That is, if we look for the area under the curve between any two numbers on the original scale (height in this example), this quantity differs for different normal curves.

- There is not a single answer that applies for all normal curves.

However, when looking for areas under the curve between two numbers expressed as z-scores (the number of SDs above or below the mean), then the areas are the same for any normal curve!

- Notice that for women, $\mu_X - 1\sigma_X = 65 - 1(2.5) = 62.5$ and $\mu_X = 65 + 1\sigma_X = 65 + 1(2.5) = 67.5$ correspond to z-scores of -1 and +1:

$$\frac{62.5 - \mu_X}{\sigma_X} = \frac{62.5 - 65}{2.5} = \frac{-2.5}{2.5} = -1$$
$$\frac{67.5 - \mu_X}{\sigma_X} = \frac{67.5 - 65}{2.5} = \frac{2.5}{2.5} = 1$$

- The probability of women's height falling between z-scores of ± 1 is 68.26%.

- And for men, $\mu_Y - 1\sigma_Y = 70 - 1(3) = 67$ and $\mu_Y + 1\sigma_Y = 70 + 1(3) = 73$ correspond to z-scores of -1 and +1:

$$\frac{67 - \mu_Y}{\sigma_Y} = \frac{67 - 70}{3} = \frac{-3}{3} = -1$$
$$\frac{73 - \mu_Y}{\sigma_Y} = \frac{73 - 70}{3} = \frac{3}{3} = 1$$

- The probability of men's height falling between z-scores of ± 1 is 68.26%.

The point of all of this is that by transforming to $z-$scores we can compute probabilities for any normal distribution (no matter its mean and variance) using a single reference distribution — the standard normal distribution $N(0, 1)$.

Fact: If $X \sim N(\mu, \sigma^2)$, then

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

For a normally distributed r.v. $X \sim N(\mu, \sigma^2)$ often we will be interested in finding probabilities like

$$P(c < X < d), \quad \text{or} \quad P(X \geq c), \quad \text{or} \quad P(X < c) \quad \text{etc.}$$

where c, d are given constants.

- E.g., we might want to know the percentage of women with heights between 60 and 65 inches, or the percentage of women with heights greater than or equal to 68 inches, or less than 61 inches, etc.

  How do we use $Z-$scores to get such probabilities?

Facts about inequalities:

1. $X \leq c$ if and only if $X \pm d \leq c \pm d$.
   - Here $\leq$ can be replaced by any other inequality or equality $(\geq, >, <, =)$ and the statement would still be true.
   - This means that

$$P(X \leq c) = P(X \pm d \leq c \pm d), \quad \text{e.g.,} \quad P(X \leq 4) = P(X - 3 \leq 4 - 3),$$
$$\text{and,} \quad P(X > 7) = P(X + 2 > 7 + 2).$$

2. $c \leq X \leq d$ if and only if $[c + b] \leq [X + b] \leq [d + b]$.
   - Again, $\leq$ can be replaced by any other inequality or equality $(\geq, >, <, =)$.
   - This means that

$$P(c \leq X \leq d) = P([c + b] \leq [X + b] \leq [d + b]),$$

   E.g.,

$$P(3 \leq X \leq 7) = P([3 - 2] \leq [X - 2] \leq [7 - 2]) = P(1 \leq X - 2 \leq 5),$$
$$\text{and,} \quad P(1 > X > -3) = P([1 + 9] > [X + 9] > [-3 + 9]) = P(10 > X + 9 > 6).$$

3. For c any constant and b a constant which is ≥ 0,

$$X \leq c \quad \text{if and only if} \quad bX \leq bc.$$

If b is a negative number, then multiplying by b reverses the inequality:

$$X \leq c \quad \text{if and only if} \quad bX \geq bc.$$

- Again, $\leq$ can be replaced by any other inequality or equality and the statement would still be true.
- This means that

$$P(X \leq c) = \begin{cases} P(bX \leq bc), & \text{if } b \geq 0, \\ P(bX \geq bc), & \text{if } b < 0 \end{cases}$$

E.g.,

$$P(X \leq 5) = P(3X \leq 3(5)) = P(3X \leq 15)$$

$$\text{and,} \quad P(X > -1) = P(-2X < (-2)(-1)) = P(-2X < 2).$$

4. Result 3 extends to double inequalities. That is, for $b \geq 0$,

$$c \leq X \leq d \quad \text{if and only} \quad bc \leq bX \leq bd$$

and for $b < 0$,

$$c \leq X \leq d \quad \text{if and only} \quad bc \geq bX \geq bd.$$

- Again, $\leq$ can be replaced by any other inequality or equality and the statement would still be true.
- This means that

$$P(c \leq X \leq d) = \begin{cases} P(bc \leq bX \leq bd), & \text{if } b \geq 0, \\ P(bc \geq bX \geq bd), & \text{if } b < 0 \end{cases}$$

E.g.,

$$P(3 \leq X \leq 9) = P((2)(3) \leq 2X \leq (2)(9)) = P(6 \leq 2X \leq 18)$$

$$\text{and,} \quad P(3 \geq X \geq -1) = P((-1)(3) \leq -X \leq (-1)(9)) = P(-3 \leq -X \leq -9).$$

- In summary, one can add or subtract any number or multiply or divide by any non-negative number on both (all) sides of an inequality without changing the inequality. If we multiply or divide by a negative number that switches the direction of the inequality.

These results allow us to use the standard normal distribution $N(0,1)$ to compute probabilities associated with a normal distribution $N(\mu, \sigma^2)$ for any μ and σ^2.

Examples:

i. To detect whether patients have had a stroke, one measure which is sometimes used is the cerebral blood flow (CBF) in the brain. Stroke patients tend to have lower levels of CBF than healthy patients.

Assume that in the general population, X =CBF follows a $N(75, 17^2)$ distribution. A patient is classified as "probable stroke" if his or her CBF is less than 40. *What proportion of healthy patients will be mistakenly classified as probable stroke victims?*

Answer: $X \sim N(\mu, \sigma^2)$ where $\mu = 75$, $\sigma = 17$. We want to find $P(X < 40)$:

$$\begin{aligned} P(X < 40) &= P(X - \mu < 40 - \mu) \\ &= P\left(\frac{X-\mu}{\sigma} < \frac{40-\mu}{\sigma}\right) \\ &= P\left(Z < \frac{40-\mu}{\sigma}\right) \quad \text{where } Z \sim N(0,1) \\ &= P\left(Z < \frac{40-75}{17}\right) = P(Z < -2.06) \end{aligned}$$

Now the probability $P(Z < c)$ for any number c can be computed from a computer program. For instance, in Minitab we select

Calc $\rightarrow$ Probability Distributions $\rightarrow$ Normal...

and then select "cumulative probability" (which gives the probability to the left of c), set the mean and standard deviation to 0 and 1, respectively, and input c in the field "input constant". Hitting OK gives the answer:

$$P(X < 40) = P(Z < -2.06) = .0197$$

- Note that -2.06 is just the Z score associated with 40.
- Note that Minitab allows you to set the mean and the standard deviation to anything you want. So, we actually could have computed $P(X < 40)$ here directly without transforming to Z scores by setting the mean and standard deviation to 75 and 17, respectively, and setting "input constant" to 40.
- Other computer programs also have normal probability functions. E.g., in Excel, the function NORMDIST(c,μ,σ,TRUE) gives $P(X < c)$ for $X \sim N(\mu, \sigma^2)$.
- While transforming to Z scores is not necessary with one of these computer functions, it is necessary for using a standard normal probability table.

Standard normal tables are given in a variety of formats. Some give $P(Z < c)$ for selected values $c \geq 0$, some give $P(-c < Z < c)$ for selected values $c \geq 0$, and others give $P(0 < Z < c)$ for selected values $c \geq 0$. (See handout).

Any of these formats can be used to compute any desired normal probability if we use some logic and the facts that

a. the normal distribution is symmetric, so (for example)

$$P(Z < -c) = P(Z > c) \quad \text{for any positive constant } c, \text{ and}$$

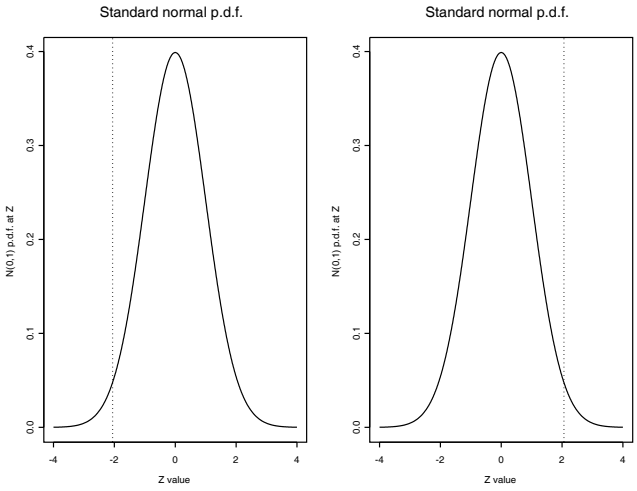
b. the area under the normal distribution is 1, so that

$$P(Z > c) = 1 - P(Z \leq c) = 1 - P(Z < c) \quad \text{for any constant } c.$$

- When computing normal probabilities from a table it is very useful to draw a picture in order to figure out exactly how to use the table and these facts to get the desired probability.

Back to the example:

We want $P(Z < -2.06)$. Picture:



- To use the first table that gives $P(Z < c)$ for $c \geq 0$ we reason as follows:

$$\begin{aligned} P(X < 40) &= P(Z < -2.06) = P(Z > 2.06) \\ &= 1 - P(Z \leq 2.06) = 1 - P(Z < 2.06) \\ &= 1 - .98030 = .0197 \end{aligned}$$

- To use the second table that gives $P(-c < Z < c)$ for $c \geq 0$ we reason as follows:

$$\begin{aligned} P(X < 40) = P(Z < -2.06) &= \frac{1}{2}\{1 - P(-2.06 < Z < 2.06)\} \\ &\approx \frac{1}{2}\{1 - P(-2.05 < Z < 2.05)\} = \frac{1}{2}(1 - .9596) = .0202 \end{aligned}$$

  which is slightly off because 2.06 didn't appear in our table and we had to use 2.05 instead.

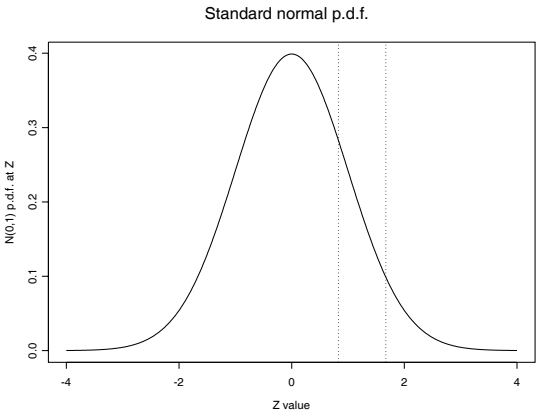
- To use the third table that gives $P(0 < Z < c)$ for $c \geq 0$ we reason as follows:

$$\begin{aligned} P(X < 40) &= P(Z < -2.06) = P(Z > 2.06) \\ &= \frac{1}{2} - P(0 < Z < 2.06) = \frac{1}{2} - .4803 = .0197 \end{aligned}$$

ii. Suppose that a mild hypertensive is defined as a person whose distolic blood pressure is between 90 and 100 mm Hg (inclusive). Suppose also that 35–44 year-old males have diastolic blood pressure which is normally distributed with mean 80 and variance 144.

What is the probability that a randomly selected 35–44 year old male is hypertensive?

I.e., if $X \sim N(80, 144)$, find $P(90 \leq X \leq 100)$.



