it is that the sample mean will be close to the population mean.



That is why psychological experiments often need many trials

Three Different Distributions

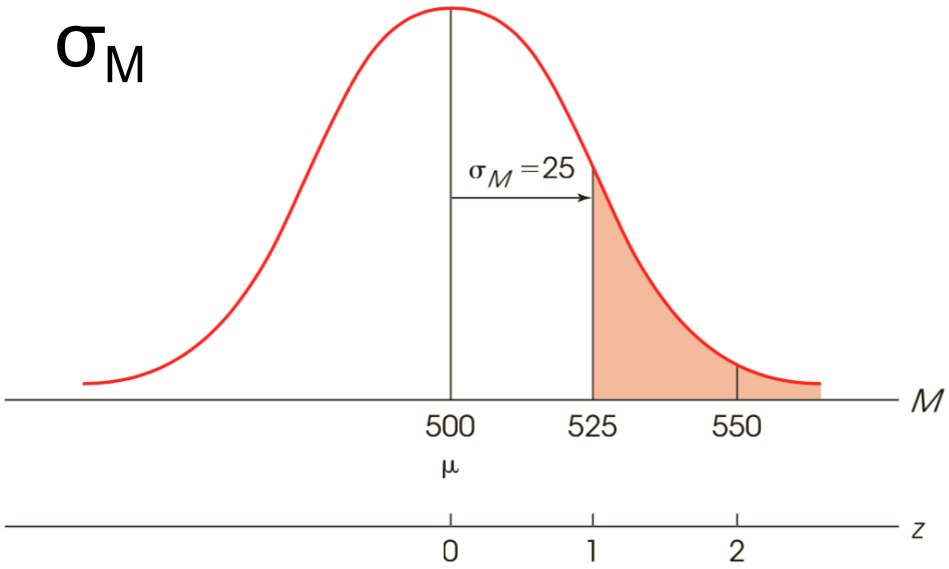
- Population distribution
- Sample distribution
- Distribution of sample means (sampling distribution)



z-Scores for Sample Means

- Within the distribution of sample means, the location of each sample mean can be specified by a z-score,

$$z = \frac{M - \mu}{\sigma_M}$$



Example: A sample of $n = 4$ scores is selected from a normal distribution with a mean of $\mu = 40$ and a standard deviation of $\sigma = 16$.

(1) Find the z-score for a sample mean of $M = 42$.

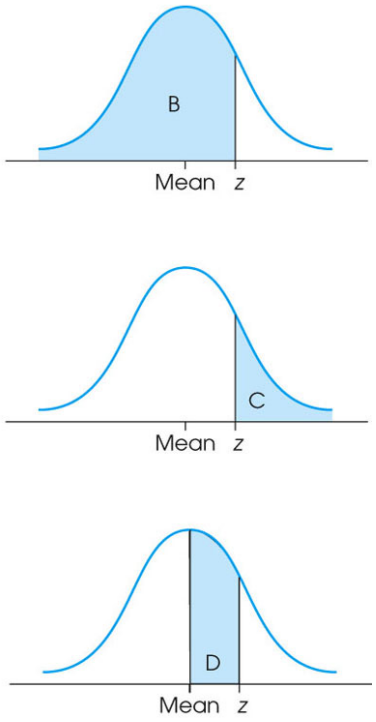
(2) Determine the probability of obtaining a sample mean larger than $M = 42$.

(1) $\sigma_M = \sigma / \sqrt{n} = 16 / \sqrt{4} = 8$

$$z = \frac{M - \mu}{\sigma_M} = \frac{42 - 40}{8} = 0.25$$

(2) $p = 0.4013$

(A) z	(B) Proportion in body	(C) Proportion in tail	(D) Proportion between mean and z
0.00	.5000	.5000	.0000
0.01	.5040	.4960	.0040
0.02	.5080	.4920	.0080
0.03	.5120	.4880	.0120
0.21	.5832	.4168	.0832
0.22	.5871	.4129	.0871
0.23	.5910	.4090	.0910
0.24	.5948	.4052	.0948
0.25	.5987	.4013	.0987
0.26	.6026	.3974	.1026
0.27	.6064	.3936	.1064
0.28	.6103	.3897	.1103
0.29	.6141	.3859	.1141
0.30	.6179	.3821	.1179
0.31	.6217	.3783	.1217
0.32	.6255	.3745	.1255
0.33	.6293	.3707	.1293
0.34	.6331	.3669	.1331

