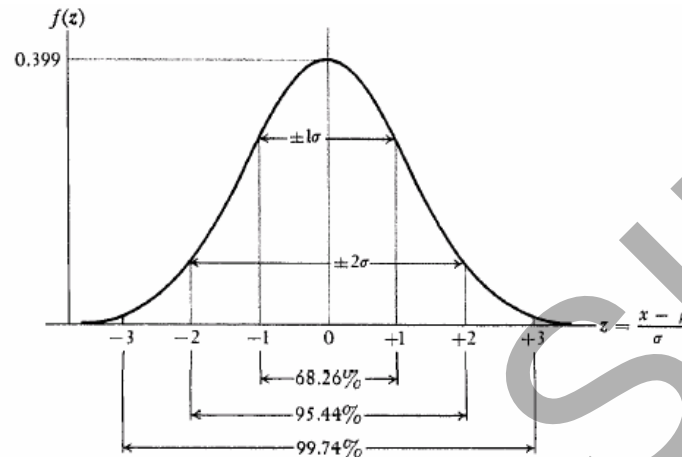# Normal distribution

The normal distribution is the most widely known and used of all distributions. Because the normal distribution approximates many natural phenomena so well, it has developed into a standard of reference for many probability problems.

I.   Characteristics of the Normal distribution

- Symmetric, bell shaped
- <u>Continuous</u> for all values of X between -∞ and ∞ so that each conceivable interval of real numbers has a probability other than zero.
- -∞ ≤ X ≤ ∞
- Two parameters, μ and σ. Note that the normal distribution is actually a family of distributions, since μ and σ determine the shape of the distribution.

- The rule for a normal density function is

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

- The notation $N(\mu, \sigma^2)$ means normally distributed with mean μ and variance σ^2. If we say $X \sim N(\mu, \sigma^2)$ we mean that X is distributed $N(\mu, \sigma^2)$.
- About 2/3 of all cases fall within one standard deviation of the mean, that is

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = .6826.$$

- About 95% of cases lie within 2 standard deviations of the mean, that is
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = .9544$$

II.   Why is the normal distribution useful?

- Many things actually are normally distributed, or very close to it. For example, height and intelligence are approximately normally distributed; measurement errors also often have a normal distribution
- The normal distribution is easy to work with mathematically. In many practical cases, the methods developed using normal theory work quite well even when the distribution is not normal.
- There is a very strong connection between the size of a sample N and the extent to which a sampling distribution approaches the normal form. Many sampling distributions based on large N can be approximated by the normal distribution even though the population distribution itself is definitely not normal.

III.   The standardized normal distribution.

a.   General Procedure. As you might suspect from the formula for the normal density function, it would be difficult and tedious to do the calculus every time we had a new set of parameters for μ and σ. So instead, we usually work with the standardized normal distribution, where μ = 0 and σ = 1, i.e. N(0,1). That is, rather than directly solve a problem involving a normally distributed variable X with mean μ and standard deviation σ, an indirect approach is used.

1.   We first convert the problem into an equivalent one dealing with a normal variable measured in standardized deviation units, called a standardized normal variable. To do this, if X ~ N(μ, σ²), then

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

2.   A table of standardized normal values (Appendix E, Table I) can then be used to obtain an answer in terms of the converted problem.

3.   If necessary, we can then convert back to the original units of measurement. To do this, simply note that, if we take the formula for Z, multiply both sides by σ, and then add μ to both sides, we get

$$X = Z\sigma + \mu$$

4.   The interpetation of Z values is straightforward. Since σ = 1, if Z = 2, the corresponding X value is exactly 2 standard deviations above the mean. If Z = -1, the corresponding X value is one standard deviation below the mean. If Z = 0, X = the mean, i.e. μ.

b.   Rules for using the standardized normal distribution. It is very important to understand how the standardized normal distribution works, so we will spend some time here going over it. Recall that, for a random variable X,

$$F(x) = P(X \leq x)$$

Appendix E, Table I (Or see Hays, p. 924) reports the cumulative normal probabilities for normally distributed variables in standardized form (i.e. Z-scores). That is, this table reports $P(Z \leq z) = F(z)$. For a given value of Z, the table reports what proportion of the distribution lies below that value. For example, $F(0) = .5$; half the area of the standardized normal curve lies to the left of $Z = 0$. Note that only positive values of Z are reported; as we will see, this is not a problem, since the normal distribution is symmetric. We will now show how to work with this table.