Answer:

$$P(90 \leq X \leq 100) = P(90 - \mu \leq X - \mu \leq 100 - \mu)$$

$$= P\left(\frac{90-\mu}{\sigma} \leq \underbrace{\frac{X-\mu}{\sigma}}_{=Z} \leq \frac{100-\mu}{\sigma}\right) = P\left(\frac{90-80}{12} \leq Z \leq \frac{100-80}{12}\right)$$

$$= P(.83 \leq Z \leq 1.67) = P(Z \leq 1.67) - P(Z < .83)$$

$$= P(Z < 1.67) - P(Z < .83) = .95254 - .79673 = .15581$$

- Our book actually gives a fourth form of normal table (Table A.3 in Appendix A) which gives $P(Z > c)$ for selected values $c \geq 0$. To use that table for this problem we would notice that

$$\begin{aligned} P(90 \leq X \leq 100) &= P(Z < 1.67) - P(Z < .83) \\ &= [1 - P(Z \geq 1.67)] - [1 - P(Z \geq .83)] \\ &= (1 - .047) - (1 - .203) = .156 \end{aligned}$$

iii. Glaucoma is an eye disease characterized by high intraocular pressure (IOP). Suppose that the distribution of X =IOP in the general population is $N(\mu, \sigma^2)$ where $\mu = 16$ mm Hg, and $\sigma = 3$ mm Hg.

If the normal (i.e. healthy) range of IOP is defined as between 12 and 20 mm Hg, what percentage of the general population would fall in this range?

Answer:

$$P(12 \leq X \leq 20) = P\left(\frac{12-\mu}{\sigma} \leq \underbrace{\frac{X-\mu}{\sigma}}_{=Z} \leq \frac{20-\mu}{\sigma}\right)$$

$$= P\left(\frac{12-16}{3} \leq Z \leq \frac{20-16}{3}\right) = P(-1.33 < Z < 1.33)$$

$$= 2P(0 < Z < 1.33) = 2\left[\frac{1}{2} - P(Z \geq 1.33)\right] = 2(.5 - .092) = 0.816$$

or 81.6%.

Normal Percentiles:

Sometimes, we'd like to work backward and figure out what value of X is associated with a particular normal probability, rather than what normal probability is associated with a particular value of X, for a normal r.v. $X \sim N(\mu, \sigma^2)$.

- That is, we'd sometimes like to find the p^{th} percentile for a random variable $X \sim N(\mu, \sigma^2)$ for any given values μ and σ^2.

Fact: For $X \sim N(\mu, \sigma^2)$ the $100p^{\text{th}}$ percentile of the distribution of X (x_p, say) is related to z_p, the $100p^{\text{th}}$ percentile of the standard normal distribution, via

$$x_p = \mu + z_p\sigma. \qquad (*)$$

- Here, z_p can be looked up in a normal table like the first one in the handout by finding p in the body of the table, and then finding z_p from the margins of the table.

Examples:

iv. Recall that for 35–44 year old men, X =diastolic blood pressure follows a $N(80, 12)$ distribution. What is the 95^{th} percentile of diastolic blood pressure in this population?

We want $x_{.95}$. To get it, first find $z_{.95}$ and then use the relationship given by (*).

Using the first normal table in the handout, we look up .95 in the body of the table. .95 doesn't appear there, but .94950 and .95053 do, which gives

$$z_{.94950} = 1.64 \quad \text{and} \quad z_{.95053} = 1.65$$

Therefore, $z_{.95}$ should be about half way between 1.64 and 1.65 or

$$z_{.95} = 1.645.$$

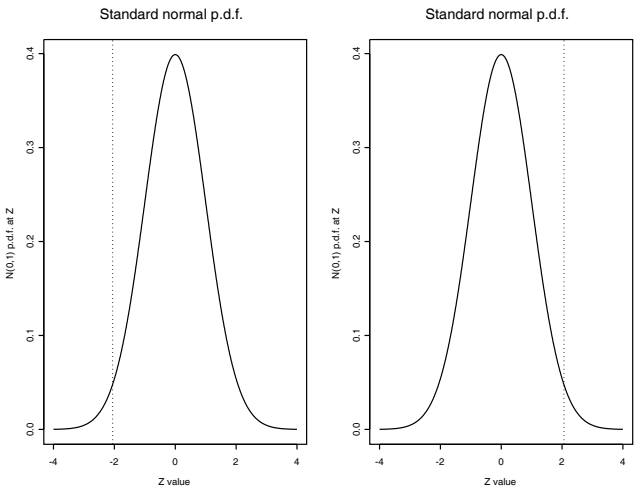
- An exact value for z_p for any p can be obtained via a computer program. For example, in Minitab we follow the steps given before, but select "Inverse cumulative probability" rather than "Cumulative probability", and then set "Input constant" to p.
- Using Minitab we can find that the exact value for $p = .95$ is $z_{.95} = 1.64485$.

Now we use the relationship (*) to get the 95th percenitle for X:

$$x_{.95} = \mu + z_{.95}\sigma = 80 + 1.64485(12) = 99.7 \quad \text{mm Hg.}$$

- Note that the table in the back of our book gives $P(Z > c)$ rather than $p = P(Z \leq c)$, but since $P(Z > c) = 1 - P(Z \leq c) = 1 - p$ we can obtain z_p from the table in our book by looking up $1 - p$ in the body of the table.
- E.g., looking up $1 - .95 = .05$ in that table, we again find that $z_{.95} = 1.645$.

v. Find the 10th percentile of diastolic blood pressure among 35–44 year old males.



From the above picture, it is clear that $z_{.10} = -z_{.90}$ or, more generally,

$$z_p = -z_{1-p}$$

Using the normal table in the back of our book, we look up .10 in the body of the table to give $z_{.90} = 1.28$, so $z_{.10} = -1.28$ and

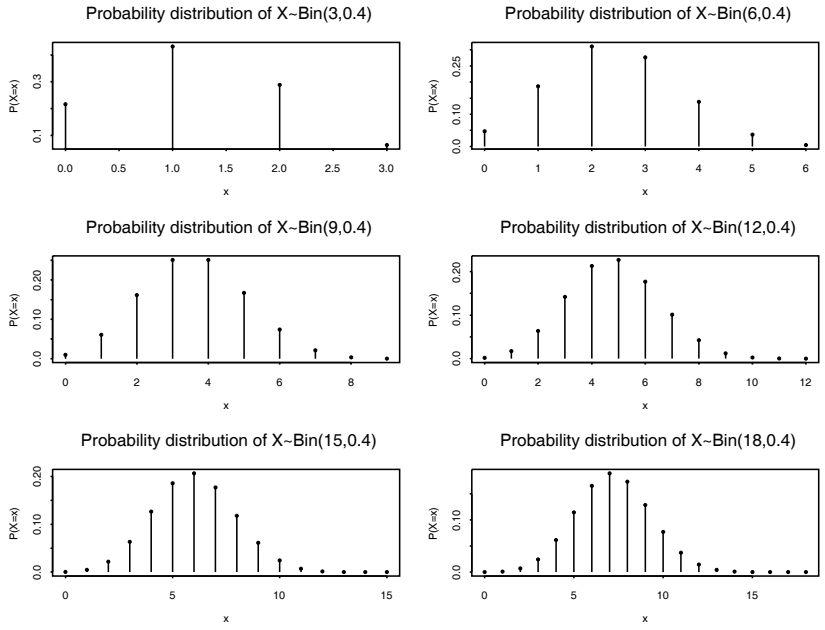
$$x_{.10} = \mu + z_{.10}\sigma = 80 + (-1.28)(12) = 64.6.$$

Normal Approximation to the Binomial:

Recall that if X = the number of successes out of n independent, identical trials with constant success probability p, then X has a binomial distribution.

- We will write this as
$$X \sim Bin(n, p)$$

Let's look at the binomial probability distribution for a particular value of p, $p = .4$, say, as n gets bigger. Below we plot the $Bin(n, p = .4)$ probability distribution for $n = 3, n = 6, n = 9, n = 12, n = 15$, and $n = 18$.



- Notice that the binomial distribution looks more and more normal as n gets large!
  - There is one important difference: binomial is discrete, normal is continuous. But this becomes less and less of a factor as n gets large, and, as we'll see, we can adjust for this difference anyway.

So, the binomial looks more and more similar to a normal distribution as n gets large, but which normal distribution is the best approximation to the distribution of $X \sim Bin(n, p)$?

The answer is: the normal distribution with the same mean and variance as X.

That is, for n large,

$$Bin(n, p) \quad \text{is well approximated by} \quad N(np, np(1-p)).$$

Example: Suppose again that 55% of UGA undergrads are women. Suppose I take a random sample of $n = 20$ undergrads. What's the probability that $X =$number of women in the sample turns out to be 12?

Based on $X \sim Bin(n, p)$ where $n = 20$, $p = .55$, we can compute this probability exactly. Using the binomial probability function,

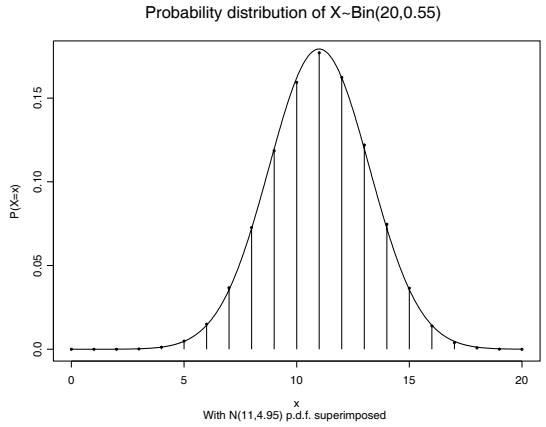
$$P(X = 12) = \binom{20}{12} .55^{12}(1 - .55)^{20-12} = .1623.$$

- Notice, though, that this is a relatively hard calculation. E.g., $20! = 2.4329 \times 10^{18}$.

Since $X \sim Bin(n, p)$, its mean and variance are

$$E(X) = np = 20(.55) = 11, \quad \text{and} \quad \text{var}(X) = np(1-p) = 4.95$$

So, the distribution of X should be well approximated by a $N(11, 4.95)$. Here's the actual binomial probability distribution with a $N(11, 4.95)$ superimposed:



Let $Y \sim N(\mu_Y, \sigma_Y^2)$, where $\mu_Y = 11$, $\sigma_Y^2 = 4.95$. Then the normal approximation to the binomial probability we want is

$$\begin{aligned} P(X = 12) \approx P(11.5 < Y < 12.5) &= P\left(\frac{11.5 - \mu_Y}{\sigma_Y} < \frac{Y - \mu_Y}{\sigma_Y} < \frac{12.5 - \mu_Y}{\sigma_Y}\right) \\ &= P\left(\frac{11.5 - 11}{\sqrt{4.95}} < Z < \frac{12.5 - 11}{\sqrt{4.95}}\right) \\ &= P(.22 < Z < .67) \\ &= P(Z < .67) - P(Z < .22) \\ &= .74857 - .58706 = .16151 \end{aligned}$$

which agrees with the true answer when rounded to three decimal places.

Another example: to find the binomial probability of 15 of more women in the sample, we would use the approximation

$$P(X \geq 15) \approx P(Y \geq 14.5), \quad \text{where} \quad Y \sim N(11, 4.95)$$

- Remember, n must be large for this approximation to work well. In fact, it will only work well if n is large and p is not too close to 0 or 1.

Rule of thumb: the normal approximation to a $Bin(n, p)$ distribution can be expected to work well if $np \geq 5$ and $n(1 - p) \geq 5$.

Sampling Distribution of the Mean*

Sampling Distributions:

Sample statistics, such as the sample mean, sample standard deviation, sample median, etc., are random variables.

Why?

- Because they are computed on a *random* sample.

Therefore, if we were to repeat the process of taking a random sample, any sample statistic (the mean, say) would vary from sample to sample, in a way that is random, because the sampling was done at random.

Of course in practice, we generally only draw one random sample, but any statistic from that sample is still a random quantity.

