Review of Probability

Ming-Ching Luoh

2005.9.14

Probabilities, Sample Space amd Ramdom Variables

- Outcomes: The mutually exclusive potential *results* of a *random process*.
- Probability: The proportion of the time that the outcome occurs.
- Sample space: The set of all possible outcomes.
- Event: A subset of the sample space.
- Random variables: A random variable is a numerical summary of a random outcome.

Probability distribution of a Discrete Random Variable

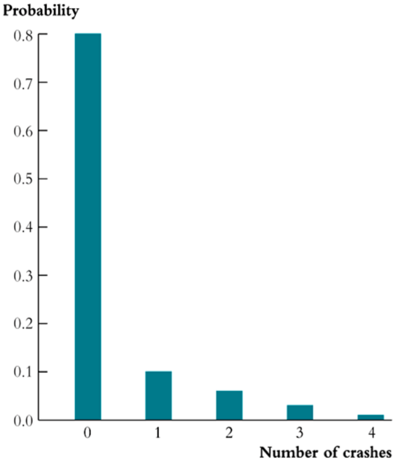
- Probability distribution.
- Probabilities of events.
- Cumulative probability distribution.

TABLE 2.1 Probability of Your Computer Crashing *M* Times

	Outcome (number of crashes)				
	0	1	2	3	4
Probability distribution	0.80	0.10	0.06	0.03	0.01
Cumulative probability distribution	0.80	0.90	0.96	0.99	1.00

FIGURE 2.1 **Probability Distribution of the Number of Computer Crashes**

The height of each bar is the probability that the computer crashes the indicated number of times. The height of the first bar is 0.80, so the probability of 0 computer crashes is 80%. The height of the second bar is 0.1, so the probability of 1 computer crash is 10%, and so forth for the other bars.



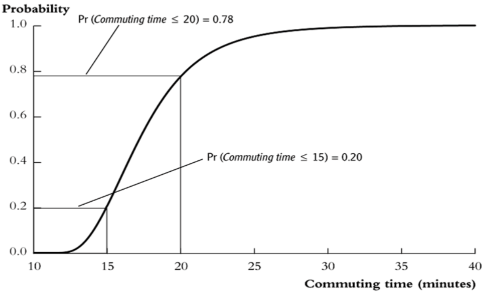
One Example: The **Bernoulli distribution.**
Let G be the gender of the next new person you meet, where $G = 0$ indicates that the person is male and $G = 1$ indicates that she is female.
The outcomes of G and there probabilities are

$$\begin{aligned} G &= 1 \text{ with probability } p \\ &= 0 \text{ with probability } 1 - p \end{aligned}$$

Probability distribution of a Continous Random Variable

- Cumulative probability distribution.

FIGURE 2.2 **Cumulative Distribution and Probability Density Functions of Commuting Time**

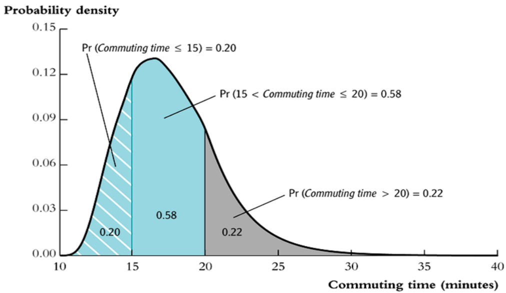


(a) Cumulative distribution function of commuting time

Figure 2.2a shows the cumulative probability distribution (or c.d.f.) of commuting times. The probability that a commuting time is less than 15 minutes is 0.20 (or 20%), and the probability it is less than 20 minutes is 0.78 (78%). Figure 2.2b shows the probability density function (or p.d.f.) of commuting times. Probabilities are given by areas under the p.d.f. The probability that a commuting time is between 15 and 20 minutes is 0.58 (58%), and is given by the area under the curve between 15 and 20 minutes.

- Probability density function (p.d.f.).

