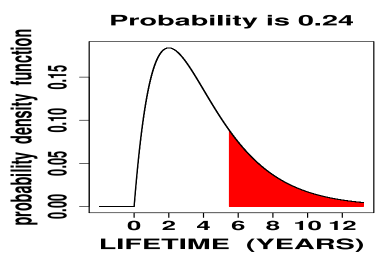
6 The Normal Distribution

Continuous distributions and density curves

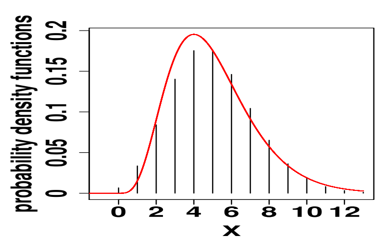
Rules for a continuous histogram.

1. The area of a histogram is 1.
2. The **probability** of the random variable taking a value in the interval from "a" to "b" is the **area** under the probability distribution curve within this interval.
3. The probability density function is nonnegative (cannot have negative probability).

Example: Let X be the lifetime of a computer CPU in years, as shown in the graph below. Determine the probability that a new CPU lasts at least 5.5 years.



□



Hence, in these above graphs of probability density functions, the *relative frequency* is represented by the **area** under the curve, NOT the **height** of the curve.

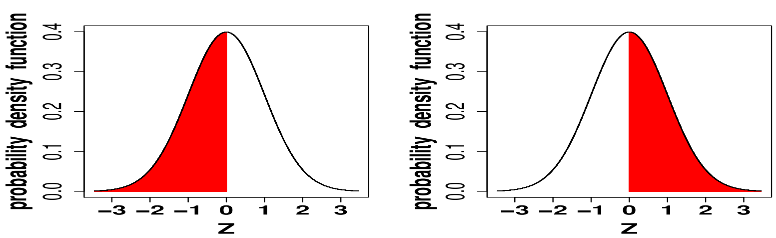
What is the total area under the density curve?

6.1 The Normal Curve

The normal distribution is bell-shaped and symmetric.

Notation: $Z \sim N(\mu_z = 0,\ \sigma_z = 1)$.

Example: Compute $P(Z < 0)$, $P(Z \leq 0)$, $P(Z > 0)$, and $P(Z \geq 0)$.



□

Example: *Using the standard normal table.* Let Z be a standard normal random variable.

(a) Determine $P(Z < 1.26)$.

(b) Determine $P(Z > 1.26)$.

(c) Determine $P(Z < -1.26)$.

(d) Determine $P(Z > -1.26)$.

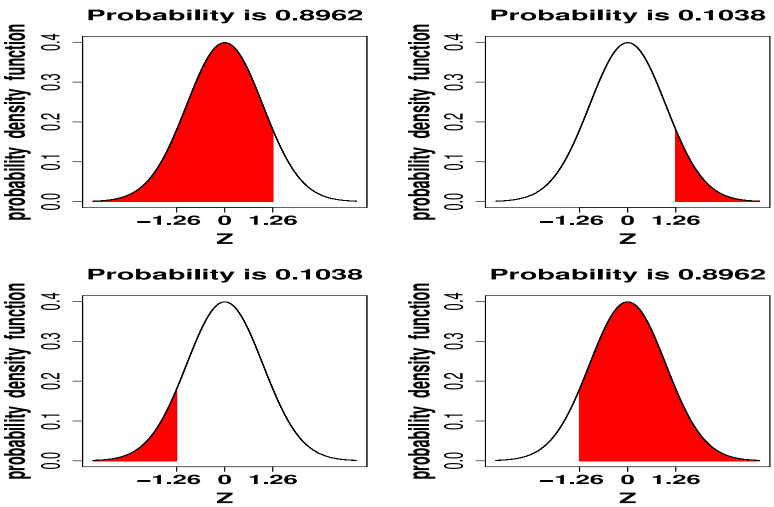
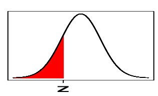
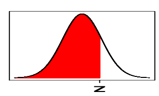


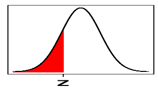
Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7

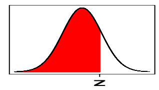




z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7





z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

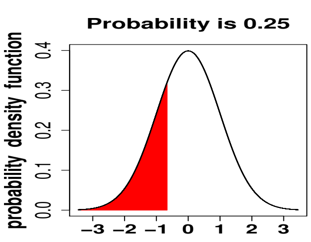
□

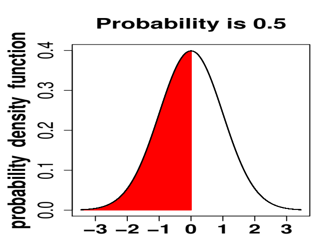
Example: *Using the standard normal table in reverse.* Let Z be a standard normal random variable.

(a) Determine the 25th percentile of Z.

(b) Determine the 50th percentile of Z.

(c) Determine the 75th percentile of Z.





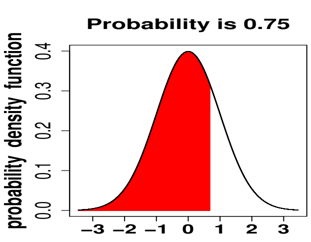
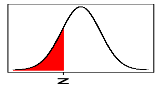
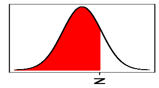
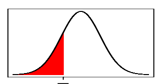
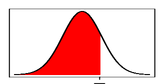


Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7

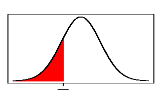
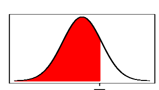
z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

□

Finding Probabilities for Bell-Shaped Distributions

Here, we focus on the **normal distribution**, which exists in many applications (at least approximately).

Recall the **empirical rule** from section 3.2.

Empirical Rule

If a large number of observations are sampled from an approximately normal distribution, then (usually)

1. Approximately 68% of the observations fall within **one** standard deviation, σ, of the mean, μ.
2. Approximately 95% of the observations fall within **two** standard deviations, σ, of the mean, μ.
3. Approximately 99.7% of the observations fall within **three** standard deviations, σ, of the mean, μ.

Suppose X has a normal distribution with mean μ and standard deviation σ.

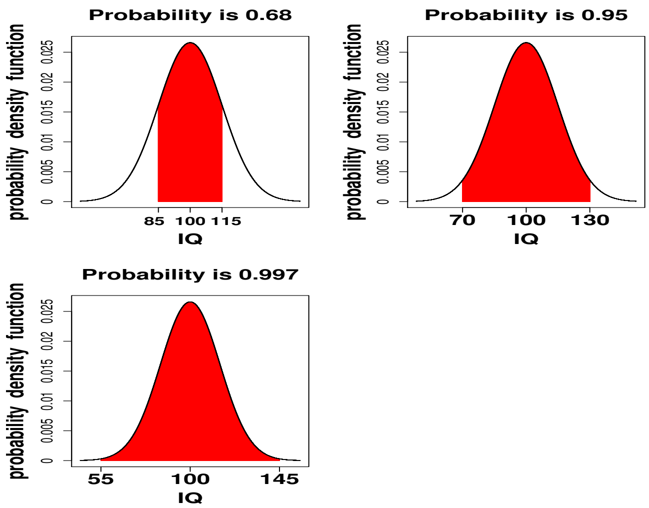
Notation: $X \sim N(\mu, \sigma)$

$$P(\mu - \sigma < X < \mu + \sigma) = 0.68$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$$

Example: IQ scores of normal adults on the Weschler test have a symmetric bell-shaped distribution with a mean of 100 and standard deviation of 15.