- Hence, any sample statistic has
  - a probability distribution,
  - an expected value, or long run average over all the possible random samples we could possibly take, and
  - a population variance, or long run variance over all of the possible random samples we could take.

The probability distribution of a sample statistic is called the **sampling distribution** of that statistic.

That sampling distribution has a (population) mean, variance, standard deviation, etc. The estimated standard deviation of a statistic is called the **standard error** of the statistic.

- Right now we will focus on the sample mean, its sampling distribution, and standard error, but it is important to realize that any statistic has a sampling distribution and standard error.

---

* Read Ch.8 of our text.

The sample mean:

- The sample mean of observations $x_1, x_2, \ldots, x_n$ has an expected value and variance that depends upon the expected value and variance of $x_1, x_2, \ldots, x_n$.
  - E.g., if $x_1, x_2, \ldots, x_n$ are big, we expect the sample mean to be big. If $x_1, x_2, \ldots, x_n$ vary a lot, we would expect their mean to be highly variable too.

Consider a random sample of observations $x_1, x_2, \ldots, x_n$, where each x_i has mean μ and variance σ^2. Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ denote the sample mean of the x_i's.

- We assume that the observations $x_1, \ldots, x_n$ are independent of each other. This is typically satisfied as a consequence of random sampling.

Then without knowing the probability distribution of the x_i's, we cannot make an exact statement about the entire probability distribution of $\bar{x}$, but we can say that the sampling distribution of $\bar{x}$

— has mean μ, and
— has variance σ^2/n.

- True for $x_1, \ldots, x_n$ drawn from any probability distribution with mean μ and variance σ^2.

These results makes sense:

- The sample mean should be centered at around the same place as the x_i's, and

- The sample mean should have variance that depends upon σ^2, the variance of the x_i's, but which also should be smaller than the variance of the x_i's.

- Notice that $\text{var}(\bar{x}) = \frac{\sigma^2}{n}$ depends on n. When the sample size is large, the sample mean has small variance.

If we know the full probability distribution of the x_i's then we can say more. In particular:

If $x_1, \ldots, x_n$ are each normally distributed with mean μ and variance σ^2 (i.e., if $x_i \sim N(\mu, \sigma^2)$ for each i) then

$$\bar{x} \sim N(\mu, \sigma^2/n).$$

Central Limit Theorem:

So, we have seen that if the x_i's have mean μ and variance σ^2, then it is *always* true that

$$\mathrm{E}(\bar{x}) = \mu, \qquad \mathrm{var}(\bar{x}) = \sigma^2/n$$

and if the x_i's are also normal, then $\bar{x}$ is normal too.

One of the most important theoreical results in statistics, the **central limit theorem** allows us to go even farther:

If the x_i's have mean μ and variance σ^2, then *regardless of the distribution of the x_i's*, their sample mean is *approximately* normally distributed if the sample size n is sufficiently large. I.e.,

$$\bar{x} \dot{\sim} N(\mu, \sigma^2/n), \quad \text{for large enough } n$$

or, if we standardize $\bar{x}$ (i.e., switch to Z scores):

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0, 1), \quad \text{for large enough } n$$

- This remarkable results is the most important reason why the normal distribution plays such a key role in statistics.
- Among other things, the CLT allows statistical inference procedures based on sample means (e.g., we typically use the sample mean to make inferences on an unknown population mean) can be based on the normal distribution (even if the original observations are not normally distributed).

Example — Body Weights

Although human heights are pretty close to normally distributed, weights are not. Especially in the US, the distribution of body weight is skewed right. That is, there are more very heavy people than there are very light people.

Suppose that among US males, the average weight is 78 kg with a standard deviation of 13 kg, and the distribution is skewed right.

Let $x_1, \ldots, x_n$ be a random sample of the weights of $n = 30$ US males and let $\bar{x} = \frac{1}{30}\sum_{i=1}^{30} x_i$ be the sample mean weight.

Then even though the weights of individual subjects are not normally distributed, the CLT implies that $\bar{x}$ is approximately distributed as

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(78, \frac{13^2}{30}\right) = N(78, 5.63).$$

Suppose we were to take samples of size $n = 30$ repeatedly, and compute the sample mean each time. What proportion of those sample means would be between 2 kg of the population mean weight (between 76 and 80 kg)?

- We can translate this question to, What is $P(76 \leq \bar{x} \leq 80)$?

Without knowing the exact distribution of weight, we don't know the exact distribution of $\bar{x}$, so we can't compute this probability exactly.

However, assuming that $n = 30$ is large enough, $\bar{x}$ is approximately $N(78, 5.63)$, so we can approximate this probability as follows:

$$\begin{aligned}
P(76 \leq \bar{x} \leq 80) &\approx P(76 - \mu \leq \bar{x} - \mu \leq 80 - \mu) \\
&= P\left(\frac{76-\mu}{\sigma/\sqrt{n}} \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq \frac{80-\mu}{\sigma/\sqrt{n}}\right) \\
&= P\left(\frac{76-78}{13/\sqrt{30}} \leq Z \leq \frac{80-78}{13/\sqrt{30}}\right) \\
&= P(-.84 \leq Z \leq .84) = 1 - 2P(Z > .84) = 1 - 2(.200) = .6
\end{aligned}$$

or approximately 60% of the samples will have sample means between 76 and 80 kg (within 2kg of the true mean).

Now what body weight cuts off the upper 5% of the sampling distribution of the sample mean, for $n = 30$?

- I.e., what is the 95th percentile of the sampling distribution of $\bar{x}$?
  - This would be the weight such that, if the x_i's each have mean 78 kg and $\sigma = 13$ kg, we would expect to observe a sample mean weight at least this small only 5% of the time.

Again, assuming that $n = 30$ is large enough for the CLT to hold, then the sampling distribution is approximately normal with mean 78 and standard deviation $13/\sqrt{30} = 2.373$.

So,

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx Z \sim N(0,1) \quad \Rightarrow \quad \bar{x} - \mu \approx Z(\sigma/\sqrt{n})$$

$$\bar{x} \approx \mu + Z(\sigma/\sqrt{n})$$

Therefore, the 95th percentile of the distribution of $\bar{x}$ is related to the 95th percentile of the standard normal distribution via

$$\bar{x}_{.95} \approx \mu + z_{.95}(\sigma/\sqrt{n}) = 78 + z_{.95}(2.373)$$

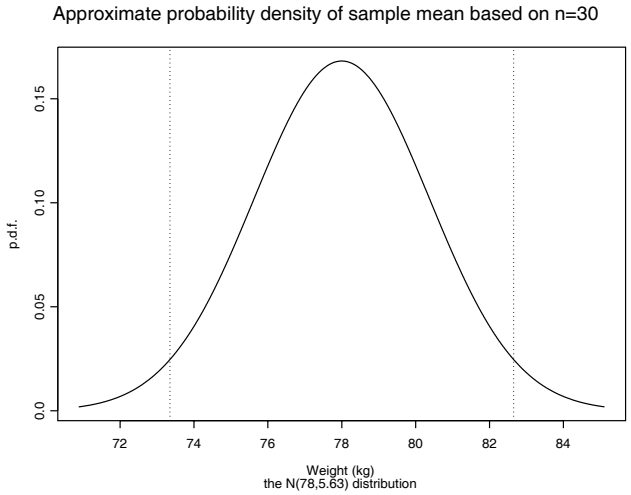
We can get $z_{.95}$ by looking up $1 - .95 = .05$ in Table A.3 in the back of our book, which yields $z_{.95} = 1.645$, so

$$\bar{x}_{.95} = 78 + (1.645)(2.373) = 81.9.$$

- So, the 95th percentile of the distribution of $\bar{x}$, the sample mean weight based on a sample of size $n = 30$ is 81.9 kg.
- This means that when taking a sample of size 30 of weights that have true mean 78 and true sd 13, 95% of the time we would expect a sample mean less than 81.9 kg.
- Based on this result, what would you conclude if you took a sample of size 30 and found the mean to be 82.4 kg (say)?
  - You'd either have gotten a very unusual sample, or
  - the sample really didn't come from a distribution with mean 78 and standard deviation 13 in the first place.

Now what weights enclose 95% of the sample means of size $n = 30$?

- That is, what are the weights (kg) such that 95% of the time we would expect to get sample mean weights between those values?



- This translates into finding $x_{.025}$ and $x_{.975}$ the 2.5th and 97.5th percentiles of the weight distribution.

By looking up .025 in Table A.3, we can find that

$$z_{.975} = 1.96 \quad \text{and} \quad z_{.025} = -1.96.$$

Therefore,

$$\bar{x}_{.975} \approx \mu + z_{.975}(\sigma/\sqrt{n}) = 78 + 1.96(2.373) = 82.65$$

and

$$\bar{x}_{.025} \approx \mu + z_{.025}(\sigma/\sqrt{n}) = 78 + (-1.96)(2.373) = 73.35$$

- So if weights have population mean 78 and population sd 13, we expect the sample mean of 30 observations to fall between 73.35 and 82.56 kg about 95% of the time.
  - Again, if we took a single sample of size 30 and calculated a sample mean outside of this range, we'd either have observed an unusual result or we might be tempted to conclude that the weights didn't have mean 78 and sd 13 in the first place.

Confidence Intervals*

Another way to look at the previous calculation is that we have used the fact that

$$P(-1.96 \leq Z \leq 1.96) = .95$$

and the CLT to infer that

$$P\left(-1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \approx .95$$

$$\Rightarrow \quad P\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) \approx .95$$

$$\Rightarrow \quad P\left(-\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx .95$$

$$\Rightarrow \quad P\left(+\bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \geq \mu \geq +\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}\right) \approx .95$$

$$\Rightarrow \quad P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx .95$$

- Because of the above probability statement, we say that the interval computed as

  $$(\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n})$$

  forms a (approximate) 95% confidence interval for μ.

- Note what is random here. μ is a population mean. It is a fixed (unknown) constant. $\bar{x}$ is random, because it is computed on a random sample.

- Therefore, we are attaching a probability to where $\bar{x}$ lies, not where μ lies.

- The interpretation here is that if we were to repeat the process by which the upper and lower limits were calculated (drawing the sample, computing the sample mean, etc.) 95% of the time we would get an interval that covers the true population mean μ.

---

* Read Ch.9 of our text.

- The confidence interval that we just introduced is an example of one of the methods of statistical inference.

Statistical Inference:

The typical paradigm for statistical inference is that we are interested in some population characteristic or parameter:
  - e.g., the average cholesterol level of 40-49 year old American females,
  - the proportion of the US voting age population that approves of the job that the president is doing,
  - the population variance in the cost of a certain medical procedure at US hospitals.

So, we collect a random sample representative of the population of interest, and use variables measured on the sample to *infer* what is true of the corresponding population parameter.

There are two main aspects of statistical inference: estimation and hypothesis testing.

1. Estimation

   a. Point estimation. In point estimation, we simply use a sample statistic to give a numerical estimate of the corresponding population value (parameter).

   - E.g., sample mean to estimation population mean, sample proportion to estimate population proportion, sample sd to estimate the population sd.

   - Good estimates should be unbiased (on target) and have small variance (be precise).

   b. Interval estimation. Almost all point estimates are likely to be wrong. They may be close to the quantity being estimated, but there is almost certainly some error (hopefully small). Confidence intervals quantify the uncertainty or error in our estimate by finding an interval within which the population parameter can be expected to lie with high probability.

   - Hopefully, that interval is narrow, meaning we're highly confident that there is little error in our estimate; i.e., it is a precise estimate.

   - Confidence intervals must be interpreted carefully.

2. Hypothesis testing. In hypothesis testing we make a decision about the population parameter based upon what we know about the corresponding sample estimate.

   - E.g., we decide whether the population mean is equal to a certain value;
   - we decide whether the population variance is equal to a certain value;
   - we decide whether two population proportions are equal to each other, etc.