FIGURE 2.2 Cumulative Distribution and Probability Density Functions of Commuting Time



(b) Probability density function of commuting time

Figure 2.2a shows the cumulative probability distribution (or c.d.f.) of commuting times. The probability that a commuting time is less than 15 minutes is 0.20 (or 20%), and the probability it is less than 20 minutes is 0.78 (78%). Figure 2.2b shows the probability density function (or p.d.f.) of commuting times. Probabilities are given by areas under the p.d.f. The probability that a commuting time is between 15 and 20 minutes is 0.58 (58%), and is given by the area under the curve between 15 and 20 minutes.

Expected Value and the Mean

Suppose the random variable Y takes on k possible values, $y_1, \ldots, y_k$, where y_1 denotes the first value, y_2 denotes the second value, etc., and that the probability that Y takes on y_1 is p_1, the probability that Y takes on y_2 is p_2, and so forth. The expected value of Y, denoted $E(Y)$, is

$$E(Y) = y_1 p_1 + y_2 p_2 + \cdots + y_k p_k = \sum_{i=1}^{k} y_i p_i, \qquad (2.4)$$

where the notation "$\sum_{i=1}^{k} y_i p_i$" means "the sum of $y_i p_i$ for i running from 1 to k." The expected value of Y is also called the mean of Y or the expectation of Y and is denoted μ_Y.

Expected value of a Bernoulli random variable

$$\mathrm{E}(G) = 1 \times p + 0 \times (1-p) = p$$

Expected value of a continuous random variable

Let f(Y) is the p.d.f of random variable Y, then the expected value of Y is

$$\mathrm{E}(Y) = \int Y f(Y) dY$$

Variance and Standard Deviation

The variance of the discrete random variable Y, denoted σ_Y^2, is

$$\sigma_Y^2 = \text{var}(Y) = E[(Y-\mu_Y)^2] = \sum_{i=1}^{k} (y_i - \mu_Y)^2 p_i. \qquad (2.6)$$

The standard deviation of Y is σ_Y, the square root of the variance. The units of the standard deviation are the same as the units of Y.

Variance of a Bernoulli random variable

The mean of the Bernoulli random variable G is $\mu_G = p$, so its variance is

$$\begin{aligned} \text{Var}(G) &= \sigma_G^2 = (1-p)^2 \times p + (0-p)^2 \times (1-p) \\ &= p(1-p) \end{aligned}$$

The standard deviation is $\sigma_G = \sqrt{p(1-p)}$.

- The expected value of Y^r is called the r^{th} **moments** of the random variable Y.
  That is the r^{th} moment of Y is $\mathrm{E}(Y^r)$.
- The mean of Y, $\mathrm{E}(Y)$, is also called the first moment of Y.

Mean and Variance of a Linear Function of a Random Variable

Suppose X is a random variable with mean μ_X and variance σ_X^2, and

$$Y = a + bX$$

Then the mean and variance of Y are

$$\begin{aligned} \mu_Y &= a + b\mu_X \\ \sigma_Y^2 &= b^2\sigma_X^2 \end{aligned}$$

and the standard deviation of Y is $\sigma_Y = b\sigma_X$.

The **joint probability distribution** of two discrete random variables, say X and Y, is the probability that the random variables simultaneously take on certain values, say x and y. The joint probability distribution can be written as the function $\Pr(X = x, Y = y)$.

TABLE 2.2 Joint Distribution of Weather Conditions and Commuting Times

	Rain ($X = 0$)	No Rain ($X = 1$)	**Total**
Long Commute ($Y = 0$)	0.15	0.07	0.22
Short Commute ($Y = 1$)	0.15	0.63	0.78
Total	0.30	0.70	1.00

The **marginal probability distribution** of a random variable Y is just another name for its probability distribution.