□

Again, consider $X \sim N(\mu, \sigma)$.

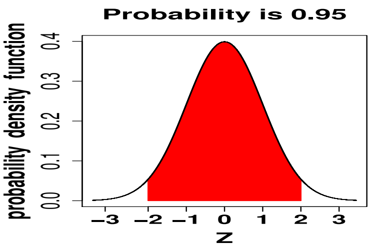
$$Z = \frac{X - \mu}{\sigma}$$

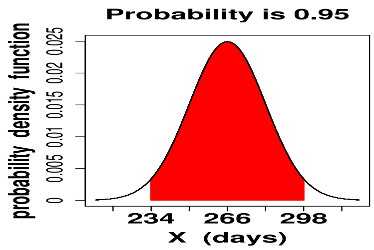
Reverse table look-up uses $X = \mu + \sigma Z$

$$X \leftrightarrow Z \leftrightarrow \text{probability}$$

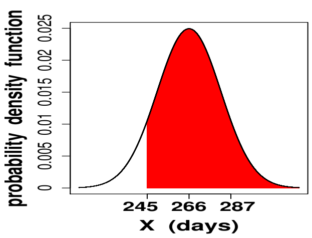
Example: The length of human pregnancies from conception to birth varies according to a distribution which is approximately normal with mean 266 days and standard deviation 16 days.

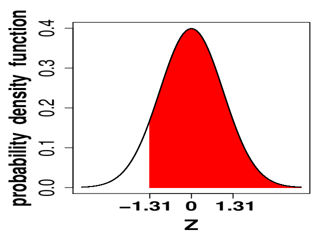
(a) Show the empirical rule regarding 95%.





(b) What proportion of pregnancies last more than 245 days?





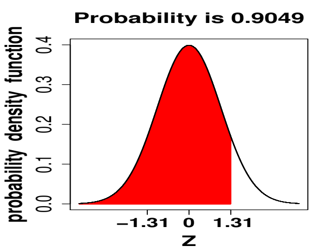
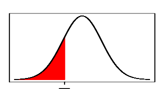
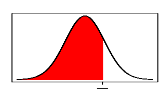


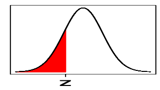
Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7

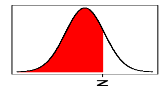




z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7





z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(c) What proportion of pregnancies last between 245 and 285 days?

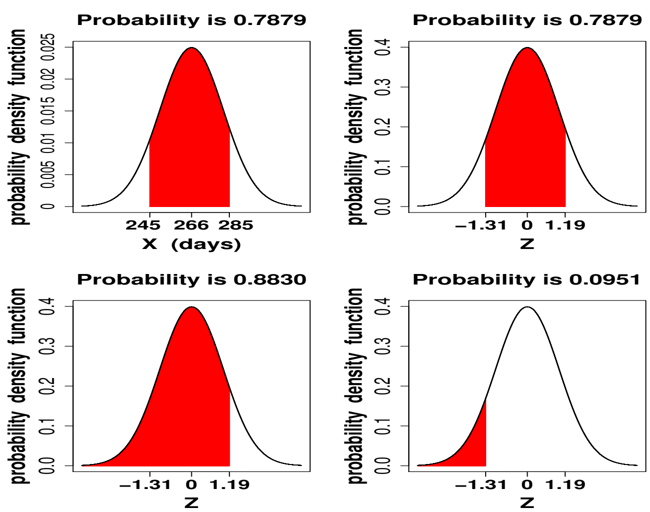
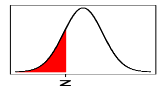
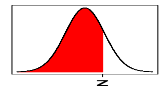


Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(d) How long do the longest 20% of pregnancies last?

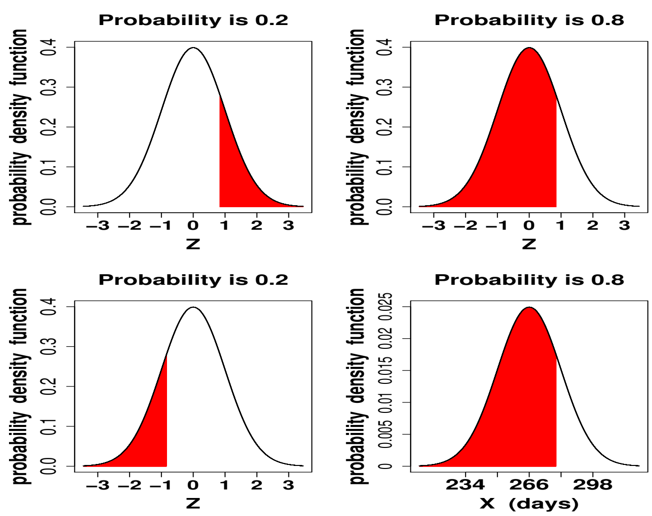
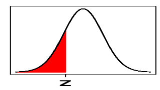
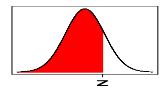
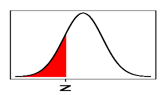
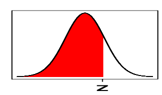


Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7

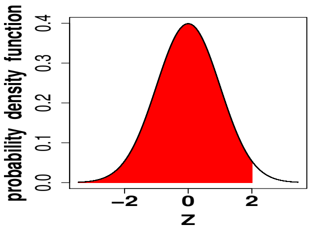
z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

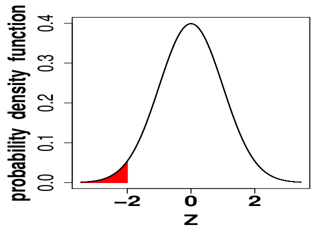
Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Example: Let $X \sim N(\mu, \sigma)$. Using the standard normal table, verify the empirical rule regarding 95%. In other words, compute $P(\mu - 2\sigma < X < \mu + 2\sigma)$ to four significant digits.