   - There is always the possibility that our decision will be wrong, but in statistical hypothesis testing, we know the probability that we have made the wrong decision.

- Hypothesis testing and confidence interval estimation are really flipsides of the same coin. That is, they are two different ways to look at the problem of statistical inference.

  - They always give compatible results, but in some cases it may be more useful to frame an inference problem in terms of interval estimation and in other cases it may be more useful to conduct hypothesis tests.

Point Estimation:

A statistic T is an **unbiased** estimator of a parameter τ if

$$\mathrm{E}(T) = \tau$$

Otherwise, T is said to be biased.

- A statistic can always be thought of as an estimator of its expected value, or long run average.

- All things being equal, we would always prefer an unbiased estimator over a biased one.

The **precision** of an estimator refers to the amount of variance in its sampling distribution.

- The more variance in an estimator, the more spread out its values, the less precise it is.

Bias and precision can be understood through the following picture:

- **Accuracy** of an estimator combines bias and precision. An accurate estimator is one that has low bias and high precision.

Estimation of a population mean:

Suppose we have a random sample from a normal distribution. That is, let $x_1, \ldots, x_n$ be independent random variables, each with a $N(\mu, \sigma^2)$ distribution.

- For now, suppose we know the value of σ^2, the variance of each x_i.

Based on a sample of size n we wish to make inference on μ.

A natural estimate of μ is the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

- *Why?* Because its expected value if μ. Recall $\mathrm{E}(\bar{x}) = \mu$.
- $\bar{x}$ is a point estimate of μ.
- Because $x_1, \ldots, x_n$ were assumed normal, $\bar{x} \sim N(\mu, \sigma^2/n)$.
- Even if $x_1, \ldots, x_n$ were not normal, though, the CLT implies that $\bar{x} \dot{\sim} N(\mu, \sigma^2/n)$ (approximately normal) if n is large.

Precision of $\bar{x}$:

Remember, the precision of a statistic is related to that statistic's variance (the variance of its sampling distribution). The variance of $\bar{x}$ is $\mathrm{var}(\bar{x}) = \sigma^2/n$, so

- $\bar{x}$ is more precise when the sample size is large because that makes σ^2/n small; and
- $\bar{x}$ is more precise when σ^2, the variance of the original data, is small, because that also makes σ^2/n small.

Back on p.120, we went through some calculations to show that for $\bar{x}$ computed from a random sample with mean μ and variance σ^2

$$P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} - 1.96\frac{\sigma}{\sqrt{n}}\right) \approx .95$$

- This probability becomes exact if the sample is drawn from a normal distribution.

Therefore, we say that

$$(\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n}) = \bar{x} \pm 1.96\sigma/\sqrt{n}$$

is a 95% confidence interval for μ.

- It is an exact 95% interval for samples drawn from a normal distribution, an approximate 95% interval for samples drawn from non-normal distributions.

- Interpretation: If we were to

  - draw a random sample of size n,
  - compute $\bar{x}$
  - construct the interval $\bar{x} \pm 1.96\sigma/\sqrt{n}$
  - repeat this process many, many times,

  then 95% of these intervals will contain μ.

The precision of $\bar{x}$ is reflected in the width of interval, which is $2(1.96)\sigma/\sqrt{n}$.

- E.g., again suppose we have a random sample $x_1, \ldots, x_n$ of size n of the weights (kg) of US males, and suppose that $\mathrm{E}(x_i) = \mu$ (unknown) and $\mathrm{var}(x_i) = \sigma^2$ (known) for each i (each subject).

  Then here are the widths of approximate 95% confidence intervals for μ for different values of n and σ:

Population SD (σ)	Sample Size (n)			
	15	30	60	120
8	8.1	5.7	4.0	2.9
13	13.2	9.3	6.6	4.7
18	18.2	12.9	9.1	6.4
23	23.3	16.5	11.6	8.2

- Notice the confidence intervals get narrower (more precise) as
  - Sample size increases.
  - Population SD decreases.

- Notice that the 95% confidence is

$$\bar{x} \pm \underbrace{z_{1-.05/2}}_{=z_{.975}=1.96} \frac{\sigma}{\sqrt{n}}$$

  How do we get a 90% interval or a 99% interval?

General formula for CI for μ: For a random sample $x_1, ldots, x_n$ from a normal distribution with common mean μ and common known variance σ^2, a $100(1-\alpha)$ confidence interval for μ is given by

$$\bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

- This confidence interval is exact for normal distributions, approximate for non-normal distributions by the CLT.

- Beware that our book uses the notation z_p for the value of the standard normal that cuts off $100p\%$ in the upper tail. We use z_p to denote the value that cuts off $100p\%$ in the lower tail.

Example — Birthweights of SIDS Babies

In 1976–77 there were 78 cases of crib death (SIDS) in King Co., WA.

- The average birthweight in this sample was $\bar{x} = 2994$ g.
- Based on nationwide surveys of millions of deliveries, the mean birthweight in the US is 3300 g, with a standard deviation of 800 g.
- Suppose that this sample of $n = 78$ babies is a random sample from the total population of SIDS cases (it's not, but we'll assume so for illustration purposes).

Find a 95% confidence interval for the population mean birthweight of SIDS cases in the US.

Since we have specified that we want a 95% interval,

$$100 \times (1 - \alpha)\% = 95\% \quad \Rightarrow \alpha =$$

Therefore,

$$1 - \alpha/2 = \qquad \text{and} \qquad z_{1-\alpha/2} =$$

Thus, the 95% interval for μ is

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} =$$

$$=$$

- If we assume that birthweights are normally distributed, then this is an exact 95% CI for μ. Otherwise, it is an approximate 95% CI for μ.

- Interpretation (short version): On the basis of these data, we are 95% confident that the population mean birthweight for SIDS infants in the US is covered by the interval

$$(\qquad , \qquad)$$

- It is conventional to form 95% intervals. However, that is just tradition without any theoretical basis. Sometimes we may want other confidence levels.

Suppose we had wanted a 90% confidence interval for μ.

Then $100(1-\alpha)\% = 90\%$ which implies that

$$\alpha = \qquad \text{and} \qquad 1-\alpha/2 =$$

so that

$$z_{1-\alpha/2} =$$

Thus the 90% interval is given by

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} =$$

$$=$$

- Note that this interval is narrower than a 95% interval.
  - As the confidence level goes up, the width of the confidence interval increases as well.
  - Intuition: for me to be very highly confident that my interval covers μ, I have to make my interval wide.

One-Sided Confidence Intervals:

So far, we have just talked about two sided confidence intervals: confidence intervals with a lower and upper bound that straddle the population mean with some pre-specified probability (95%, say).

In some situations, we are interested only in finding an upper bound, which will fall above the population mean with some probability. Or perhaps, a lower bound, which falls below the mean with some probability.

Example — Cholesterol Level

High cholesterol is considered a risk factor for heart disease. There is little concern about low cholesterol levels — basically, the lower, the better. So, we might be interested in estimating the mean cholesterol level in the normal (healthy) population, and placing an upper bound on that mean, such that we can be 95% sure that the population mean falls below that upper bound.

- This would be useful for deciding whether a patient with a given cholesterol level has elevated cholesterol relative to the healthy population.
- Suppose that the population standard deviation for cholesterol level among healthy people is known to be 25 mg/dL. Cholesterol levels are known to be somewhat skewed right.
- Suppose that a random sample of 28 normal adults was obtained and the sample mean cholesterol level was 168.3 mg/dL.

Obtain a 95% upper bound on μ, the mean cholesterol level in the healthy population.

Answer: If we assume that the sample size is large enough for the CLT to hold, then

$$\bar{x} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right).$$

or, if we switch to Z scores, this statement is equivalent to

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0, 1).$$

We know that

$$P(Z \geq -1.645) = .95$$

because $-1.645 = z_{.05}$, the 5th percentile of the standard normal distribution.

Therefore,

$$\begin{aligned} .95 = P(Z \geq -1.645) &\approx P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq -1.645\right) \\ &= P\left(\bar{x} - \mu \geq -1.645\sigma/\sqrt{n}\right) \\ &= P\left(-\mu \geq -\bar{x} - 1.645\sigma/\sqrt{n}\right) \\ &= P\left(\mu \leq \bar{x} + 1.645\sigma/\sqrt{n}\right) \end{aligned}$$

That is,

$$P(\mu \leq \bar{x} + 1.645\sigma/\sqrt{n}) \approx .95 \qquad (*)$$

so that $\bar{x} + 1.645\sigma/\sqrt{n}$ is a 95% upper confidence bound for μ.

- Note that if cholesterol levels had been normal to begin with, then (*) would have been an exact equality, and our confidence bound would have been an exact 95% bound.
- Since cholesterol level was non-normal, we used the CLT to establish the approximate relationship given by (*) and our bound is an approximate 95% bound.

So, in the example, the upper bound is given by

$$\bar{x} + 1.645\sigma/\sqrt{n} = 168.3 + 1.645(25)/\sqrt{24} = 176.7$$

- So, we can be 95% confident that the population mean cholesterol level for healthy adults falls below 176.7.

The general formula for a $100(1-\alpha)\%$ upper confidence bound on μ based on a sample of size n from a population with standard deviation σ is

$$\bar{x} + z_{1-\alpha}\text{s.e.}(\bar{x}) = \bar{x} + z_{1-\alpha}\sigma/\sqrt{n}.$$

A $100(1-\alpha)\%$ lower confidence bound on μ is given by

$$\bar{x} - z_{1-\alpha}\text{s.e.}(\bar{x}) = \bar{x} - z_{1-\alpha}\sigma/\sqrt{n}.$$

<u>The case when σ is unknown:</u>

To this point, we have assumed that we know the population sd σ.

Occasionally, this may be the case, but typically, σ is unknown and must be estimated from the data just as μ must.

What should we expect this to do to our confidence intervals?

- Well if we have to estimate an additional parameter σ, one should expect that that would introduce additional uncertainty, and make our confidence intervals wider. As we'll see, this intuition is correct.

Student's t distribution:

In the case when σ was known, we based our confidence interval for μ on the fact that for a random sample from a $N(\mu, \sigma^2)$ distribution, the sample mean follows a normal distribution:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \qquad (\dagger)$$

- (This result is only approximately true for a sample from a non-normal distribution when n is sufficiently large.)

That is, in the σ known case, we considered the distribution of $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ to derive a confidence interval for μ

In the σ unknown case, therefore, a natural starting point is to consider the distribution of

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where we've replaced σ by its sample estimate, the sample standard deviation s.

This makes sense, but once we replace σ by s, this quantity no longer follows a standard normal distribution.

In fact, it can be shown that t follows a distribution that looks like the normal, but is more spread out. That distribution is called **Student's t distribution**.

- This distribution is named after Student, which is the pseudonym of the author who discovered it.

- Student's t distribution is more spread out because having to estimate σ introduces additional uncertainty (error) and makes t a more variable quantity than Z.

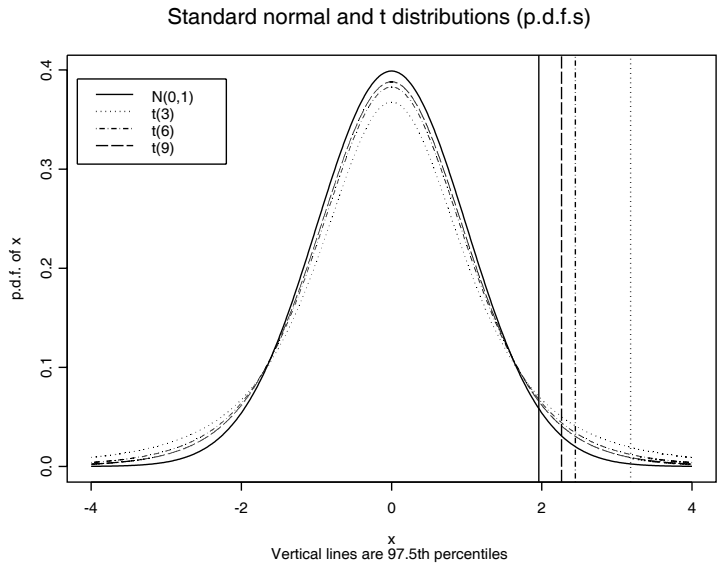
  How much more variable?