$$\Pr(Y = y) = \sum_{i=1}^{l} \Pr(X = x_i, Y = y)$$

TABLE 2.2 Joint Distribution of Weather Conditions and Commuting Times

	Rain ($X = 0$)	No Rain ($X = 1$)	**Total**
Long Commute ($Y = 0$)	0.15	0.07	0.22
Short Commute ($Y = 1$)	0.15	0.63	0.78
Total	0.30	0.70	1.00

Conditional distribution of Y given $X = x$ is

$$\Pr(Y = y | X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)}$$

Condtional expectation of Y given $X = x$ is

$$E(Y|X = x) = \sum_{i=1}^{k} y_i \Pr(Y = y_i | X = x)$$

TABLE 2.3 Joint and Conditional Distributions of Computer Crashes (M) and Computer Age (A)

A. Joint Distribution

	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	Total
Old computer ($A = 0$)	0.35	0.065	0.05	0.025	0.01	0.50
New computer ($A = 1$)	0.45	0.035	0.01	0.005	0.00	0.50
Total	0.8	0.1	0.06	0.03	0.01	1.00

B. Conditional Distributions of M given A

	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	Total
$\Pr(M \mid A = 0)$	0.70	0.13	0.10	0.05	0.02	1.00
$\Pr(M \mid A = 1)$	0.90	0.07	0.02	0.01	0.00	1.00

The mean of Y is the weighted average of the conditional expectation of Y given X, weighted by the probability distribution of X.

$$E(Y) = \sum_{i=1}^{l} E(Y|X = x_i) \Pr(X = x_i)$$

Stated differently, the expectation of Y is the expectation of the conditional expectation of Y given X, that is,

$$E(Y) = E[E(Y|X)],$$

where the inner expectation is computed using the conditional distribution of Y given X and the outer expectation is computed using the marginal distribution of X.

This is known as the **law of iterated expectations**.

Proof that $\mathrm{E}(Y) = \sum_{i=1}^{l} \mathrm{E}(Y|X = x_i) \Pr(X = x_i)$

$$
\begin{aligned}
\mathrm{E}(Y) &= \sum_{j=1}^{k} y_j \Pr(Y = y_j) = \sum_{j=1}^{k} y_j \sum_{i=1}^{l} \Pr(Y = y_j, X = x_i) \\
&= \sum_{j=1}^{k} y_j \sum_{i=1}^{l} \Pr(Y = y_j | X = x_i) \Pr(X = x_i) \\
&= \sum_{i=1}^{l} \underline{\sum_{j=1}^{k} y_j \Pr(Y = y_j | X = x_i)} \Pr(X = x_i) \\
&= \sum_{i=1}^{l} \mathrm{E}(Y|X = x_i) \Pr(X = x_i)
\end{aligned}
$$

Conditional variance. The variance of Y conditional on X is the variance of the conditional distribution of Y given X.

$$\text{Var}(Y|X=x) = \sum_{i=1}^{k} [y_i - \text{E}(Y|X=x)]^2 \Pr(Y = y_i | X = x)$$

Independence

- Two random variable X and Y are **independently distributed,** or **independent**, if knowing the value of one of the variables provides no information about the other.
- That is, X and Y are independent if for all values of x and y,

$$\Pr(Y = y|X = x) = \Pr(Y = y)$$

State dfferently, X and Y are independent if

$$\frac{\Pr(X = x, Y = y)}{\Pr(X = x)} = \Pr(Y = y)$$

$$\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y)$$

That is, the joint distribution of two independent random variables is the product of their marginal distributions.

Covariance.

One measure of the extent to which two random variables move together is their covariance.

$$
\begin{aligned}
& \mathrm{Cov}(X, Y) = \sigma_{XY} \\
= \; & \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] \\
= \; & \sum_{i=1}^{l} \sum_{j=1}^{k} (x_i - \mu_x)(y_j - \mu_Y) \Pr(X = x_i, Y = y_j)
\end{aligned}
$$

Correlation.

The correlation is an alternative measure of dependence between X and Y that solves the "unit" problem of covariance.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

The random variables X and Y are said to be **uncorrelated** if $\text{Corr}(X, Y) = 0$. The correlation is always between -1 and 1.