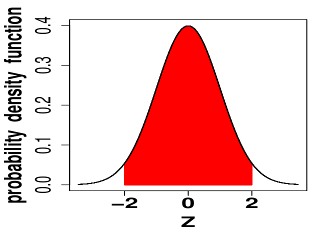
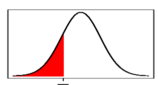
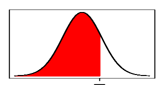


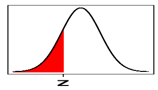
Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7

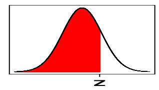




z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7





z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Example: (off the charts)

(a) Determine $P(Z < -4)$

(b) Determine $P(Z > -4)$

(c) Determine $P(Z < 6)$

(d) Determine $P(Z < -6)$

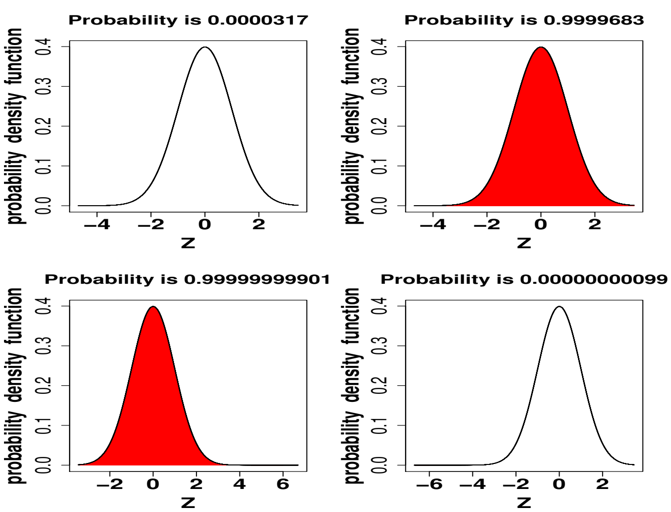
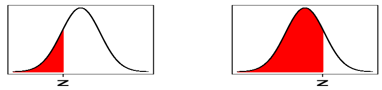
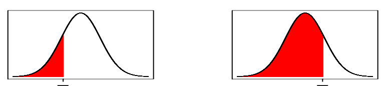


Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7



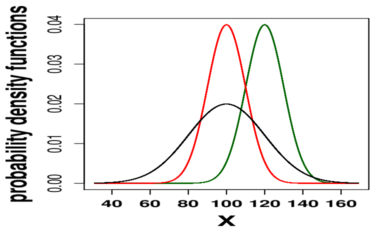
z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
−3.7 or less	.0001									
−3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
−3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7



z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7 or more	.9999									

Example: Compare means and standard deviations in the graphs below.



6.2 Sampling Distributions and the Central Limit Theorem

The Sampling Distribution (of a Statistic)

Definition: (Recall) A **statistic** is a quantity computed from a sample.

Example:

Recall from section 5.1:

Definition: The **probability distribution** of a **discrete** random variable X consists of the possible values of X along with their associated probabilities.

Definition: The *probability distribution* of a **statistic** is called its **sampling distribution**.

Hence, the **sampling distribution** of a **discrete** *statistic* consists of the possible values of the *statistic* along with their associated probabilities.

Example: Consider a population consisting of three cards, which are labeled as $\boxed{2}$, $\boxed{3}$, and $\boxed{4}$. Let x be the value of a card drawn.

(a) Determine the **probability distribution** of X.

(b) Graph the *probability distribution* of X.

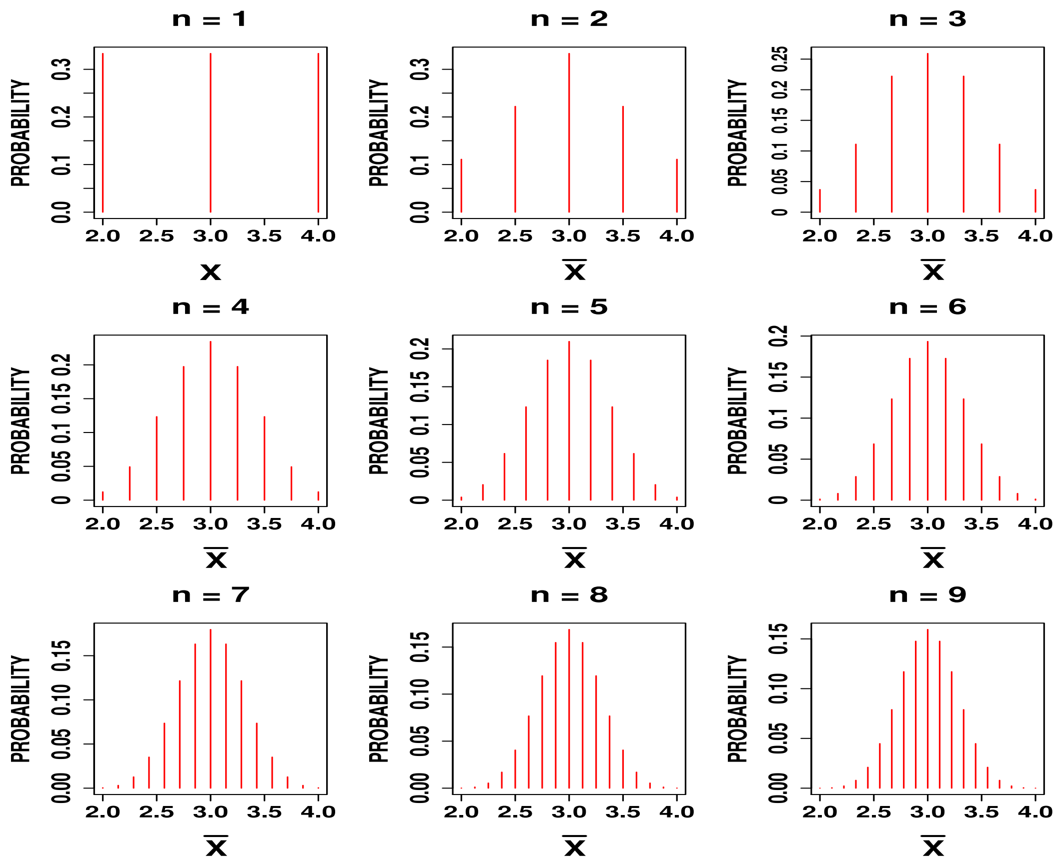
(c) Determine the *mean* of X.

(d) Let $\bar{X}$ be the sample mean, based on **two** observations independently sampled (i.e., **with** replacement) from this population. Determine the **sampling distribution** of $\bar{X}$.