- That depends upon how precise s is as an estimate of σ, which is determined by the sample size n, or equivalently, by the divisor $n-1$ in the formula for s, which is called the **degrees of freedom** of the t distribution.

- That is, there is a distinct t distribution for every possible value of the degrees of freedom $n-1$.

  - I.e., the t distribution is a parametric distribution with parameter $n-1$, called its *degrees of freedom*.

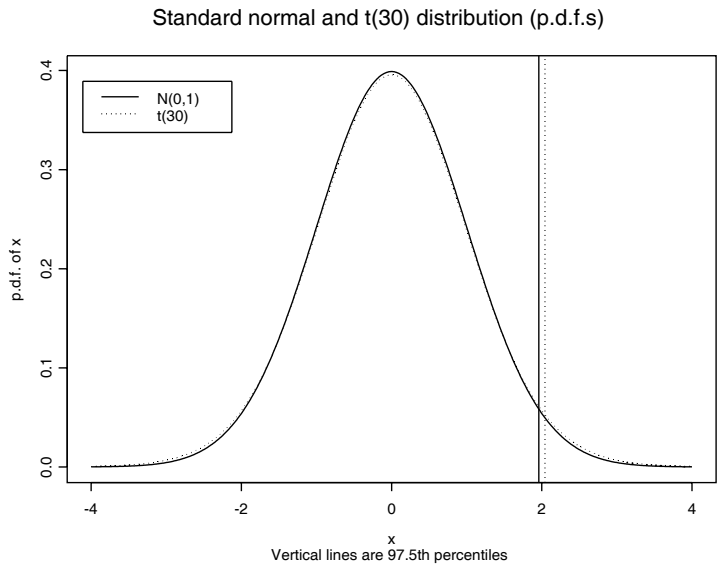
- As n grows, s becomes a better estimate of σ, and the t distribution get's less spread out relative to the normal. Here are t distributions for degrees of freedom equal to $3, 6, 9$ relative to a standard normal distribution.



- We denote the t distribution with d degrees of freedom by $t(d)$. Here we have the $t(3)$, $t(6)$, and $t(9)$ degrees of freedom as well as the $N(0,1)$.

- Notice that the spread in the t distribution decreases with the degees of freedom $n-1$.

- In fact, if $n-1$ is large enough, the $t(n-1)$ and $N(0,1)$ become almost indistinguishable. Here is the $t(30)$ compared to the $N(0,1)$:



- On these plots we've also plotted vertical lines for $z_{.975} = 1.96$, the 97.5th percentile of the standard normal distribution as well as, the corresponding 97.5th percentiles for the t distributions: $t_{.975}(3), t_{.975}(6)$, $t_{.975}(9)$, and, in the second plot, $t_{.975}(30)$.

Recall $z_{.975} = 1.96$ was the multiplier for obtaining a 95% CI for μ in the σ known case. In that case the 95% CI for μ was given by

$$\bar{x} \pm \underbrace{z_{.975}}_{=1.96} \text{s.e.}(\bar{x}) = \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

In the σ unknown case, a 95% CI for μ based on a sample of size n is given by

$$\bar{x} \pm t_{.975}(n-1)\text{s.e.}(\bar{x}) = \bar{x} \pm t_{.975}(n-1)\frac{s}{\sqrt{n}}$$

- Thus, in the σ unknown case, our multipler chages from $z_{.975}$ to $t_{.975}(n-1)$.
  - As we can see from the plots, $t_{.975}(n-1)$ is a bigger number, especially when $n-1$ is small, so we get a wider interval.

- Note that for degrees of freedom $n-1=30$, the z and t multipliers are very close. This observation has led to the often given rule of thumb that for $n-1 \geq 30$ we can use the z multiplier in place of the t to form a confidence interval for μ even when σ is unknown.
    - This replacement introduces some error in the calculation, but not much, especially for $n-1$ much larger than 30.

General Formula: In general, for a sample of size n drawn from a normally distributed population with mean μ and variance σ^2, a $100(1-\alpha)\%$ CI for μ is given by

$$\bar{x} \pm t_{1-\alpha/2}(n-1)\text{s.e.}(\bar{x}) = \bar{x} \pm t_{1-\alpha/2}(n-1)s/\sqrt{n}$$

- This interval is approximately correct even if the sample is drawn from a non-normal population, as long as the sample size is large.

Example – Lead Content in Boston Drinking Water

Recall the following data on the lead content (mg/liter) in 12 samples of drinking water in the city of Boston, MA.

$$.035, .060, .055, .035, .031, .039, .038, .049, .073, .047, .031, .016$$

Assuming that lead content is normally distributed, form a 90% CI for μ the mean lead content in Boston drinking water.

Answer: In this case, we do not know the population mean or population standard deviation, so they must be estimated from the sample data:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} = \frac{1}{n}(.035 + \cdots + .016) = .0424$$

$$s = \sqrt{\frac{1}{n-1}\{(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2\}} = \frac{1}{11}\{.035^2 + \cdots + .016^2) - 12(.0424^2)\} = .0153$$

The standard error of $\bar{x}$ is

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{.0153}{\sqrt{12}} = .00441$$

Here we want a 90% interval, so $100(1-\alpha) = 90$, or

$$\alpha = \qquad \text{and} \qquad 1 - \alpha/2 =$$

Since the population sd is unknown, we must use the t distribution to form our interval. The sample size is $n = 12$, so the appropriate degrees of freedom is $n - 1 = 12 - 1 = 11$. Going to the back of our book, Table A.4, we find

$$t_{1-\alpha/2}(n-1) = t \quad (11) =$$

Therefore, a 90% CI for μ is given by

$$\begin{aligned} \bar{x} \pm t_{1-\alpha/2}(n-1)\text{s.e.}(\bar{x}) &= .0424 \pm (\quad)(.00441) \\ &= (\quad , \quad) \end{aligned}$$

- We are 90% confident that the mean lead content in Boston drinking water lies between and .
- Note that this is an exact 90% interval because the sample was drawn from a normal population. However, this would be an approximate 90% CI if lead content was not normally distributed, provided that the sample size was large enough for the CLT to hold.

  How large does the sample size have to be for the CLT to hold?

- Tough question. It depends upon how close to normal the population was that we sampled from.
  - If we drew a sample from a very non-normal population (highly skewed and/or highly discrete) then it requires a larger sample size in order for sample means from that population to follow a normal sampling distribution.
  - If the population we drew from to begin with is nearly normal, though, a much smaller sample size may suffice.
- The sample size necessary for the CLT to hold can be quite small in cases — as small as $n = 5$ sometimes — but to be safe, we generally need samples of size 25 or more to be fairly confident that normal distribution will provide a good approximation to the sampling distribution of $\bar{x}$.

Hypothesis Testing*

The other main aspect of statistical inference (besides point and interval estimation) is hypothesis testing.

In hypothesis testing we make a decision about the true state of the population based upon what we know concerning the sample.

- This decision is guided by probability.

A good metaphor for the approach used in statistical hypothesis testing is the American legal system.

"Innocent until proven guilty" means

— we assume innocence
— we collect and examine evidence to contradict innocence
— if evidence is strongly againsth innocence (beyond a reasonable doubt) we reject innocence and conclude the alternative, guilt.
— if not, we haven't proven innocence, only failed to proved guilt and assumption of innocence is maintained.

- The prosecutor's hypothesis is that the defendant is guilty, so he/she assumes the oppositive and tries to disprove it.

In statistical hypothesis testing, the researcher plays the role of the prosecutor. His/her research hypothesis is "guilt" so he/she assumes the opposite, which is called the **null hypothesis**, and is typically represented as H_0 ("H naught").

- For example,

$$H_0 : \text{defendant is innocent}$$

or

$$H_0 : \text{no association between obesity and diabetes}$$

* Read Ch.10 of our text.

The hypothesis that the researcher is trying to prove is called the **alternative hypothesis,** denoted H_A, or sometimes, H_1.

- For example,

  $$H_A : \text{defendant is not innocent (guilty)}$$

  or

  $$H_A : \text{there is an association between obesity and diabetes}$$

- The alternative hypothesis is always framed in such a way that it is the only other possibility under consideration, if the null hypothesis is not true. That is, the alternative hypothesis is

  $$H_A : \text{not } H_0$$

Typically, we are interested in the true state of nature in the population; operationalized in terms of the true value of some parameter, or parameters.

The simplest case: we want to test a hypothesis about a population mean.

Example — Birthweights of SIDS Cases

- Based on nationwide surveys of millions of deliveries, the mean birthweight in the US is 3300 g, with a standard deviation of 800 g.
- We want to investigate whether the population mean birthweight of SIDS cases is different from that of the general population.
- Recall that in 1976–77, there were 78 SIDS cases in King County, WA. The sample mean birthweight among the King Co. cases was $\bar{x} = 2994$ g.
- We will assume that these cases are a random sample from the population of SIDS cases nationwide (strong, questionable assumption). We will also assume that SIDS birthweights are normally distributed, with population sd $\sigma = 800$, the same as in the general population.

Is the mean birthweight among SIDS cases different than in the general population?

Let μ be the population mean birthweight among SIDS cases in the US. The null hypothesis is what we want to disprove. In this case, then, our null hypothesis is

$$H_0 : \mu = 3300g$$

- The value that we assume for μ under the null hypothesis is called the **null value** for μ and is denoted as μ_0. That is, our null hypothesis is of the form

$$H_0 : \mu = \mu_0, \qquad \text{where } \mu_0 = 3300 \text{ g}$$

How about the alternative hypothesis?

Here, there are three possibilities for the truth:

$$\mu < \mu_0, \quad \mu = \mu_0, \quad \text{or} \quad \mu > \mu_0$$

In a **one-sided alternative hypothesis** situation, the researcher/analyst makes an *a priori* assumption and dismisses either $\mu < \mu_0$ or $\mu > \mu_0$ as out of the realm of possibility.

- In the SIDS example, the researcher may be willing to assume *a priori* that there is no possible way that the mean SIDS birthweight could be greater than in the general population. In that case

$$H_A : \text{not } H_0$$

becomes

$$H_A : \mu < \mu_0, \qquad \text{where } \mu_0 = 3300 \text{ g}$$

In a **two-sided alternative hypothesis** situation, the researcher/analyst makes no such *a priori* assumption, so that the alternative hypothesis becomes

$$H_A : \mu \neq \mu_0, \qquad \text{where } \mu_0 = 3300 \text{ g}$$

- We will concentrate on one-sided alternatives first, and then discuss how things change when we instead use a two-sided alternative.

Type I and II Errors

In performing a hypothesis test there are two possible states of nature and two possible conclusions that can be made:

Conclusion	State of Nature: H_0 is true	State of Nature: H_0 is false
Fail to Reject H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

We can make errors in two ways:

I We can incorrectly reject H_0 — A Type I Error
II We can incorrectly fail to reject H_0 — A Type II Error

- Ideally, we rarely make errors of wither type.

Let

$$\alpha = P(\text{we make a Type I error})$$

$$\beta = P(\text{we make a Type II error})$$

We would like to simultaneouly minimize both α and $\boldsymbol{\beta}$.

- However, the only one of these that we have complete control over is α. β depends upon how false the null hypothesis is.
  - *Why?* Because if the true value of μ is far from μ_0, its a lot easier to reject $H_0 : \mu = \mu_0$, than if μ is closer to μ_0.