The Mean and Variance of Sums of Random Variables

$$
\begin{aligned}
\mathrm{E}(X) + \mathrm{E}(Y) &= \mathrm{E}(X) + \mathrm{E}(Y) = \mu_X + \mu_Y \\
\mathrm{Var}(X + Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y) \\
&= \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}
\end{aligned}
$$

Means, Variances, and Covariances of Sums of Random Variables

Let X, Y, and V be random variables, let μ_X and σ_X^2 be the mean and variance of X, let σ_{XY} be the covariance between X and Y (and so forth for the other variables), and let a, b, and c be constants. The following facts follow from the definitions of the mean, variance, and covariance:

$$E(a + bX + cY) = a + b\mu_X + c\mu_Y, \quad (2.29)$$

$$\text{var}(a + bY) = b^2\sigma_Y^2, \quad (2.30)$$

$$\text{var}(aX + bY) = a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2, \quad (2.31)$$

$$E(Y^2) = \sigma_Y^2 + \mu_Y^2, \quad (2.32)$$

$$\text{cov}(a + bX + cV, Y) = b\sigma_{XY} + c\sigma_{VY}, \text{ and} \quad (2.33)$$

$$E(XY) = \sigma_{XY} + \mu_X\mu_Y. \quad (2.34)$$

$$|\text{corr}(X,Y)| \leq 1 \text{ and } |\sigma_{XY}| \leq \sqrt{\sigma_X^2\sigma_Y^2} \text{ (correlation inequality).} \quad (2.35)$$

The probability density function of a normal distributed random variable (the **normal p.d.f.**) is

$$f_Y(y) = \frac{1}{\sigma_Y\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]$$

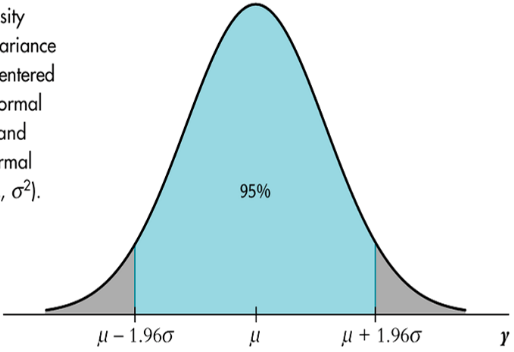
where $\exp(x)$ is the exponential function of x.
The factor $\frac{1}{\sigma_Y\sqrt{2\pi}}$ ensures that

$$\Pr(-\infty \leq Y \leq \infty) = \int_{-\infty}^{\infty} f_Y(y)dy = 1$$

The normal distribution with mean μ and variance σ^2 is expressed as $N(\mu, \sigma^2)$.

FIGURE 2.3 The Normal Probability Density

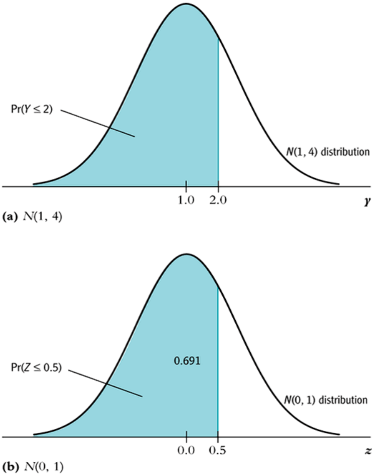
The normal probability density function with mean μ and variance σ^2 is a bell-shaped curve, centered at μ. The area under the normal p.d.f. between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95. The normal distribution is denoted $N(\mu, \sigma^2)$.



- The **standard normal distribution** is the normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ and is denoted $N(0, 1)$.
- The standard normal distribution is often denoted by Z and its cumulative distribution function is denoted by Φ. Accordingly, $\Pr(Z \leq c) = \Phi(c)$, where c is a constant.