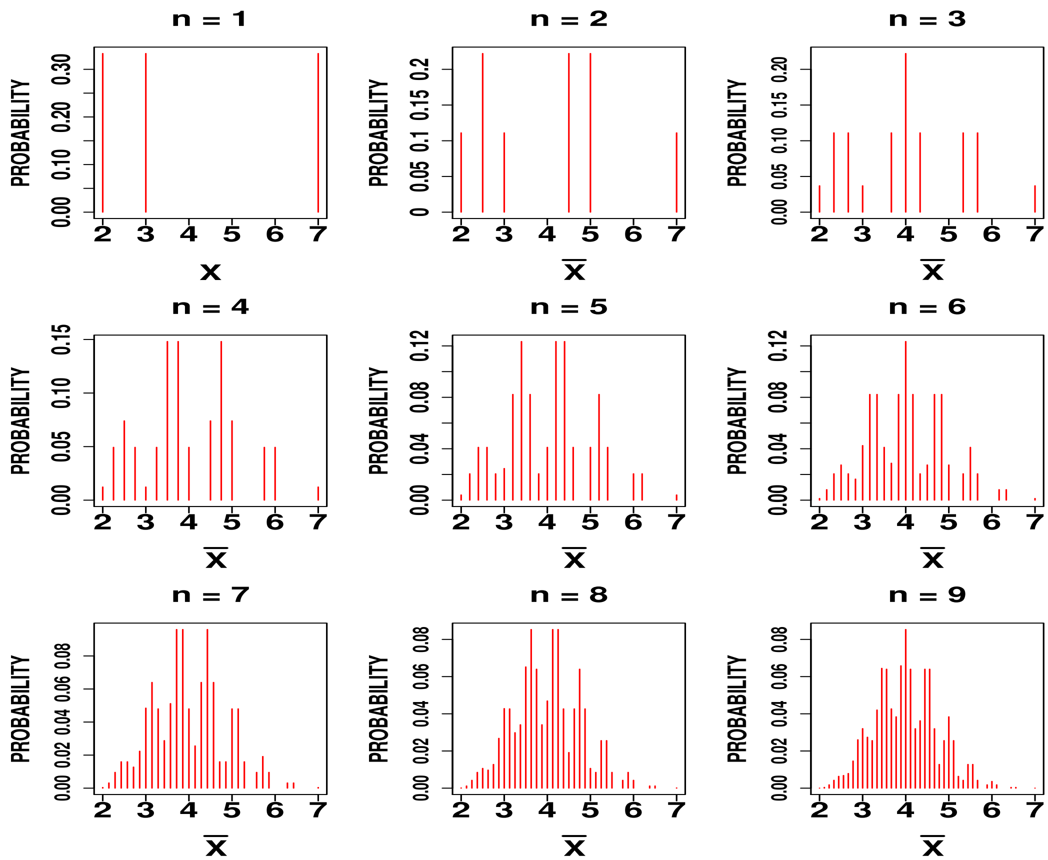
(e) Graph the *sampling distribution* of $\bar{X}$.

(f) Determine the *mean* of $\bar{X}$.

(g) Additional graphs of the *sampling distribution* of $\bar{X}$ are below, based on independent observations and sample size n.



(h) Repeat part (g), using cards labeled $\boxed{2}$, $\boxed{3}$, and $\boxed{7}$.



□

Case A: Sample **with** replacement. *Hence, observations are independent.*

Case B: Sample **without** replacement, but the population size is quite large compared to n; i.e., $N \geq 20n$. *Hence, observations are nearly independent.*

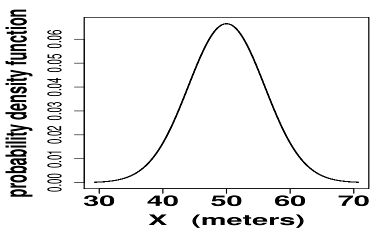
(a) $\mu_{\bar{X}} = \mu$ *always.*

(b) $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ (called the **standard error** of $\bar{X}$), exactly for Case A and approximately for Case B.

(c) (A version of the Central Limit Theorem) The sample mean, $\bar{X}$, is approximately normally distributed for Cases A and B (and positive finite σ), for **large** n (usually $n > 30$, if neither tail of the distribution is too heavy).

(d) (A special case) The sample mean, $\bar{X}$, is approximately normally distributed for Cases A and B (and positive finite σ), if the **original population** is approximately **normally distributed** (for **any** sample size n).

Example: Suppose $X \sim N(\mu = 50 \text{ meters}, \sigma = 6 \text{ meters})$. Sample nine independent observations of X.



(a) Determine the *mean* of $\bar{X}$.

(b) Determine the *standard deviation* of $\bar{X}$; i.e., the *standard error* of $\bar{X}$.

(c) Determine the probability that $\bar{X}$ exceeds 51 meters.

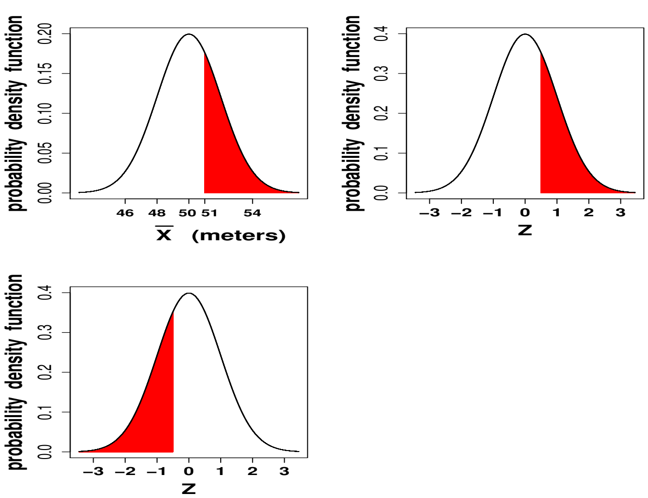
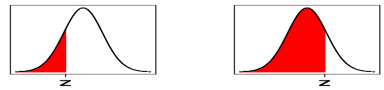


Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7



z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

□

Example: Suppose personal income, X, in a large country has mean $\mu = \$40,000$ and standard deviation $\sigma = \$30,000$. Sample **without** replacement.

(a) Determine $P(\bar{X} > \$44,000)$, for $n = 64$.

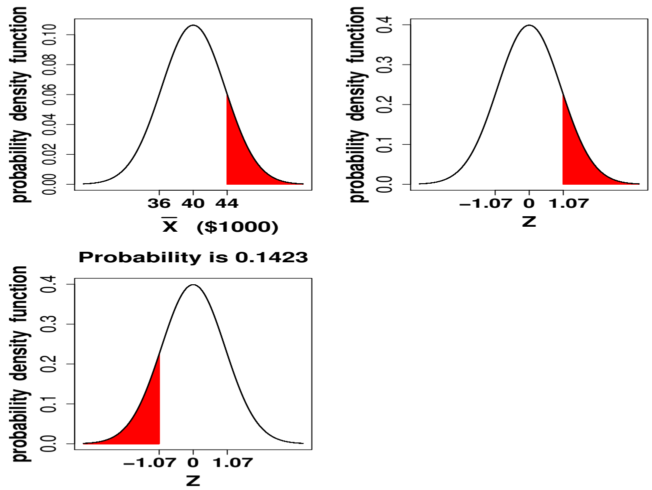
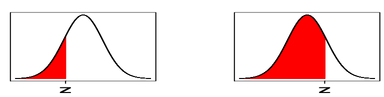


Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7



z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(b) Determine $P(\bar{X} > \$44,000)$, for $n = 100$.