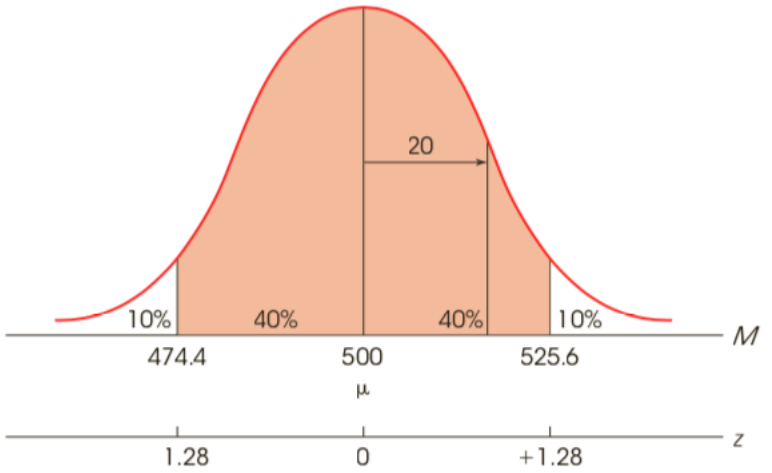


Example: The distribution of SAT scores forms a normal distribution with a mean of $\mu = 500$ and a standard deviation of $\sigma = 100$. What is the exact range of values that includes the middle 80% of the distribution of sample means for $n = 25$?

$$\sigma_M = \sigma / \sqrt{n} = 100 / \sqrt{25} = 20$$

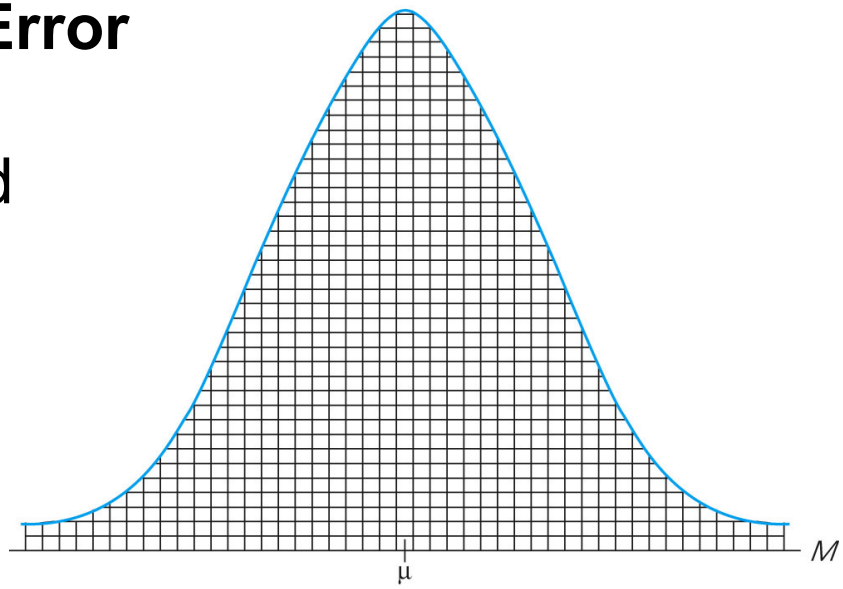
The middle 80% of the z distribution corresponds to $z_{0.1} = -1.28$ and $z_{0.9} = +1.28$. The 80% range of the distribution of sample means for $n = 25$ is

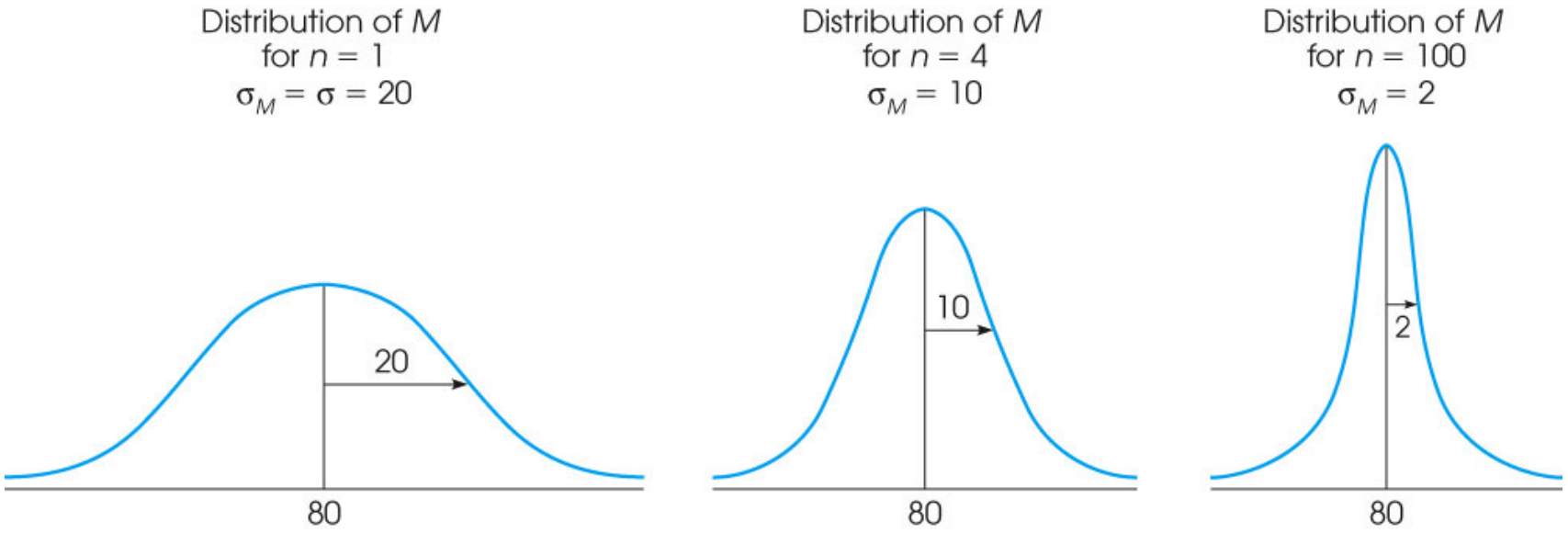
$$\begin{aligned} &\left[z_{0.1}\sigma_M + \mu, z_{0.9}\sigma_M + \mu\right] \\ &= \left[-1.28 \times 20 + 500, 1.28 \times 20 + 500\right] \\ &= \left[474.4, 525.6\right] \end{aligned}$$