FIGURE 2.4 Calculating the Probability that $Y \leq 2$ When Y is Distributed $N(1, 4)$

To calculate $\Pr(Y \leq 2)$, standardize Y, then use the standard normal distribution table. Y is standardized by subtracting its mean ($\mu = 1$) and dividing by its standard deviation ($\sigma_Y = 2$). The probability that $Y \leq 2$ is shown in Figure 2.4a, and the corresponding probability after standardizing Y is shown in Figure 2.4b. Because the standardized random variable, $\frac{Y-1}{2}$, is a standard normal (Z) random variable, $\Pr(Y \leq 2) = \Pr\left(\frac{Y-1}{2} \leq \frac{2-1}{2}\right) = \Pr(Z \leq 0.5)$. From Appendix Table 1, $\Pr(Z \leq 0.5) = 0.691$.



The bivariate normal distribution.

The **bivariate normal p.d.f.** for the two random variables X and Y is

$$
\begin{aligned}
& g_{X,Y}(x,y) \\
= \quad & \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \\
\times \quad & \exp\left\{\frac{1}{-2(1-\rho_{XY}^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{XY}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)\right.\right. \\
& \left.\left. + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}
\end{aligned}
$$

where ρ_{XY} is the correlation between X and Y.

Important properties for normal distribution.

1. If X and Y have a bivariate normal distribution with covariance σ_{XY}, and if a and b are two constants, then

$$aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY})$$

2. The marginal distribution of each of the two variables is normal. This follows by setting $a = 1$, $b = 0$ in 1.
3. If $\sigma_{XY} = 0$, then X and Y are independent.

- The **chi-squared distribution** is the distribution of the sum of m squared **independent** standard normal random variables.
- The distribution depends on m, which is called the degrees of freedom of the chi-squared distribution.
- A chi-squared distribution with m degrees of freedom is denoted χ^2_m.

- The $F_{m,\infty}$ **distribution** is the distribution of a random variable with a chi-squared distribution with m degrees of freedom, divided by m.
- Equivalently, the $F_{m,\infty}$ distribution is the distribution of the average of m squared standard normal random variables.

The **Student t distribution** with m degrees of freedom is defined to be the distribution of the ratio of a *standard normal random variable*, divided by the *square root* of an *independently distributed chi-squared random variable with m degrees of freedom divided by m.*

That is, let Z be a standard normal random variable, let W be a random variable with a chi-squared distribution with m degrees of freedom, and let Z and W be independently distributed. Then

$$\frac{Z}{\sqrt{\frac{W}{m}}} \sim t_m$$

When m is 30 or more, the Student t distribution is well approximated by the standard normal distribution, and t_∞ distribution equals the standard normal distribution Z.

Simple random sampling is the simplest sampling scheme in which n objects are selected at *random* from a **population** and each member of the population is equally likely to be included in the sample.
Since the members of the population included in the sample are selected at random, the values of the observations $Y_1, \cdots, Y_n$ are themselves random.

i.i.d. draws.

Because $Y_1, \cdots, Y_n$ are randomly drawn from the same population, the marginal distribution of Y_i is the same for eacn $i = 1, \cdots, n$. $Y_1, \cdots, Y_n$ are said to be **identically distributed.**

When $Y_1, \cdots, Y_n$ are drawn from the same distribution and are indepently distributed, they are said to be **independently and identically distributed**, or **i.i.d.**

The sample average of the n observations $Y_1, \cdots, Y_n$ is

$$\bar{Y} = \frac{1}{n}(Y_1 + \cdots, Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Because $Y_1, \cdots, Y_n$ are random, their average is random and has a probability distribution.
The distribution of $\bar{Y}$ is called the **sampling distribution** of $\bar{Y}$.

Mean and Variance of $\bar{Y}$

Suppose $Y_1, \cdots, Y_n$ are i.i.d. and let μ_Y and σ_Y^2 denote the mean and variance of Y_i. Then

$$
\begin{aligned}
\mathrm{E}(\bar{Y}) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(Y_i) = \mu_Y \\
\mathrm{Var}(\bar{Y}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\mathrm{Cov}(Y_i, Y_j) \\
&= \frac{\sigma_Y^2}{n}
\end{aligned}
$$

Two approaches to characterizing sample distributions.