NOTE: While memorization may be useful, you will be much better off if you gain an intuitive understanding as to why the rules that follow are correct. Try drawing pictures of the normal distribution to convince yourself that each rule is valid.

RULES:

1. **P(Z ≤ a)**
   = F(a)   (use when a is positive)
   = 1 - F(-a)   (use when a is negative)
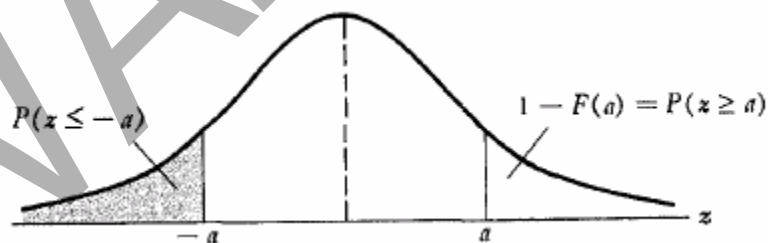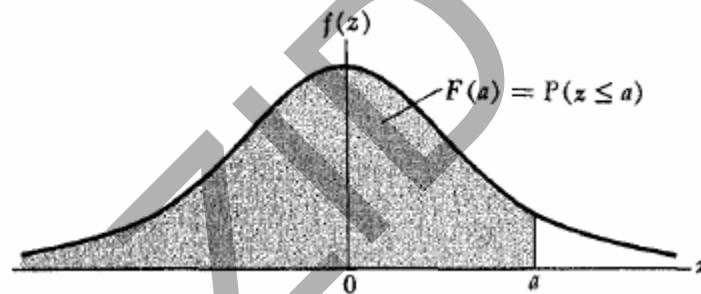
EX: Find $P(Z \leq a)$ for a = 1.65, -1.65, 1.0, -1.0

To solve: for positive values of a, look up and report the value for F(a) given in Appendix E, Table I. For negative values of a, look up the value for F(-a) (i.e. F(absolute value of a)) and report 1 - F(-a).

$P(Z \leq 1.65) = F(1.65) = .95$
$P(Z \leq -1.65) = F(-1.65) = 1 - F(1.65) = .05$

$P(Z \le 1.0) = F(1.0) = .84$
$P(Z \le -1.0) = F(-1.0) = 1 - F(1.0) = .16$

You can also easily work in the other direction, and determine what a is given $P(Z \le a)$

EX: Find a for $P(Z \le a) = .6026, .9750, .3446$

To solve: for $p \ge .5$, find the probability value in Table I, and report the corresponding value for Z. For $p < .5$, compute $1 - p$, find the corresponding Z value, and report the negative of that value, i.e. -Z.
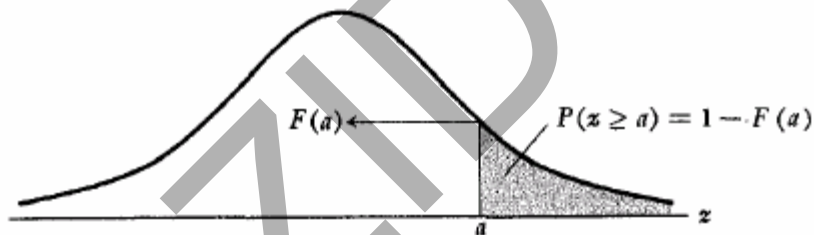$P(Z \le .26) = .6026$
$P(Z \le 1.96) = .9750$
$P(Z \le -.40) = .3446$ (since $1 - .3446 = .6554 = F(.40)$)

NOTE: It may be useful to keep in mind that $F(a) + F(-a) = 1$.