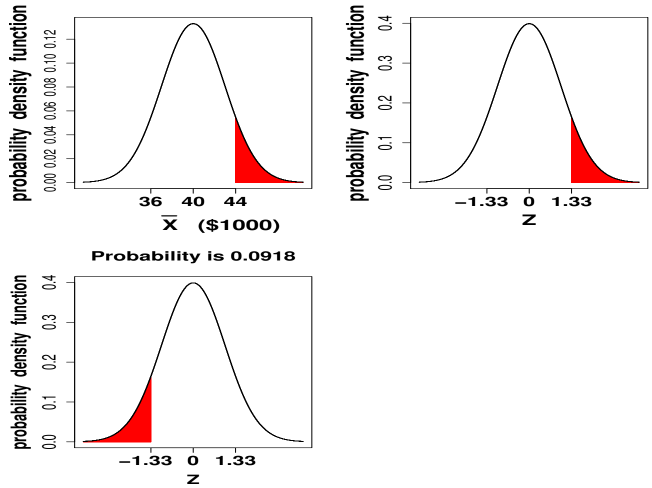
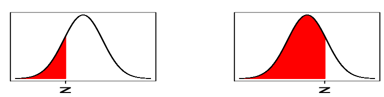
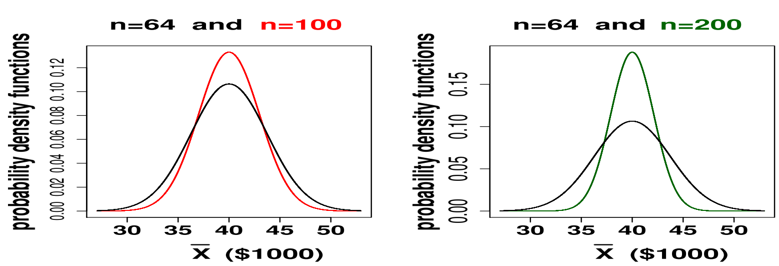


Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7



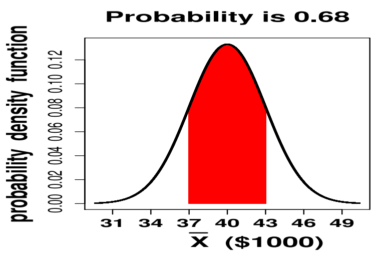
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(c) What happens to $P(\bar{X} > \$44,000)$ as we increase n to 200?

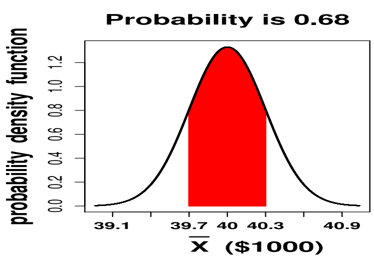


(d) Determine $P(\bar{X} > \$44,000)$, for $n = 10$.

(e) Determine the 68% part of the empirical rule for $n = 100$.



(f) Determine the 68% part of the empirical rule for $n = 10,000$.



□

6.3 The Central Limit Theorem for Proportions, $\hat{p}$

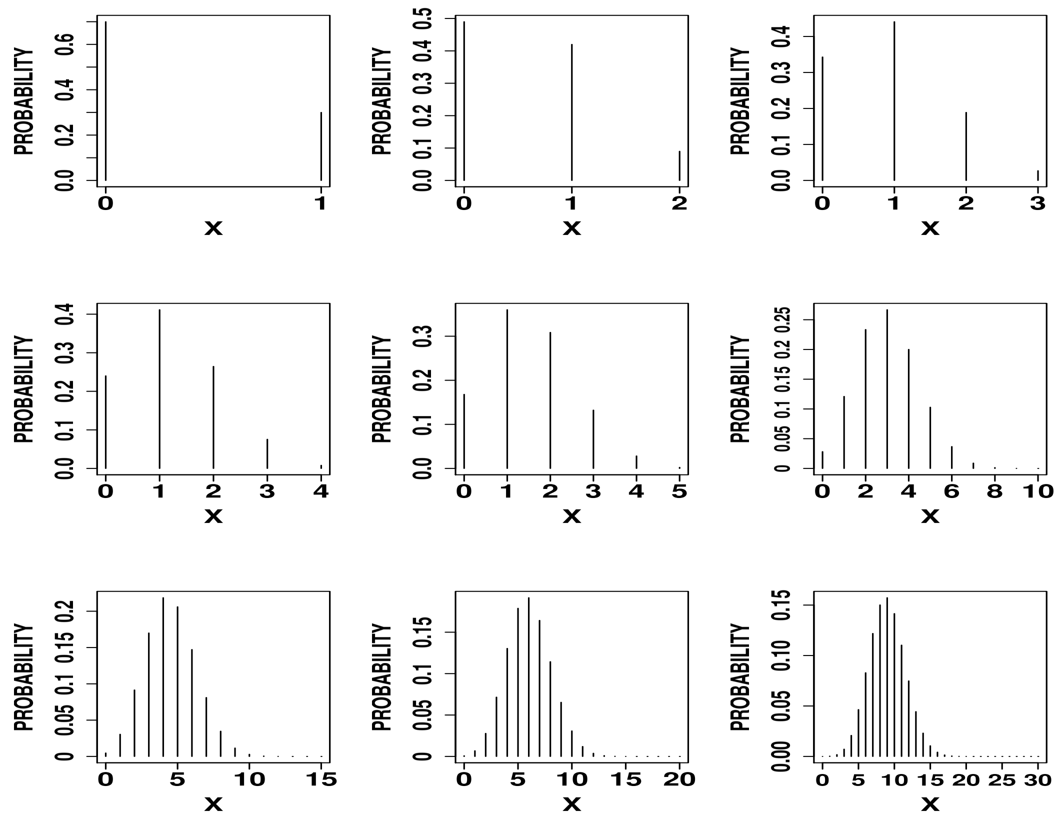
This section 6.3 includes some of the concepts from the next section 6.4, *The Normal Approximation to the Binomial Distribution.*

For large sample sizes (i.e., $np \geq 10$ and $n(1-p) \geq 10$), a *binomial random variable* and a *sample proportion* are approximately **normally distributed** by the **Central Limit Theorem**.