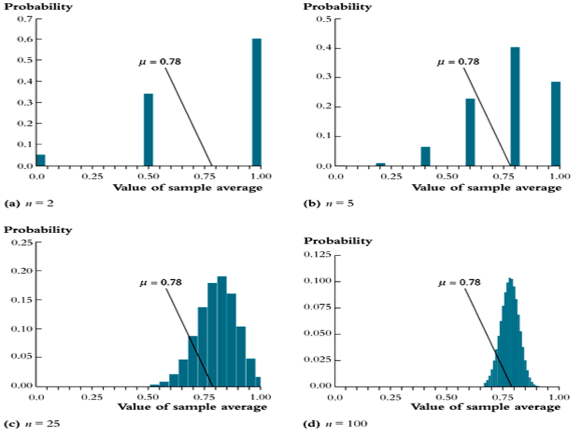
- **Exact** distribution, or finite sample distribution when the distribution of Y is known.
- **Asymptotic** distribution: large-sample approximation to the sampling distribution.

- The **law of large numbers** states that, *under general conditions*, $\bar{Y}$ will be near μ_Y with very high probability when n is large.
- The property that $\bar{Y}$ is near μ_Y with increasing probability as n increases is called **convergence in probability**, or **consistency**.
- The law of large numbers states that, under certain conditions, $\bar{Y}$ converges in probability to μ_Y, or, $\bar{Y}$ is consistent for μ_Y.

The conditions for the law of large numbers are

- $Y_i, i = 1, \cdots, n$, are i.i.d.
- The variance of Y_i, σ_Y^2, is finite.

FIGURE 2.6 Sampling Distribution of the Sample Average of n Bernoulli Random Variables



The distributions are the sampling distributions of $\overline{Y}$, the sample average of n independent Bernoulli random variables with $p = \Pr(Y_i = 1) = 0.78$ (the probability of a fast commute is 78%). The variance of the sampling distribution of $\overline{Y}$ decreases as n gets larger, so the sampling distribution becomes more tightly concentrated around its mean $\mu = 0.78$ as the sample size n increases.

Formal definitions of consistency and law of large numbers.

Consistency and convergency in probability.

Let $S_1, S_2, \cdots, S_n, \cdots$ be a sequence of random variables. For example, S_n could be the sample average $\bar{Y}$ of a sample of n observations of the randome variable Y.

The sequence of randome variables $\{S_n\}$ is said to **converge in probability** to a limit, μ, if the probability that S_n is within $\pm\delta$ of μ tends to one as $n \to \infty$, as long as the constant δ is positive.

That is,

$$S_n \xrightarrow{p} \mu \text{ if and only if } \Pr[|S_n - \mu| \geq \delta] \rightarrow 0$$

as $n \rightarrow \infty$ for every $\delta > 0$.
If $S_n \xrightarrow{p} \mu$, then S_n is said to be a **consistent estimator** of μ.

The law of large numbers.

If $Y_1, \cdots, Y_n$ are i.i.d., $\mathrm{E}(Y_i) = \mu_Y$ and $\mathrm{Var}(Y_i) < \infty$, then

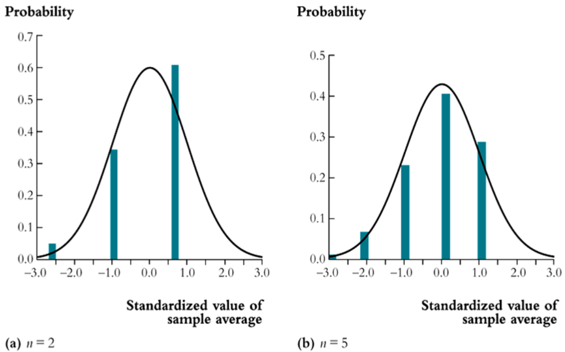
$$\bar{Y} \xrightarrow{p} \mu_Y$$

The **central limit theorem** says that, under general conditions, the distribution of $\bar{Y}$ is well approximated by a normal distribution when n is large.