2.   **$P(Z \ge a)$**
     $= 1 - F(a)$ (use when a is positive)
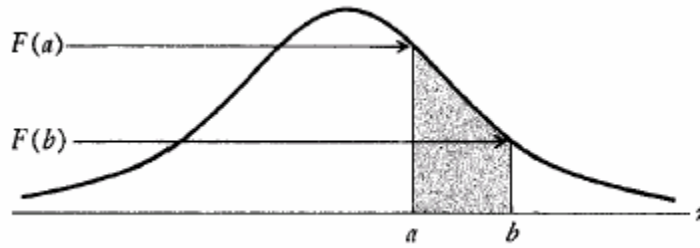   **$= F(-a)$**   (use when a is negative)

EX: Find $P(Z \ge a)$ for $a = 1.5, -1.5$

To solve: for a positive, look up F(a), as before, and subtract F(a) from 1. For a negative, just report F(-a).
$P(Z \ge 1.5) = 1 - F(1.5) = 1 - .9332 = .0668$
$P(Z \ge -1.5) = F(1.5) = .9332$

3.   P(a ≤ Z ≤ b) = F(b) - F(a)

EX: Find P(a ≤ Z ≤ b) for a = -1 and b = 1.5

To solve: determine F(b) and F(a), and subtract.
P(-1 ≤ Z ≤ 1.5) = F(1.5) - F(-1) = F(1.5) - (1 - F(1)) = .9332 - 1 + .8413 = .7745

4.   For a positive, P(-a ≤ Z ≤ a) = 2F(a) - 1

PROOF:
P(-a ≤ Z ≤ a)
= F(a) - F(-a)   (by rule 3)
= F(a) - (1 - F(a))   (by rule 1)
= F(a) - 1 + F(a)
= 2F(a) - 1

EX: find P(-a ≤ Z ≤ a) for a = 1.96, a = 2.58

P(-1.96 ≤ Z ≤ 1.96) = 2F(1.96) - 1 = (2 * .975) - 1 = .95
P(-2.58 ≤ Z ≤ 2.58) = 2F(2.58) - 1 = (2 * .995) - 1 = .99

4B.   For a positive, F(a) = [1 + P(-a ≤ Z ≤ a)] / 2

EX: find a for P(-a ≤ Z ≤ a) = .90, .975

F(a) = (1 + .90)/2 = .95, implying a = 1.65.

For P(-a ≤ Z ≤ a) = .975,

F(a) = (1 + .975)/2 = .9875, implying a = 2.24

NOTE: Suppose we were asked to find a and b for P(a ≤ Z ≤ b) = .90. There are an infinite number of values that we could use; for example, we could have a = negative infinity and b = 1.28, or a = -1.28 and b = positive infinity, or a = -1.34 and b = 2.32, etc. The smallest interval between a and b will always be found by choosing values for a and b such that a = -b. For example, for P(a ≤ Z ≤ b) = .90, a = -1.65 and b = 1.65 are the "best" values to choose, since they yield the smallest possible value for b - a.

IV.   Using the standardized normal distribution. Now that we know how to read Table I, we can give some examples of how to use standardized scores to address various questions.

EXAMPLES.

1. The top 5% of applicants (as measured by GRE scores) will receive scholarships. If GRE ~ $N(500, 100^2)$, how high does your GRE score have to be to qualify for a scholarship?

Solution. Let X = GRE. We want to find x such that

$$P(X \geq x) = .05$$

This is too hard to solve as it stands - so instead, compute

$$Z = (X - 500)/100 \qquad (\text{NOTE: } Z \sim N(0,1))$$

and find z for the problem,

$$P(Z \geq z) = .05$$

Note that $P(Z \geq z) = 1 - F(z)$ (Rule 2). If $1 - F(z) = .05$, then $F(z) = .95$. Looking at Table I in Appx E, $F(z) = .95$ for $z = 1.65$ (approximately).

Hence, $z = 1.65$. To find the equivalent x, compute

$x = (z * 100) + 500 = (1.65 * 100) + 500 = \underline{665}$. Thus, your GRE score needs to be 665 or higher to qualify for a scholarship.