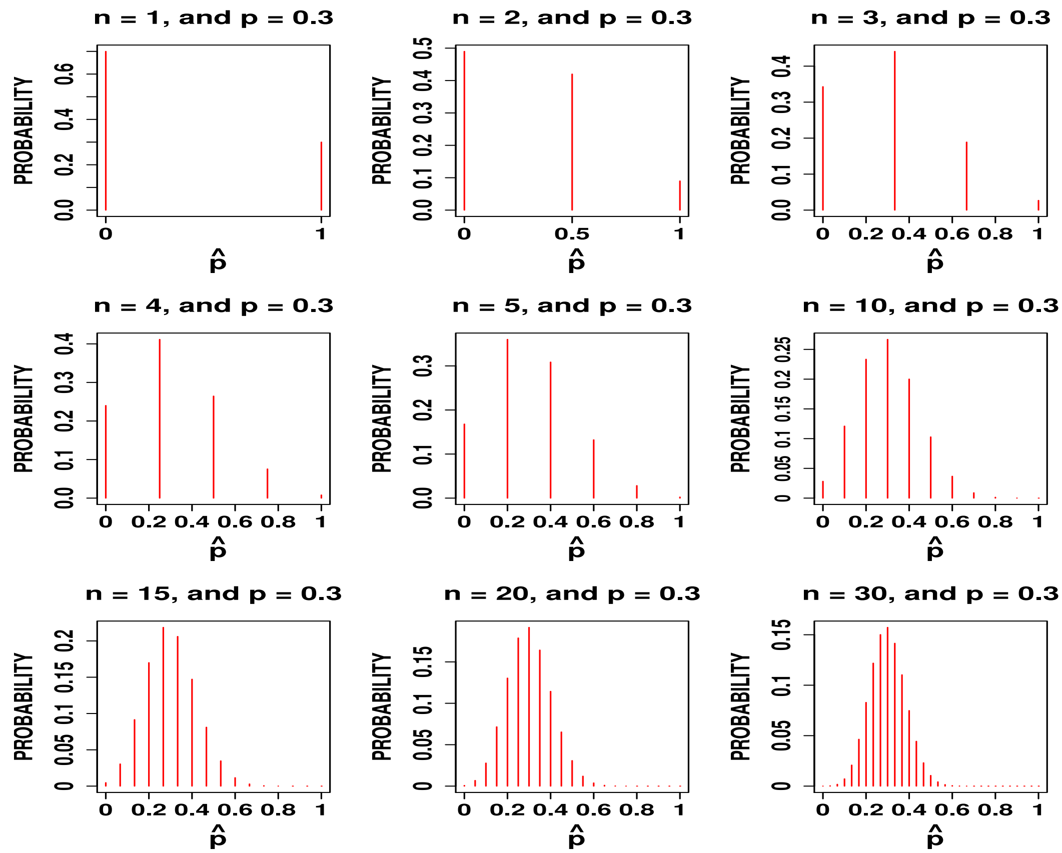
□

Example: Viewing the Central Limit Theorem.

(a) Consider the graphs below for **binomial** random variables, using $p = 0.3$ and $n = 1, 2, 3, 4, 5, 10, 15, 20$, and 30.



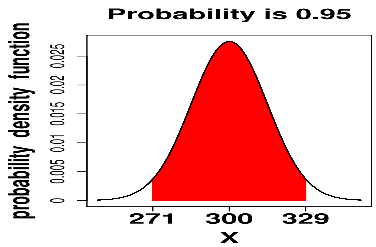
(b) Consider the graphs below for **sample proportions**, $\hat{p}$, using $p = 0.3$ and $n = 1, 2, 3, 4, 5, 10, 15, 20$, and 30.



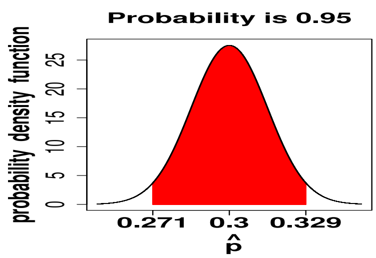
□

Example: *Revisit the Democrats.* Sample 1,000 *independent* observations from a large population which is 30% Democrat.

(a) Use the 95% part of the **empirical rule** on the *binomial random variable.*



(b) Use the 95% part of the **empirical rule** on the *sample proportion*.



□

The Sampling Distribution (of a Statistic)

Definition: (Recall) A **statistic** is a quantity computed from a sample.

Example:

The sampling distribution of a sample proportion, $\hat{p}$

Recall that a proportion is a special case of a mean.

Example: *Revisit the Democrats.* Sample *independent* observations from a large population which is 30% Democrat. Let $\hat{p}$ be the sample proportion of Democrats.

(a) State the **population distribution** in a chart, and construct the *line graph* of the **population distribution**.

Let $X = 0$ if non-Democrat, and $X = 1$ if Democrat.

Note that the *sampling distribution* of $\hat{p}$ for $n = 1$ is the same as the *population distribution* of X.

(b) For $n = 2$, state the **sampling distribution** of $\hat{p}$ in a chart, and construct the *line graph* of the **sampling distribution** of $\hat{p}$.

(c) What happens to the *sampling distribution* of $\hat{p}$ as the sample size, n, gets larger?

□

Example: *Virginians who exercise.* According to the Centers for Disease Control and Prevention, about 48% of Virginian adults achieved the recommended level of physical activity.

Recommended physical activity is defined as "reported moderate-intensity activities (i.e., brisk walking, bicycling, vacuuming, gardening, or anything else that causes small increases in breathing or heart rate) for at least 30 minutes per day, at least 5 days per week or vigorous-intensity activities (i.e., running, aerobics, heavy yard work, or anything else that causes large increases in breathing or heart rate) for at least 20 minutes per day, at least 3 days per week or both. This can be accomplished *through lifestyle activities (i.e., household, transportation, or leisure-time activities).”*

Take a sample of size $n = 100$, and let X be the number who achieved the recommended level of physical activity. What is the distribution of X?

□

Case A: Sample **with** replacement. *Hence, observations are independent.*

Case B: Sample **without** replacement, but the population size is quite large compared to n; i.e., $N \geq 20n$. *Hence, observations are nearly independent.*

If n is a small percentage of the population size, then sampling **without** *replacement is similar to sampling* **with** *replacement, since sampling the same person more than once would be quite unlikely.*

(a) $\mu_{\hat{p}} = p$ *always.*

(b) $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$ (called the **standard error** of $\hat{p}$), exactly for Case A and approximately for Case B.

(c) (A version of the Central Limit Theorem) The sample proportion $\hat{p}$ is approximately normal if {rule of thumb} $np \geq 10$ and $n(1-p) \geq 10$, for Cases A and B.

Example: *Revisit Virginians who exercise.* Determine the probability that a majority of Virginians in a sample of size 100 achieve the recommended level of physical activity.

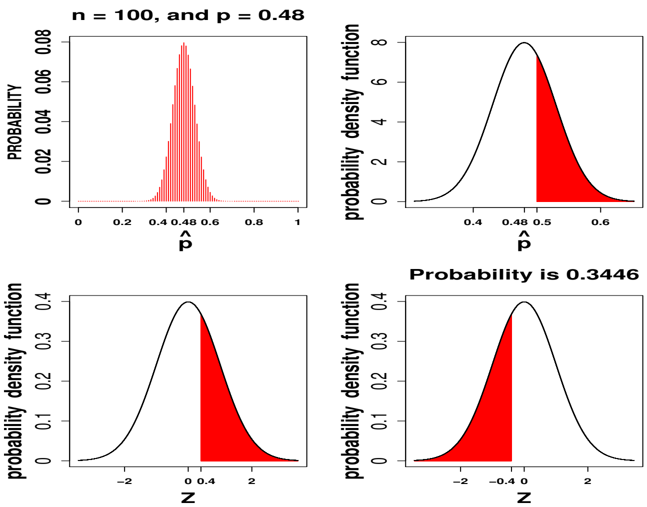
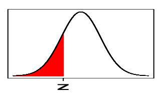
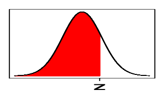


Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7



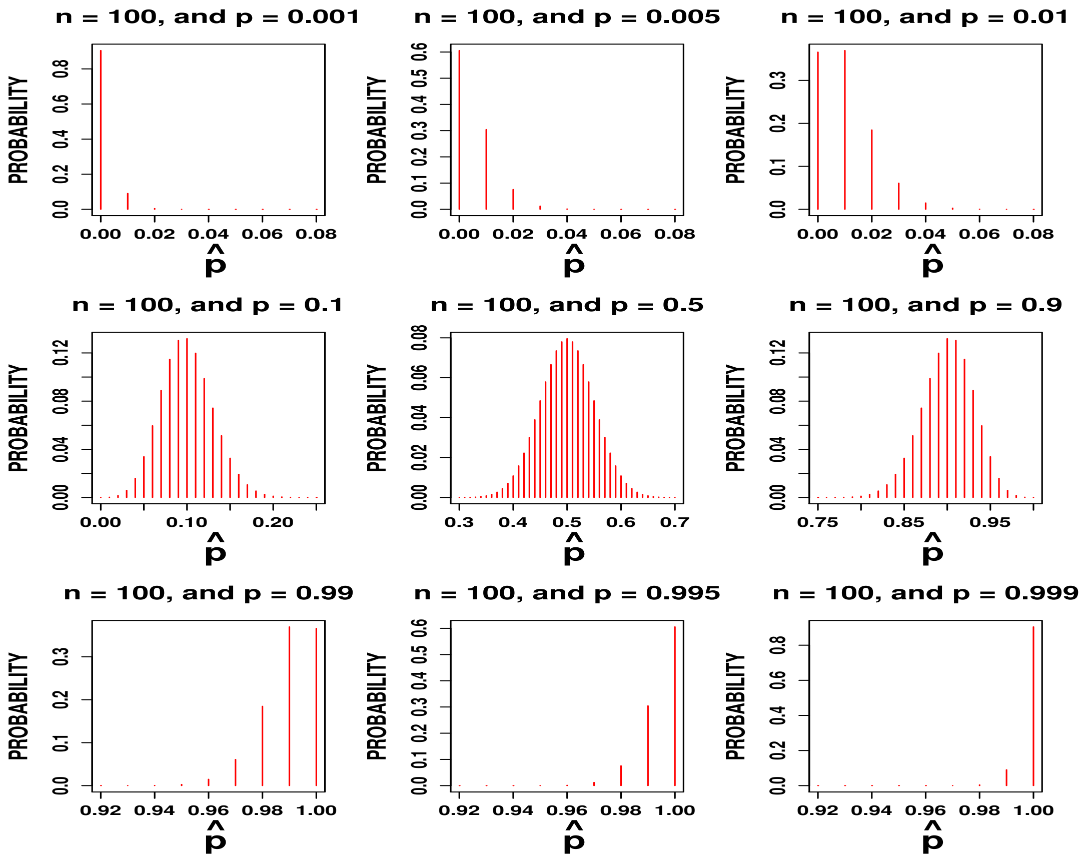


z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

□

Why is the rule of thumb needed?

Example: Consider the *sampling distribution* of $\hat{p}$, for $n = 100$ and various p.



□

Summary of Types of Distributions

The distribution of the original population is called the **population distribution**.

The distribution of a statistic, such as $\hat{p}$ or $\bar{X}$, is called the **sampling distribution**.

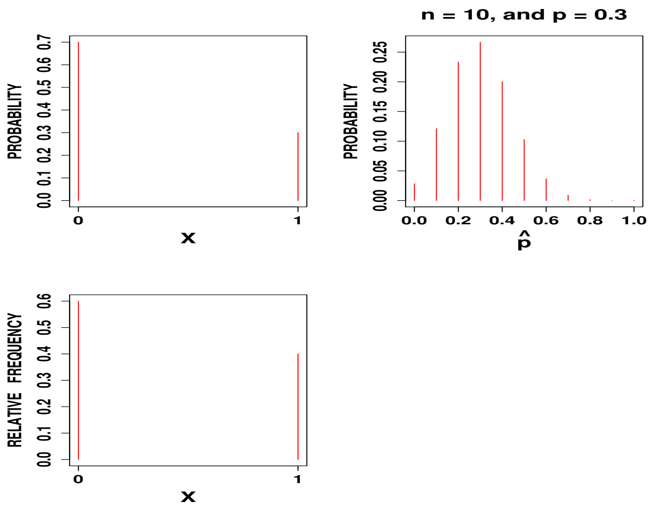
The distribution of one particular data set is called the **data distribution**.

Example: *Revisit the Democrats.* Consider a large population which is 30% Democrat.

(a) Graph the **population distribution**, where a *one* represents a Democrat and a *zero* represents a non-Democrat.

(b) Let $\hat{p}$ be the sample proportion of Democrats in a sample of size $n = 10$. Graph the **sampling distribution** of $\hat{p}$.

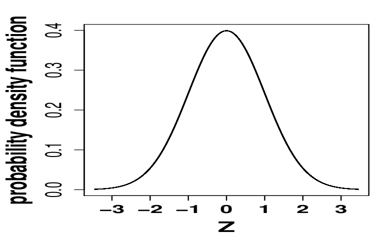
(c) In a sample of size 10, suppose that we have four Democrats, three Republicans, and three Independents. Graph the **data distribution**, where a *one* represents a Democrat and a *zero* represents a non-Democrat.



□

Brief review of formulas (for independent or nearly independent observations)

Notation: $Z \sim N(0, 1)$



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

(b) $Z = \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, if $\bar{X} \sim N(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}} = \sigma/\sqrt{n})$

Here, we need either the original population to be approximately normal or a large sample size (usually $n > 30$, if neither tail of the distribution is too heavy).

Note that $\sigma_{\bar{X}}$, the **standard deviation** of $\bar{X}$, is also called the **standard error** of $\bar{X}$.

(c) $Z = \frac{\hat{p}-\mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$

Note that $\sigma_{\hat{p}}$, the **standard deviation** of $\hat{p}$, is also called the **standard error** of $\hat{p}$.

Here, we need both $np \geq 10$ and $n(1-p) \geq 10$.

6.4 The Normal Approximation to the Binomial Distribution

Suppose $X \sim \text{Binomial}(n,\ p)$.

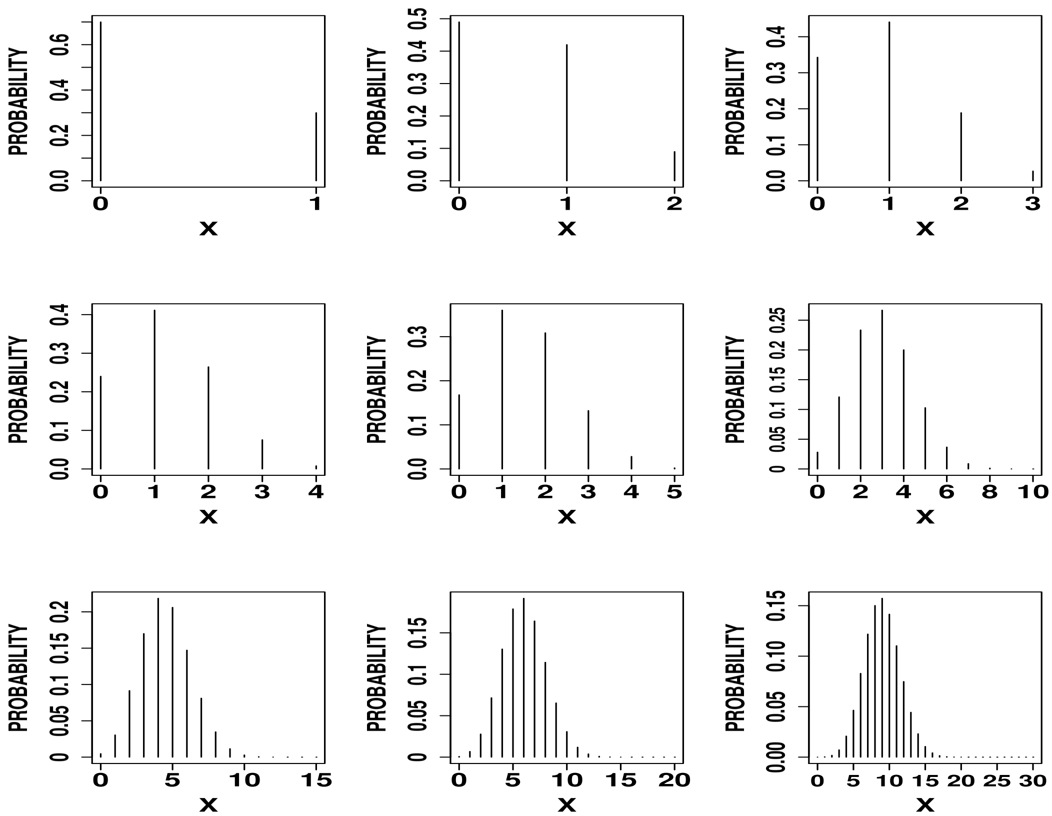
Then, $\mu_x = EX = np$ and $\sigma_x = \sqrt{np(1-p)}$.