Since the mean of $\bar{Y}$ is μ_Y and its variance if $\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n}$, when n is large the distribution of $\bar{Y}$ is approximately $N(\mu_Y, \sigma_{\bar{Y}}^2)$.

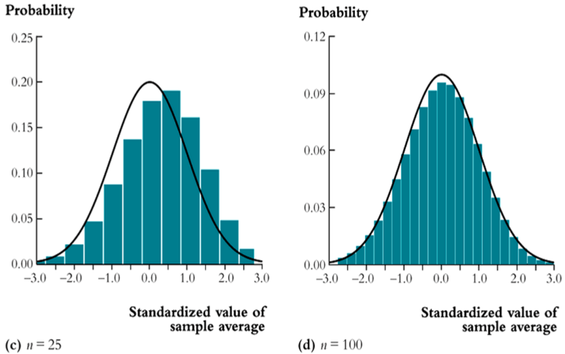
Accordingly, $\frac{\bar{Y}-\mu_Y}{\sigma_{\bar{Y}}}$ is well approximated by the standard normal distribution $N(0, 1)$.

FIGURE 2.7 Distribution of the Standardized Sample Average of n Bernoulli Random Variables with $p = .78$



The sampling distribution of $\overline{Y}$ in Figure 2.6 is plotted here after standardizing $\overline{Y}$. This centers the distributions in Figure 2.6 and magnifies the scale on the horizontal axis by a factor of $\sqrt{n}$. When the sample size is large, the sampling distributions are increasingly well approximated by the normal distribution (the solid line), as predicted by the central limit theorem.

FIGURE 2.7 Distribution of the Standardized Sample Average of n Bernoulli Random Variables with $p = .78$



The sampling distribution of $\overline{Y}$ in Figure 2.6 is plotted here after standardizing $\overline{Y}$. This centers the distributions in Figure 2.6 and magnifies the scale on the horizontal axis by a factor of $\sqrt{n}$. When the sample size is large, the sampling distributions are increasingly well approximated by the normal distribution (the solid line), as predicted by the central limit theorem.

Convergence in distribution.

Let $F_1, \cdots, F_n, \cdots$ be a sequence of cumulative distribution functions corresponding to a sequence of random variables, $S_1, \cdots, S_n, \cdots$. Then the sequence of random variables S_n is said to **converge in distribution** to S (denoted $S_n \xrightarrow{d} S$) if the distribution functions $\{F_n\}$ converge to F.

That is,

$$S_n \xrightarrow{d} S \text{ if and only if } \lim_{n \to \infty} F_n(t) = F(t),$$

where the limit holds at all points t at which the limiting distribution F is continuous.
The distribution F is called the **asymptotic distribution** of S_n.

The central limit theorem.

If $Y_1, \cdots, Y_n$ are $i.i.d.$ and $0 < \sigma_Y^2 < \infty$, then

$$\sqrt{n}(\bar{Y} - \mu_Y) \xrightarrow{d} N(0, \sigma_Y^2)$$

In other words, the asymptotic distribution of

$$\sqrt{n}\frac{\bar{Y} - \mu_Y}{\sigma_Y} = \frac{\bar{Y} - \mu_Y}{\frac{\sigma_Y}{\sqrt{n}}} = \frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}}$$

is $N(0, 1)$.

Slutsky's theorem combines consistency and convergence in distribution.