2.   Family income ~ $N(\$25000, \$10000^2)$. If the poverty level is \$10,000, what percentage of the population lives in poverty?

Solution. Let X = Family income. We want to find $P(X \leq \$10,000)$. This is too hard to compute directly, so let
$$Z = (X - \$25,000)/\$10,000.$$

If x = \$10,000, then $z = (\$10,000 - \$25,000)/\$10,000 = -1.5$. So,

$P(X \leq \$10,000) = P(Z \leq -1.5) = F(-1.5) = 1 - F(1.5) = 1 - .9332 = \underline{.0668}$. Hence, a little under 7% of the population lives in poverty.

3.   A new tax law is expected to benefit "middle income" families, those with incomes between \$20,000 and \$30,000. If Family income ~ N(\$25000, \$10000^2), what percentage of the population will benefit from the law?

Solution. Let X = Family income. We want to find P(\$20,000 X ≤ \$30,000). To solve, let

Z = (X - \$25,000)/\$10,000.

Note that when x = \$20,000, z = (\$20,000 - \$25,000)/\$10,000 = -0.5, and when x = \$30,000, z = +0.5. Hence,

P(\$20,000 ≤ X ≤ \$30,000) = P(-.5 ≤ Z ≤.5) = 2F(.5) - 1 = 1.383 - 1 = .383. Thus, about 38% of the taxpayers will benefit from the new law.
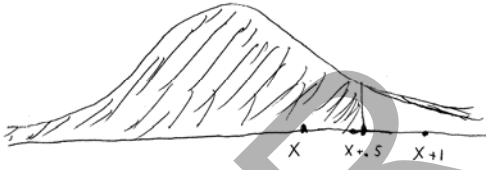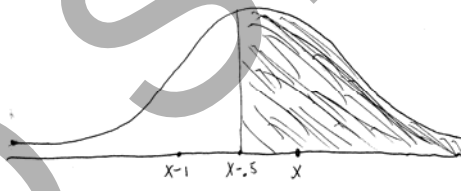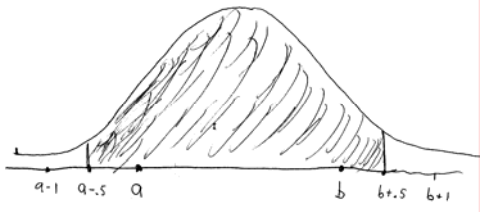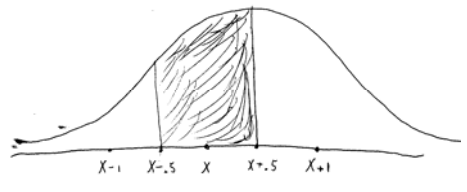
V.   The normal approximation to the binomial.

As we saw before, many interesting problems can be addressed via the binomial distribution. However, for large Ns, the binomial distribution can get to be quite awkward to work with. Fortunately, as N becomes large, the binomial distribution becomes more and more symmetric, and begins to converge to a normal distribution. That is, for a large enough N, a binomial variable X is approximately ~ N(Np, Npq). Hence, the normal distribution can be used to approximate the binomial distribution. Just how large N needs to be depends on how close p is to 1/2, and on the precision desired, but fairly good results are usually obtained when Npq ≥ 3.