- So, we construct our test in such a way that α is small.

Example — Birthweights of SIDS Cases (Continued)

Suppose that we are interested in the one-sided alternative, so we want to test

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_A : \mu < \mu_0, \quad \text{where } \mu_0 = 3300 \text{ g}$$

- That is, we're willing to dismiss the possibility that SIDS cases might have birthweights greater than the general population.

Given that we don't know μ, how do we decide in favor of H_0 or H_A?

Answer: we look at how much smaller $\bar{x}$ is than μ_0.

- If $\bar{x}$ is much smaller than $\mu_0 = 3300$ then there's strong evidence against H_0 and we reject H_0 in favor of H_A.
  - Suppose $\bar{x}$ had been 1100 g. That seems very far from $\mu_0 = 3300$, so we would have little trouble concluding in favor of H_A.
  - However, what if $\bar{x}$ had been 3250 g? That's smaller than $\mu_0 = 3300$, but is it small enough enough to conclude that $\mu < 3300$?

How much smaller than 3300 must $\bar{x}$ be before we're willing to conclude that $\mu < 3300$?

In the legal system the evidence against the null hypothesis of innocence must be "beyond a reasonable doubt."

In hypothesis testing, "beyond a reasonable doubt" is α, the probability that we reject H_0 when it is really true (the probability of convicting an innocent person).

- We set this probability low, to some pre-specified level called the **significance level** of the test.
  - The conventional choice for the significance level is $\alpha = .05$, but this is just convention. Other values such as $\alpha = .1$ or $\alpha = .01$ are also sometimes used.
- *How do set the significance level low?* By requiring $\bar{x}$ to be smaller than μ_0 by enough so that such an extreme value would be unlikely, if the null hypothesis were true.

So, we look at how unlikely it is, given that the null hypothesis is true, to have observed an $\bar{x}$ at least as far from μ_0 as the one we got.

- This probability, the probability of obtaining a result at least as unlikely as the one obtained, given that the null hypothesis is true, is called the ***p*-value** of the test.

Then if the p-value is small enough (smaller than the pre-specified significance level α), then we reject H_0.

In the SIDS example, suppose we decide to test $H_0 : \mu = 3300$ using significance level $\alpha = .05$.

- That is, we are going to require that $\bar{x}$ be fairly unusually small, something that occurs 5% of the time, assuming that the null hypothesis is true, before we decide that the null hypothesis isn't really true.

Recall that in our sample of $n = 78$ cases, the sample mean was $\bar{x} = 2994$ g. This is less than $\mu_0 = 3300$, but how unlikely is it to get a sample mean that's as small as 2994 given that the population mean is $\mu = 3300$ (given that H_0 is true)?

- That is, what's the p-value here?

Since we assumed that SIDS birthweights are $N(\mu, \sigma^2)$, where $\sigma = 800$, then based on a sample size of $n = 78$,

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(3300, \frac{800^2}{78}\right)$$

assuming that $H_0 : \mu = 3300$ is true.

Therefore, our p-value is

$$\begin{aligned}
p &= P(\bar{x} \leq 2994) = P(\bar{x} - \mu \leq 2994 - \mu) \\
&= P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq \frac{2994 - \mu}{\sigma/\sqrt{n}}\right) \\
&= P\left(Z \leq \frac{2994 - \mu}{800/\sqrt{78}}\right) \\
&= P\left(Z \leq \frac{2994 - 3300}{800/\sqrt{78}}\right) \quad \text{assuming } H_0 : \mu = 3300 \text{ is true} \\
&= P(Z \leq -3.38) = P(Z \geq 3.38) = .00036
\end{aligned}$$

- So, under the null hypothesis, we would expect to get a sample mean at least as small as the one we got with probability .00036 (or only .036% of the time).
- Therefore, either $H_0 : \mu = 3300$ is true and we've observed a very unusual event, or $H_0 : \mu = 3300$ is not true.
- Since, $p = .00036$ is less than our chosen significance level of $\alpha = .05$, its such an unusual event under H_0, that we're willing to reject $H_0 : \mu = 3300$ in favor of $H_A : \mu < 3300$.

Steps in a statistical hypothesis test:

1. State the research question in terms of the null and alternative hypotheses
   - In the previous SIDS example, $H_0 : \mu = \mu_0$ versus $H_A : \mu < \mu_0$, where $\mu_0 = 3300$ g.
2. Specify the significance level.
   - In the SIDS example, we used $\alpha = .05$.
3. Choose an appropriate test statistic.
   - In the SIDS example, since we're testing a hypothesis on the population mean μ, we based our test on the sample mean $\bar{x}$.
   - Specifically, however, we looked at how much smaller $\bar{x}$ was than μ_0 relative to the standard error of $\bar{x}$, $\sigma/\sqrt{n}$. That is, we looked at the test statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

4. Collect the data and compute the necessary sample statistics and test statistic.
   - We collected the data and computed $\bar{x}$ and then the test statistic z, which turned out to be $z = -3.38$.
5. Calculate the p-value, compare it to the significance level α, and state the conclusion. It is good practice to report not only the result of the test (reject, fail to reject) but also the numeric value of the test statistic and the numeric p-value.
   - We found that $p = .00036$, so we rejected $H_0 : \mu = 3300$ in favor of $\mu < 3300$. Our conclusion was that the population mean birthweight for SIDS cases is less than 3300 g, the mean birthweight in the general population ($z = -3.38$, $p = .00036$).

We have emphasized the p-value approach to making the decision whether to reject, or fail to reject the null hypothesis.

- Compute the p-value and reject if $p < \alpha$.

This is the preferred method of conducting the test, but you should be aware that there is another, equivalent approach for making our conclusion known as the **critical value approach**.

To understand the critical value approach, think back to our SIDS example. There, we observed $\bar{x} = 2994$, which was low relative to the null value of $\mu_0 = 3300$.

This led to a test statistic of $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = -3.38$.

Notice that if $\bar{x}$ had been closer to μ_0, then

- the test statistic would have been closer to 0,
- and the p-value would have been larger.

- E.g., if $\bar{x}$ had been 3250, say, then the test statistic would have been

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{3250 - 3300}{800/\sqrt{78}} = -.55,$$

  which has p-value $p = P(Z < -.55) = .291$, which is $> \alpha = .05$ and we would not have rejected H_0.

- Thus different values of $\bar{x}$ lead to different test statistics.

The **rejection region** of a test, is the set of values of the test statistic which lead to rejection of H_0.

- Equivalently, the set of values that lead to p values less than α.

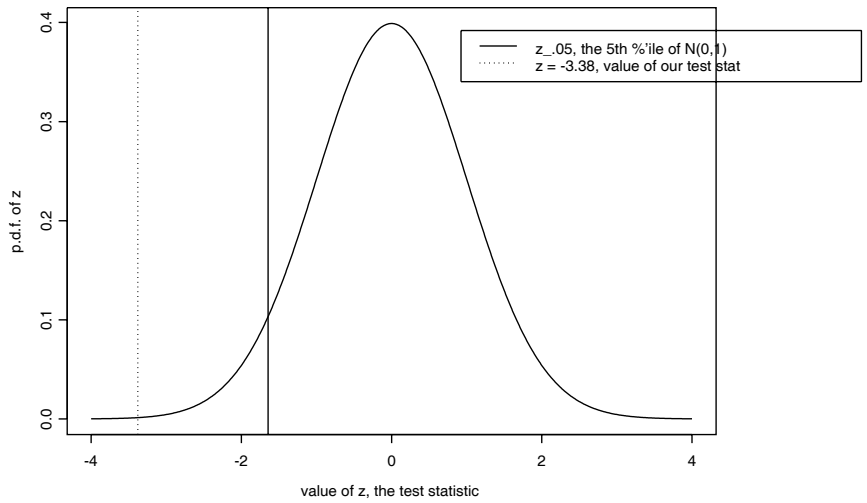
For a given level α, the **critical value** of a test statistic is the boundary of the rejection region.

- I.e., it is the value of the test statistic which is just barely large enough in magnitude to lead to rejection of H_0 at a given significance level α.

In the SIDS example, and in general for testing $H_0 : \mu = \mu_0$ versus a one-sided alternative for a normal sample with known sd σ, our test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}, \quad \text{distributed as } N(0,1) \text{ under } H_0$$

- Below is a picture of the distribution of this statistic under H_0:



The solid vertical line is $z_{.05} = -1.645$, the 5th percentile of the $N(0,1)$ distribution. That is, the area under the curve to the left of that line is .05.

- Therefore, for an $\alpha = .05$-level test, if our test statistic had turned out to < -1.645, then we would reject H_0; if it had been > -1.645 then we would have failed to reject H_0.

- Thus, $(-\infty, -1.645)$ is the rejection region of our test, $(-1.645, +\infty)$ is the acceptance region, and -1.645 is the critical value because it is the boundary of the rejection region.

Thus, instead of computing the p-value of our observed test statistic $z = -3.38$ and comparing it to $\alpha = .05$, we could instead have compared $z = -3.38$ to the critical value $z_{.05} = -1.645$. Since $z = -3.38 < z_{.05} = -1.645$, we reject H_0.

What if our alternative hypothesis had been $H_A : \mu > \mu_0$ rather than $H_A : \mu < \mu_0$?

In that case, we would have been looking for large values of our test statistic $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.

In particular, we would have rejected H_0 in favor of $H_A : \mu > \mu_0$ if

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{.95} = 1.645$$

- Notice that for either direction of the one-sided alternative we rejected H_0 if $|z| > z_{.95} = 1.645$.

General method for an α-level test of $H_0 : \mu = \mu_0$ versus a one-sided alternative based on a sample of size n from the $N(\mu, \sigma^2)$ distribution when σ^2 is known:

Critical value approach: reject H_0 if $\bar{x}-\mu_0$ is consistent with the alternative hypothesis and

$$|z| = \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{1-\alpha}.$$

Otherwise, we fail to reject.

p-value approach: reject H_0 if $p < \alpha$. Let Z denote a $N(0,1)$ random variable, and z the value of our test statistic. The p-value is computed as

$$p = \begin{cases} P(Z < z), & \text{if the alternative is } H_A : \mu < \mu_0, \\ P(Z > z), & \text{if the alternative is } H_A : \mu > \mu_0, \end{cases}$$

The Case When σ is Unknown:

If σ is unknown, the logic of testing $H_0 : \mu = \mu_0$ doesn't change at all.

However, the test statistic $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ is no longer available to us, because σ is unknown. Instead, we do the obvious thing and replace σ byits sample estimate, s.

That substitution changes our test statistic from z to t, where

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and s is the sample standard deviation.

Of course, this also changes the distribution of our test statistic. While $z \sim N(0,1)$ under H_0, $t \sim t(n-1)$ under H_0.

- This affects how we compute the p-value and critical value for our test, but not the basic logic of the testing procedure or the steps taken to implement the test.

Myocardial Infarction (Heart Attack)

- A topic of recent clinical interest is the possibility of using drugs to reduce the size of the infarct (area of tissue death due to loss of blood flow) who have had a myocardial infarction within the last 24 hours.

- Suppose we know that in untreated patients, the mean infarct size is 25 ($ck\text{-}g\text{-}EQ/m^2$). Furthermore, in 8 patients treated with a drug, the sample mean infarct size was 16 with a sample standard deviation of $s = 10$.

  Do the treated patients have smaller than average infarct size?

Let μ = population mean infarct size for patients treated with the drug. Then our hypotheses that we'd like to test are

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_A : \mu < \mu_0, \quad \text{where } \mu_0 = 25.$$

Suppose we want an $\alpha = .05$-level test.

The logic here remains the same as before. Since we're interest in a population mean μ, we examine the sample mean $\bar{x}$.