Suppose that $a_n \xrightarrow{p} a$, where a is a constant, and $S_n \xrightarrow{d} S$. Then

$$
\begin{aligned}
a_n + S_n &\xrightarrow{d} a + S, \\
a_n S_n &\xrightarrow{d} aS, \\
S_n / a_n &\xrightarrow{d} S/a, \text{ if } a \neq 0
\end{aligned}
$$

Continuous mapping theorem:

If g is a continuous function, then

- if $S_n \xrightarrow{p} a$, then $g(S_n) \xrightarrow{p} g(a)$, and
- if $S_n \xrightarrow{d} S$, then $g(S_n) \xrightarrow{d} g(S)$.

But, how large of n is "large enough?"

The answer is: it depends on the distribution of the underlying Y_i that make up the average.

At one extreme, if the Y_i are themselves normally distributed, then $\bar{Y}$ is exactly normally distributed for all n.

In contrast, when Y_i is far from normally distributed, then this approximation can require $n = 30$ or even more.

But, how large of n is "large enough?"

The answer is: it depends on the distribution of the underlying Y_i that make up the average.

At one extreme, if the Y_i are themselves normally distributed, then $\bar{Y}$ is exactly normally distributed for all n.

In contrast, when Y_i is far from normally distributed, then this approximation can require $n = 30$ or even more.

But, how large of n is "large enough?"

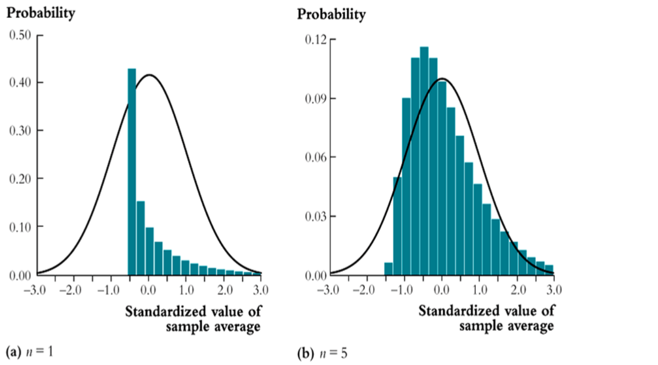
The answer is: it depends on the distribution of the underlying Y_i that make up the average.

At one extreme, if the Y_i are themselves normally distributed, then $\bar{Y}$ is exactly normally distributed for all n.

In contrast, when Y_i is far from normally distributed, then this approximation can require $n = 30$ or even more.

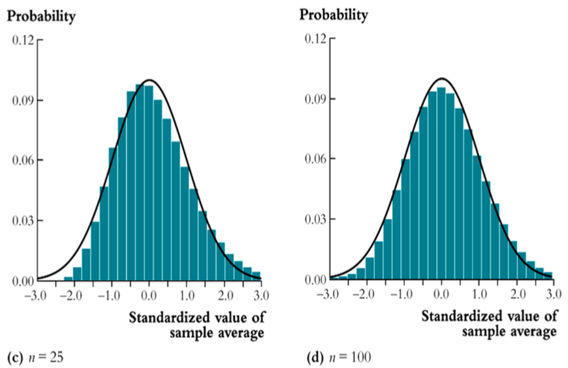
Example: A skewed distribution.

FIGURE 2.8 Distribution of the Standardized Sample Average of *n* Draws from a Skewed Distribution



The figures show the sampling distribution of the standardized sample average of *n* draws from the skewed (asymmetric) population distribution shown in Figure 2.8a. When *n* is small ($n = 5$), the sampling distribution, like the population distribution, is skewed. But when *n* is large ($n = 100$), the sampling distribution is well approximated by a standard normal distribution (solid line), as predicted by the central limit theorem.

FIGURE 2.8 Distribution of the Standardized Sample Average of *n* Draws from a Skewed Distribution



The figures show the sampling distribution of the standardized sample average of *n* draws from the skewed (asymmetric) population distribution shown in Figure 2.8a. When *n* is small ($n = 5$), the sampling distribution, like the population distribution, is skewed. But when *n* is large ($n = 100$), the sampling distribution is well approximated by a standard normal distribution (solid line), as predicted by the central limit theorem.