Rule of thumb: If $\min\{np,\ n(1-p)\}$ is sufficiently large, say, at least **10**, then X is approximately $N(\mu_x,\ \sigma_x)$.

This result follows from the Central Limit Theorem (defined in section 6.2), since a Binomial random variable is a sample sum of Bernoulli random variables.

Example: *Viewing the normal approximation to the binomial distribution.*

Consider the graphs below for **binomial** random variables, using $p = 0.3$ and $n = 1, 2, 3, 4, 5, 10, 15, 20,$ and 30.



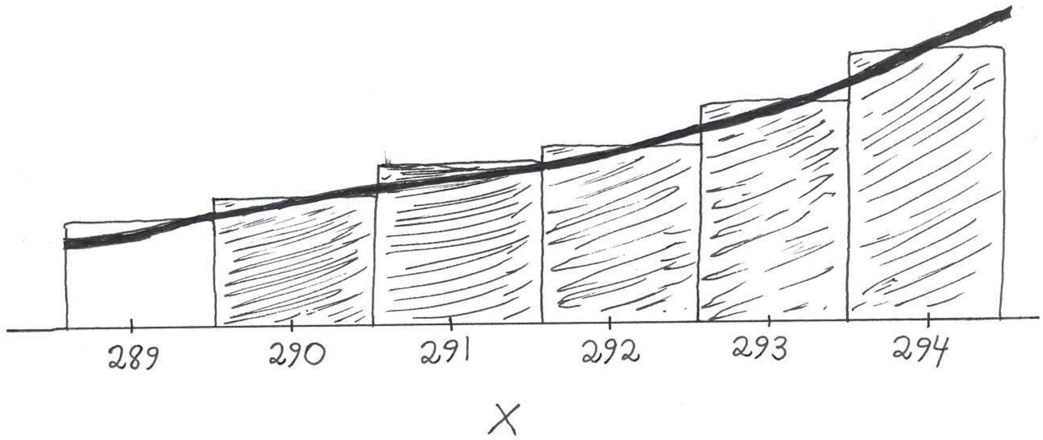
Example: Toss a coin 1000 times where $P(\text{heads}) = 0.3$, and let X be the number of heads.

(a) State the **exact** distribution of X.

(b) Compute the **mean** and **standard deviation** of X.

(c) Check the **rule of thumb**.

(d) Calculate $P(290 \leq X \leq 320)$ using the normal approximation **with continuity correction**.



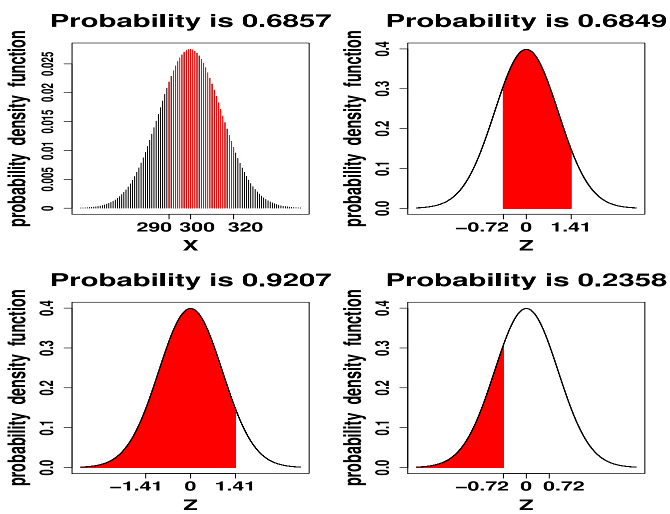
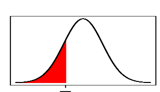
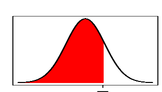


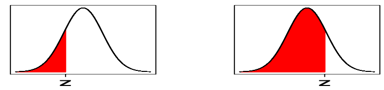
Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7





z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7



z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

□

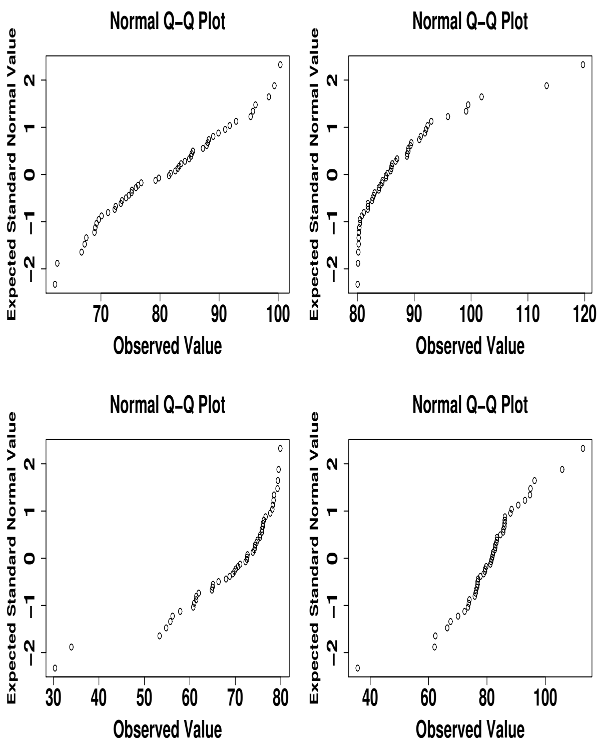
6.5 Assessing Normality

Normal-Quantile plots (also called Quantile-Quantile plots, or Q-Q plots)

How do we know if a sample is from an approximately normal population?

To construct a Q-Q plot, plot *typical* or *quantile* ordered values from a **normal distribution** against the ordered **observations**.

Example: Describe the distributions which likely generated the following Q-Q plots.



□

Other methods for detecting nonnormality include:

* checking for outliers;
* checking for skewness in a dotplot, stem-and-leaf plot, histogram, or boxplot;
* checking for more than one distinct mode in a histogram.