Of course, a binomial variable X is not distributed exactly normal because X is not continuous, e.g. you cannot get 3.7 heads when tossing 4 coins. In the binomial, P(X ≤ a) + P(X ≥ a + 1) = 1 whenever a is an integer. But if we sum the area under the normal curve corresponding to P(X ≤ a) + P(X ≥ a + 1), this area does not sum to 1.0 because the area from a to (a + 1) is missing.
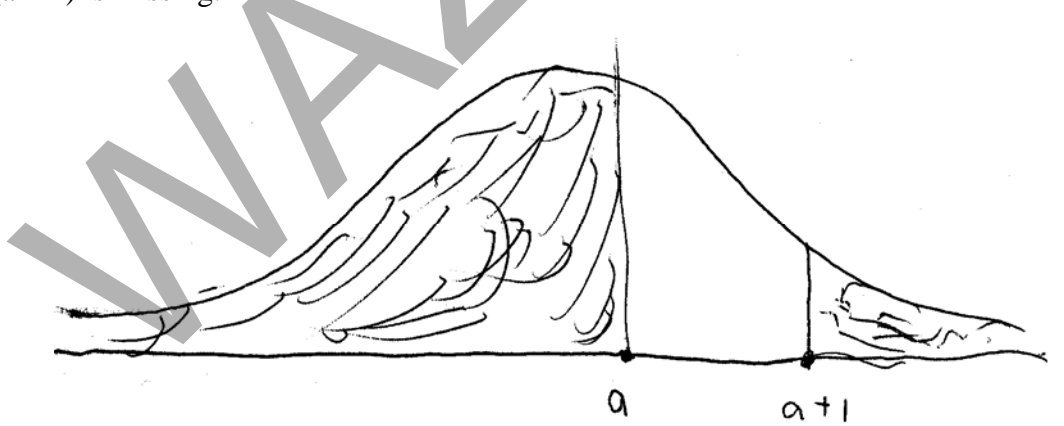
The usual way to solve this problem is to associate 1/2 of the interval from a to a + 1 with each adjacent integer. The continuous approximation to the probability P(X ≤ a) would thus be P(X ≤ a + 1/2), while the continuous approximation to P(X ≥ a + 1) would be P(X ≥ a + 1/2). This adjustment is called a correction for continuity. More specifically,

Binomial distribution	Normal approximation
$P(X \leq x) = P(X < x+1)$	$P(X \leq x + .5) = P\left(Z \leq \dfrac{x - Np + .5}{\sqrt{Npq}}\right)$
$P(X \geq x) = P(X > x - 1)$	$P(X \geq x - .5) = P\left(Z \geq \dfrac{x - Np - .5}{\sqrt{Npq}}\right)$
$P(a \leq X \leq b) = P(a - 1 < X < b + 1)$	$P(a - .5 \leq X \leq b + .5) =$ $P\left(\dfrac{a - Np - .5}{\sqrt{Npq}} \leq Z \leq \dfrac{b - Np + .5}{\sqrt{Npq}}\right)$
$P(X = x) = P(x \leq X \leq x) =$ $P(x - 1 < X < x + 1)$	$P(x - .5 \leq X \leq x + .5) =$ $P\left(\dfrac{x - Np - .5}{\sqrt{Npq}} \leq Z \leq \dfrac{x - Np + .5}{\sqrt{Npq}}\right)$

NOTE: For the binomial distribution, the values to the right of each = sign are primarily included for illustrative purposes. The equalities which hold in the binomial distribution do not hold in the normal distribution, because there is a gap between consecutive values of a. The normal approximation deals with this by "splitting" the difference.

For example, in the binomial, $P(X \leq 6) = P(X < 7)$, since 6 is the next possible value of X that is less than 7. In the normal, we approximate this by finding $P(X \leq 6.5)$. And, in the binomial, $P(X \geq 6) = P(X > 5)$, because 6 is the next value of X that is greater than 5. In the normal, we approximate this by finding $P(X \geq 5.5)$

EXAMPLES.

1. Suppose 50% of the population approves of the job the governor is doing, and that 20 individuals are drawn at random from the population. Solve the following, using both the binomial distribution and the normal approximation to the binomial.

        a.   What is the probability that exactly 7 people will support the governor?
        b.   What is the probability that 7 or fewer people will support the governor?
        c.   What is the probability that exactly 11 will support the governor?
        d.   What is the probability that 11 or fewer will support the governor?