Specifically, we calculate the p-value: the probability of observing a sample mean at least as extreme (as small, in this case) as the one we got (16), under the null hypothesis that $\mu = 25$:

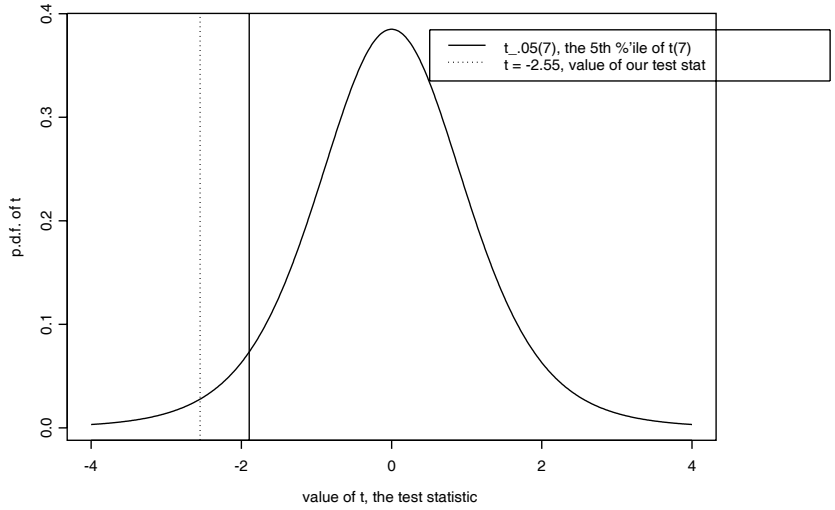
$$
\begin{aligned}
p &= P(\bar{x} \leq 16) = P(\bar{x} - \mu \leq 16 - \mu) \\
&= P\left(\underbrace{\frac{\bar{x} - \mu}{s/\sqrt{n}}}_{\sim t(n-1)} \leq \frac{16 - \mu}{s/\sqrt{n}} \right) \\
&= P\left(t(n-1) \leq \frac{16 - \mu}{s/\sqrt{n}} \right) \\
&= P\left(t(n-1) \leq \underbrace{\frac{16 - \mu_0}{s/\sqrt{n}}}_{= t, \text{ our test stat}} \right), \quad \text{assuming } H_0 : \mu = \mu_0 \text{ is true} \\
&= P\left(t(n-1) \leq \frac{16 - 25}{10/\sqrt{8}} \right) \\
&= P(t(n-1) \leq -2.55) = .0191
\end{aligned}
$$

- Here, $P(t(n-1) \leq -2.55) = .0191$ was computed in Minitab.
- Conclusion: since $p = .0191 < \alpha = .05$, we reject $H_0 : \mu = 25$ and conclude that the mean infarct size among treated patients is smaller than the average infarct size of untreated patients.

The basic steps of hypothesis testing haven't changed:

1. State the research question in terms of the null and alternative hypotheses
   - $H_0 : \mu = \mu_0$ versus $H_A : \mu < \mu_0$, where $\mu_0 = 25$.
2. Specify the significance level.
   - We used $\alpha = .05$.
3. Choose an appropriate test statistic.
   - We based our test on the sample mean $\bar{x}$ and formed a test statistic equal to
   $$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$
4. Collect the data and compute the necessary sample statistics and test statistic.
   - We collected the data and computed $\bar{x} = 16$ and then the test statistic t, which turned out to be $t = -2.55$.
5. Calculate the p-value, compare it to the significance level α, and state the conclusion.
   - We found that $p = .0191 < \alpha = .05$, so we rejected $H_0 : \mu = 25$ in favor of $\mu < 25$.

- Below is a picture of the distribution of our test statistic $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ under $H_0 : \mu = \mu_0$:



In the critical value approach, instead of comparing p to .05, we would compare our test statistic t to the critical value $t_{.05}(7)$, the 5th percentile of the t distribution on $n - 1 = 7$ degrees of freedom.

- Equivalently, we can compare $|t| = |-2.55| = 2.55$ to $t_{.95}(7)$ which is just -1 times $t_{.05}(7)$.

- From Table A.4 in the back of our book, $t_{.95}(7) = 1.895$, so since $|t| = 2.55 > 1.895$, we reject H_0 at level $\alpha = .05$.

General method for an α-level test of $H_0 : \mu = \mu_0$ versus a one-sided alternative based on a sample of size n from the $N(\mu, \sigma^2)$ distribution when σ^2 is unknown:

Critical value approach: reject H_0 if $\bar{x} - \mu_0$ is consistent with the alternative hypothesis and

$$|t| = \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| > t_{1-\alpha}(n-1).$$

Otherwise, we fail to reject.

p-value approach: reject H_0 if $p < \alpha$. Let $t(n-1)$ denote a random variable with this distribution, and t the value of our test statistic. The p-value is computed as

$$p = \begin{cases} P(t(n-1) < t), & \text{if the alternative is } H_A : \mu < \mu_0, \\ P(t(n-1) > t), & \text{if the alternative is } H_A : \mu > \mu_0, \end{cases}$$

Two-sided Alternatives:

Often, we are not willing to dismiss either $\mu > \mu_0$ or $\mu < \mu_0$, the two possible alternatives to $\mu = \mu_0$. In such cases, the appropriate set of hypotheses to test are

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad \mu \neq \mu_0$$

How does this affect our testing procedure?

Again, the answer is that it doesn't really change the logic or the steps in the procedure, it just changes how we compute the p-value or critical value.

Example — Serum Cholesterol of Asians vs. Americans

- Suppose we want to compare the mean serum-cholesterol level among recent Asian immigrants to the US with the population mean in the US.

- Suppose we assume that cholesterol levels in women aged 21-40 years in the US are normally distributed with population mean 190 mg/dL and population sd 40 mg/dL.

- Suppose that we take a random sample of $n = 100$ recent Asian immigrant women in this age range, and measure cholesterol level on these subjects. The average cholesterol level in this sample was $\bar{x} = 181.52$ mg/dL and we are willing to assume the population SD among these Asian immigrants is $\sigma = 40$, the same as it is among Americans.

Is the mean cholesterol level among recent Asian immigrant women the same as that of the coresponding general US population?

Steps for conducting hypothesis test to address this question:

1. State the research question in terms of the null and alternative hypotheses

   – Let $\mu =$ the mean cholesterol level among the Asian population. Then our hypotheses are

   $$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_A : \mu \neq \mu_0, \quad \text{where} \quad \mu_0 = 190.$$

   – This is a two-sided alternative situation because if the Asians differ from the general US population, we can't be sure that their cholesterol levels will be lower or higher.

2. Specify the significance level.

   – We'll stick with $\alpha = .05$ for now.

3. Choose an appropriate test statistic.

   - Since we're interested in μ and whether it differs from μ_0, it still makes sense to examine $\bar{x}$ and how far it differs from μ_0.
   - In addition, we know σ here.
   - Therefore, it still makes sense to base inference on the test statistic
   $$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

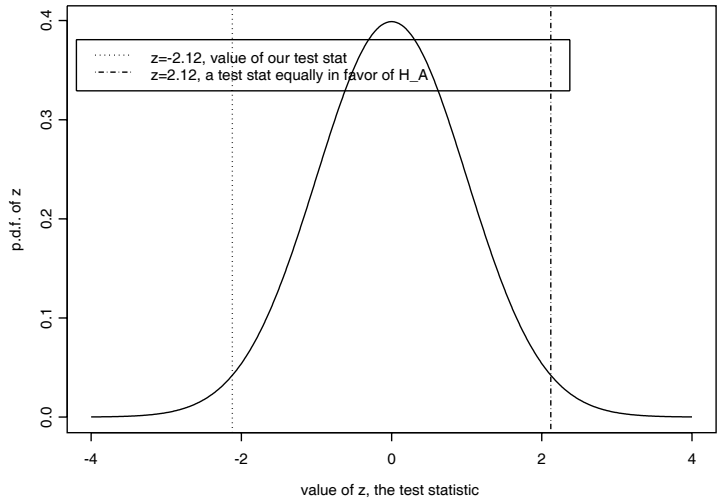
4. Collect the data and compute the necessary sample statistics and test statistic.

   - We collected the data and computed $\bar{x} = 181.52$. The test statistic is computed as
   $$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{181.52 - 190}{40/\sqrt{100}} = -2.12.$$

5. Calculate the p-value, compare it to the significance level α, and state the conclusion.

   - Here's where things differ from the one-sided alternative case.

- The p-value is the probability of getting a result at least as extreme as the one that we obtained. That is, the probability of a result which provides evidence at least as strong against the null hypothesis (in favor of the alternative). Picture:



- In the picture above, our test statistic is $t = -2.12$. Notice, that any value of the test statistic ≤ -2.12 and any value of the test statistic ≥ 2.12 would provide at least as much evidence in favor of $H_A : \mu \neq \mu_0$.

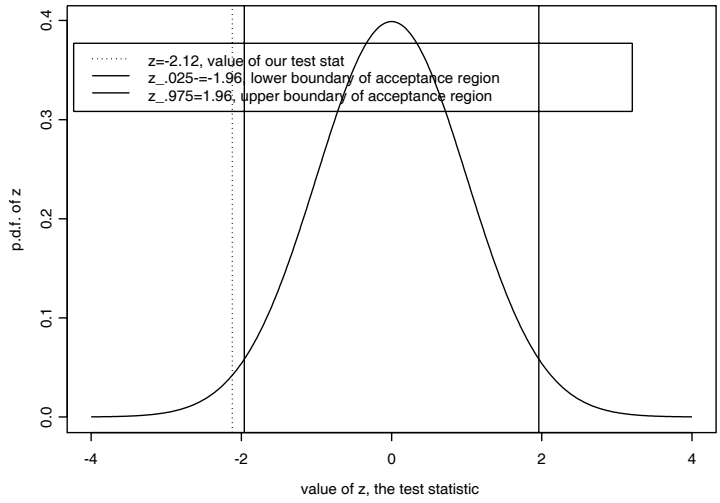
- Therefore, the p-value here is computed as

$$p = P(Z \leq -2.12) + P(Z \geq 2.12) = 2P(Z \geq 2.12) = 2(.017) = .034$$

- Notice that this is exactly twice as large as the p-value we would have obtained for a one-sided laternative $H_A : \mu < \mu_0$.

- Since $p = .034 < \alpha = .05$, we reject H_0 and conclude that recent Asian immigrant women between the ages of 21 and 40 years have different (in this case lower) mean cholesterol level that the corresponding US population ($z = -2.12$, $p = .034$).

To understand how the critical value approach differs in the two-sided alternative case, a picture is again helpful:



- The solid line at $z_{.025} = -1.96$ is the value such that 2.5% of the area under the curve falls to the left of that line.
- Since the p-value in a two-sided alternative situation is twice the probability in one-tail, a value of the test statistic equal to $z_{.025} = -1.96$ would have had a p-value of $p = 2(.025) = .05$.
- Similarly, if the value of the test statistic had been equal to $z_{.975}$ the p-value would also have been $p = 2(.025) = .05$.
- Therefore, the rejection region of our test would include all values of the test statistic $\leq z_{.025} = -1.96$ and all values $\geq z_{.975} = 1.96$.
- Thus, there are two boundaries of the rejection region and hence two critical values: $z_{.025} = -1.96$ and $z_{.975} = 1.96$.

So, based on the critical value approach, we would reject H_0 if our test statistic $z < z_{.025}$ or if $z > z_{.975}$.