Sampling Error and Standard Error

(Distinguishing between standard deviation and standard error)





Reporting Standard Error (SE)

- SE or SEM (for standard error of the mean) is frequently used to denote standard error.
- M±SE is often reported in the text, in tables, or in graphs.

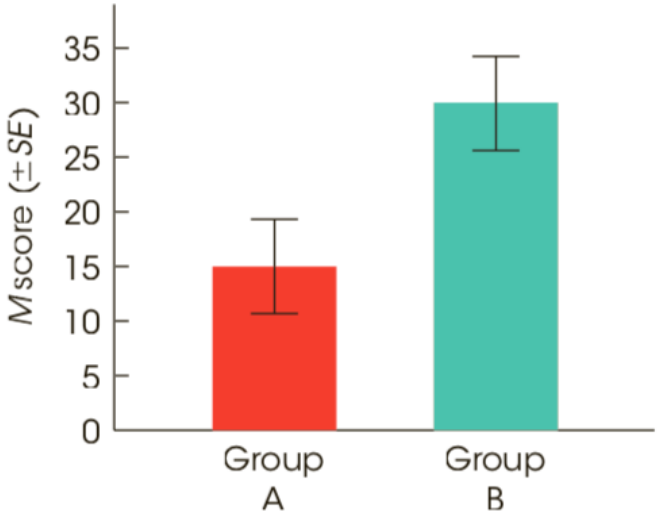
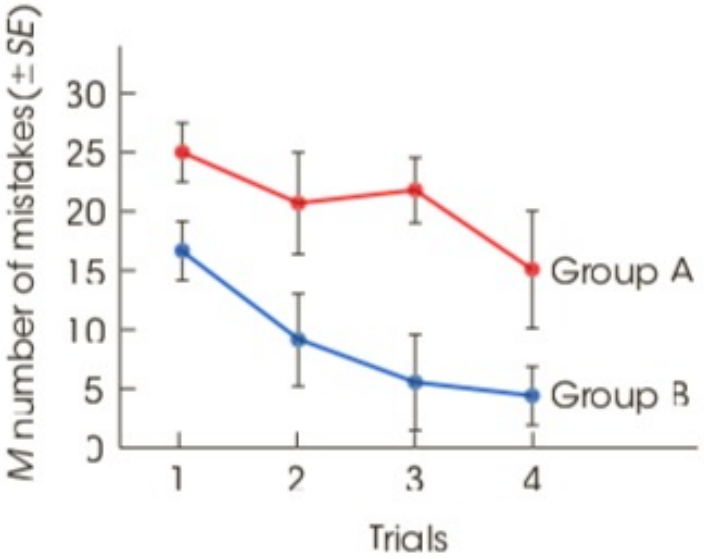


TABLE 7.3

The mean self-consciousness scores for participants who were working in front of a video camera and those who were not (controls).

	n	Mean	*SE*
Control	17	32.23	2.31
Camera	15	45.17	2.78



Summary

- Continuous distribution: PDF and CDF
- Distributions: exponential, Poisson, Gaussian
- Gaussian approximation of a binomial distribution
- Central limit theorem
- Law of large number
- Distribution of sample means: SEM vs SD