SOLUTION: Note that $N = 20$, $p = .5$, so $\mu = Np = 10$ and $\sigma^2 = Npq = 5$, $\sigma = 2.236$. Since Npq ≥ 3, it is probably safe to assume that X has approximately a N(10,5) distribution.

        a.   For the binomial, find $P(X = 7)$. Appx. E, Table II shows $P(7) = .0739$. For the normal, find $P(6.5 \leq X \leq 7.5)$. We convert 6.5 and 7.5 to their corresponding z-scores (-1.57 and -1.12), and the problem becomes finding $P(-1.57 \leq Z \leq -1.12) = F(1.57) - F(1.12) = .9418 - .8686 = .0732$.

        b.   To use the binomial distribution, find $P(X \leq 7)$. Using Appx. E, we get
$P(7) + P(6) + P(5) + P(4) + P(3) + P(2) + P(1) + P(0) =$
$.0739 + .0370 + .0148 + .0046 + .0011 + .0002 + 0 + 0 = .1316$.
        To use the normal approximation to the binomial, find $P(X \leq 7.5)$. As noted above, the z-score that corresponds to 7.5 is -1.12. $F(-1.12) = 1 - F(1.12) = 1 - .8686 = .1314$.

        c.   For the binomial, find $P(X = 11)$. Appx. E shows $P(11) = .1602$. For the normal, find $P(10.5 \leq X \leq 11.5)$. If we convert 10.5 and 11.5 to their corresponding z-scores, the problem becomes a matter of finding $P(.22 \leq Z \leq .67) = F(.67) - F(.22) = .7486 - .5871 = .1615$.

        d.   For the binomial, find $P(X \leq 11)$. From Appx E Table 2, you can determine that this is .7483. For the normal, find $P(X \leq 11.5)$. The z-score that corresponds to 11.5 is .67, and $F(.67) = .7486$.
        In all of the above, note that the results obtained using the binomial distribution and the normal approximation to the binomial are almost identical.

2. In each of 25 races, the Democrats have a 60% chance of winning. What are the odds that the Democrats will win 19 or more races? Use the normal approximation to the binomial.

Solution. $Np = 15$, $Npq = 6$, so $X \sim N(15, 6)$. Using the normal approximation to the binomial, we want to find $P(X \geq 18.5)$.

Let $Z = (X - 15)/\sqrt{6}$. When $x = 18.5$, $z = 3.5/\sqrt{6} = 1.43$. Hence,

$P(X \geq 18.5) = P(Z \geq 1.43) = 1 - F(1.43) = 1 - .9236 = \underline{.0764}$.

Hence, Democrats have a little less than an 8% chance of winning 19 or more races.

Incidentally, note that, since $N = 25$ is not included in Appendix E, Table II, it would be very tedious to calculate this using the binomial distribution.

3. In a family of 11 children, what is the probability that there will be more boys than girls? Use the normal approximation to the binomial.

Solution. $\mu = Np = 5.5$, $\sigma^2 = Npq = 2.75$, so $X \sim N(5.5, 2.75)$. If we were using the binomial distribution, we would find $P(X \geq 6)$; since we are using the normal approximation to the binomial, we find $P(X \geq 5.5)$. Hence,

$P(X \geq 5.5) = P(Z \geq 0) = .5$. (I hope you are convinced about this question by now!)

Warning (Added September 2004): I've suggested using the Correction for Continuity ever since I first taught this course in the 1980s. Several other sources also recommend this. However, it turns out that this is fairly controversial; while the correction often produces more accurate results (as it does in all the examples I have presented here) sometimes the results are less accurate (indeed, at least one old exam has a wrong answer!). If your life really depends on getting the results exactly correct, it is better to work with a computer program that can do the calculations directly using the binomial distribution rather than the normal approximation. Stata and Excel are among the programs that can do this. For example, in Stata you could work problem #2 exactly using the `bitesti` command:

```
. bitesti 25 19 0.6, detail

        N    Observed k    Expected k    Assumed p    Observed p
------------------------------------------------------------------
       25            19            15      0.60000       0.76000

 Pr(k >= 19)            = 0.073565   (one-sided test)
 Pr(k <= 19)            = 0.970638   (one-sided test)
 Pr(k <= 11 or k >= 19) = 0.151366   (two-sided test)

 Pr(k == 19)            = 0.044203   (observed)
 Pr(k == 12)            = 0.075967
 Pr(k == 11)            = 0.043410   (opposite extreme)
```