Equivalently, we reject H_0 at level $\alpha = .05$ if

$$|z| > z_{.975} = z_{1-.025} = z_{1-.05/2}.$$

General method for an α-level test of $H_0 : \mu = \mu_0$ versus a two-sided alternative $H_A : \mu \neq \mu_0$ based on a sample of size n from the $N(\mu, \sigma^2)$ distribution when σ^2 is <u>known</u>:

Critical value approach: reject H_0 if

$$|z| = \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{1-\alpha/2}.$$

Otherwise, we fail to reject.

p-value approach: reject H_0 if $p < \alpha$. Let Z denote a $N(0,1)$ random variable, and z the value of our test statistic. The p-value is computed as

$$p = 2P(Z > |z|)$$

General method for an α-level test of $H_0 : \mu = \mu_0$ versus a two-sided alternative $H_A : \mu \neq \mu_0$ based on a sample of size n from the $N(\mu, \sigma^2)$ distribution when σ^2 is <u>unknown</u>:

Critical value approach: reject H_0 if

$$|t| = \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| > t_{1-\alpha/2}(n-1).$$

Otherwise, we fail to reject.

p-value approach: reject H_0 if $p < \alpha$. Let $t(n-1)$ denote a random variable distributed at $t(n-1)$, and t the value of our test statistic. The p-value is computed as

$$p = 2P(t(n-1) > |t|)$$

Example — Serum-Creatinine

- The mean serum-creatinine level measured in 12 patients 24 hours after they received a newly proposed antibiotic was 1.2 mg/dL. The sample sd was 0.6 mg/dL.

- Suppose that it is known that the general population has a mean serum-creatinine of 1.0 mg/dL.

Does the population mean serum-creatinine level among patients treated with the antibiotic differ from that of the general population?

- We assume that serum-creatinine in the population of interest is normally distributed with mean μ and unknown variance.

- We also assume that the 12 patients are randomly sampled from the population of interest (all patients given this antibiotic).

1. $$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_A : \mu \neq \mu_0, \quad \text{where} \quad \mu_0 = 1.0$$

2. For variety's sake, let's test at $\alpha = .01$ for a change.

3. We test based on $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$.

   – We reject H_0 at level α if $|t| > t_{1-\alpha/2}(n-1)$.

   – Or, equivalently, we reject H_0 at level α if $p < \alpha$.

4. $\bar{x} = 1.2$, $s = 0.6$, so our test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.2 - 1.0}{0.6/\sqrt{12}} = 1.15$$

5. The p-value is

$$p = 2P(t(n-1) > |t|) = 2P(t(11) > 1.15) = 2(.1373) = .2746$$

   Since $p > \alpha = .01$ we fail to reject H_0.

5* Equivalently, we could compare $|t|$ to the critical value

$$t_{1-\alpha/2}(n-1) = t_{1-.01/2}(11) = t_{.995}(11) = 3.106$$

Since $|t| = 1.15 < t_{.995}(11) = 3.106$, we fail to reject H_0.

- Conclusion: There is insufficient evidence to conclude that the mean serum-creatinine level among patients treated with the antibiotic differs from the mean serum-creatinine in the general population.

Power and Sample Size

Recall from our discussion of error types when conducting a statistical hypothesis test that

$\alpha = P(\text{we make a Type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$

$\beta = P(\text{we make a Type II error}) = P(\text{not reject } H_0 \text{ when } H_0 \text{ is false})$

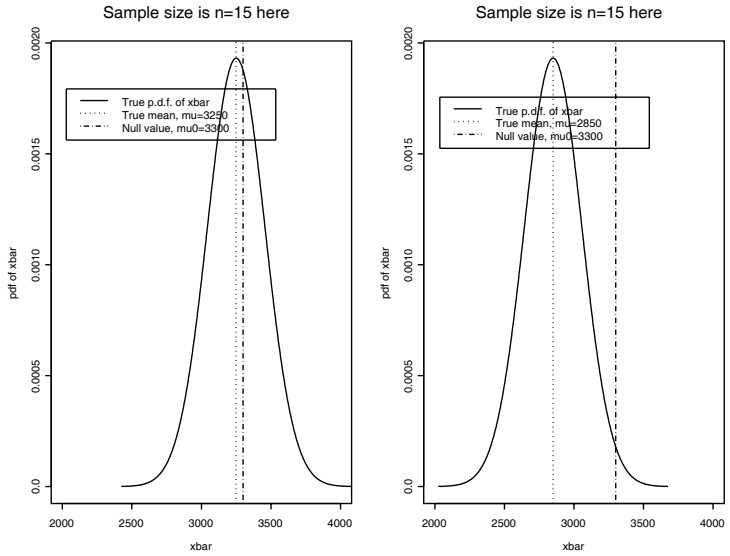
- We construct our test in such a way to ensure that α is equal to some prespecified small value (e.g., $\alpha = .05$).
- We constructed our test to control α to be small. We'd like β to be small too, but we noted that β depends upon "how false" the null hypothesis is.

Example — Birthweights of SIDS Cases

- Recall that we had a sample of n birthweights of SIDS babies with a sample mean of $\bar{x} = 2994$. We assumed that $\sigma = 800$ and we used the 1-sample z test to test

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_A : \mu < \mu_0, \quad \text{where } \mu_0 = 3300.$$

Picture:



- Here, we've assumed that the true value of μ, the mean birthweight of SIDS cases is $\mu = 3250$ on the left and $\mu = 2850$ on the right.
- We've also assumed a sample size of $n = 15$, so that the true distribution of $\bar{x}$ is

$$\bar{x} \sim N(\mu, \sigma^2/n) = N(\mu, 800^2/15) = N(\mu, 42666.67)$$

- Clearly, $\boldsymbol{\beta}$, is smaller when μ is far from μ_0 (in the plot on the right).

β is the probability of failing to reject H_0 when it is false. That is, failing to detect the truth of H_A.

- (In the current context, β is the probability of failing to detect a true difference between μ and μ_0).

Often it is more convenient to think in terms of the probability of detecting the truth of H_A (detecting a difference between μ and μ_0).

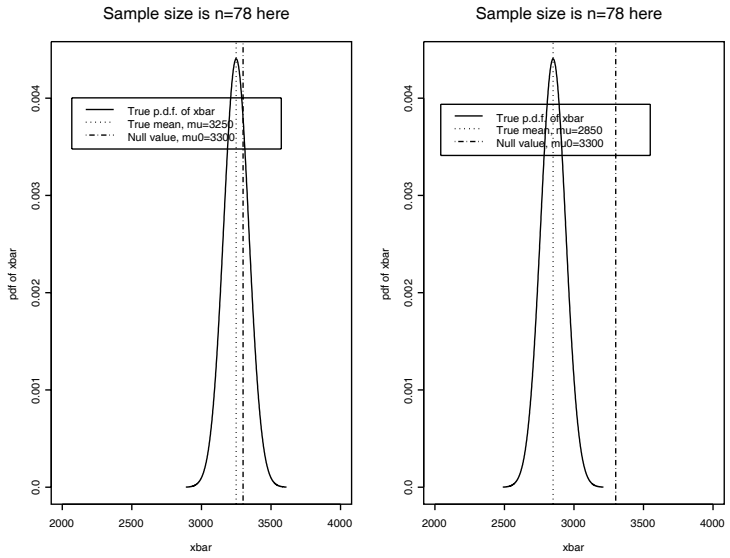
This probability is called the **power** of the test and it is simply

$$\begin{aligned}\text{power} &= P(\text{rejecting } H_0 | H_0 \text{ is false}) \\ &= 1 - P(\text{not rejecting } H_0 | H_0 \text{ is false}) = 1 - \beta\end{aligned}$$

- Thus, the further μ is from μ_0, the smaller β is and the larger the power is.
  - It is easier to reject H_0 (power is high) when H_0 is "very false" (plot on the right) than when H_0 is only slightly false (plot on the left).

As we noted, though, we can't control how false H_0 is, because we can't control the true population mean μ.

However, power also depends upon the spread in the distribution of $\bar{x}$. Suppose that instead of the picture on the previous page, we had less spread in the distribution of $\bar{x}$:



- Clearly, now it is easier to reject H_0 in both cases. This is because the spread in the distribution of $\bar{x}$ has decreased:

$$\bar{x} \sim N(\mu, \sigma^2/n) = N(\mu, 800^2/78) = N(\mu, 8205.13)$$

- That is, the less spread in the distribution of $\bar{x}$, the greater the power.
- The spread in the distribution of $\bar{x}$ is quantified by $\text{var}(\bar{x}) = \sigma^2/n$. So, this spread depends on
  - σ^2 (Power increases as σ^2 decreases.)
  - n, the sample size. (Power increases as n increases.)
- Note that we can't control σ^2, but we can control n, the sample size, when we design the study.

So, power and sample size are intimately related. A given sample size implies a certain power, and a certain power implies a certain sample size.

- Typically, at the design stage of a study, the specific hypothesis test that will be used to analyze the study is identified, and then the minimum sample size is determined so as to achieve a prespecified desired level of power.
  - Typically, it is desirable to have power of 80% or higher. Otherwise, there's a pretty good chance (20%) that we won't be able to detect the difference (effect) we are interested in even if it's real, which makes the study not worth doing.

- Of course, power depends upon a variety of other factors besides sample size. It depends on

  i. Sample size

  - The larger the sample size, the greater the power.
  - Can be controlled in the design of the study.

  ii. the true difference we are trying to detect (how false H_0 is, or the true difference $\mu - \mu_0$).

  - Bigger differences are easier to detect (result in higher power).
  - Unknown, so must be assumed.

  iii. the population SD σ.

  - The less variable the population is (smaller σ), the easier it is to detect effects (easier to to detect a signal when there's not much noise (static)).

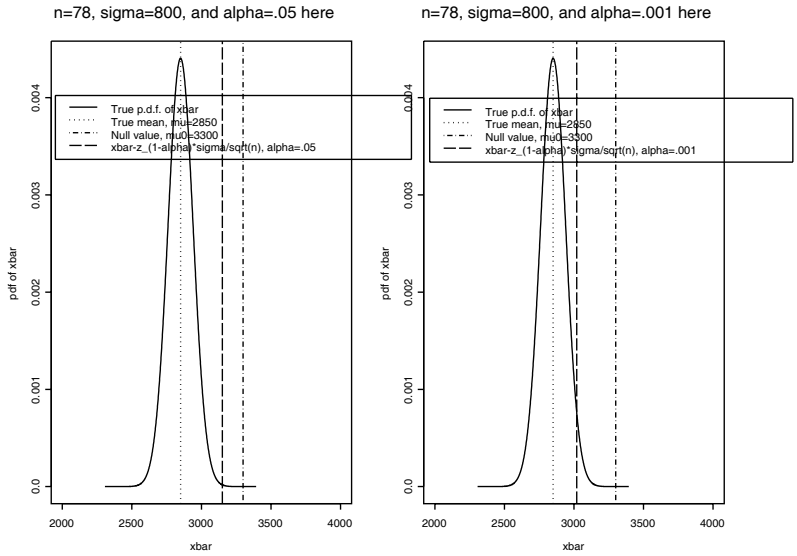
  iv. α, the significance level.

  - Similar to senstivity and specificity in diagnostic testing, there's a trade-off between α and $\boldsymbol{\beta}$ (and hence between α and power).
  - Decreasing α makes it harder to reject H_0, which decreases power (increases β).

To understand the trade-off betwen α and β, recall that in the one-sample z test of $H_0 : \mu = \mu_0$ versus $H_A : \mu < \mu_0$ we reject H_0 if

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha = -z_{1-\alpha}, \quad \text{or, equivalently, if} \quad \bar{x} < \mu_0 - z_{1-\alpha}\sigma/\sqrt{n}$$

Consider the following picture:



- In both pictures, the true mean is $\mu = 2850$, the true population SD is $\sigma = 800$, and the sample size is $n = 78$.
- In the picture on the left we are testing at $\alpha = .05$, and on the right we are testing at $\alpha = .001$.
  - Note that decreasing α makes it harder to reject $H_0 : \mu = \mu_0 = 3300$, so we need to observe a value of $\bar{x}$ which is more inconsistent with the null hypothesis. That is, we need to observe a smaller $\bar{x}$ to reject H_0 if α is small.
  - In the plot on the left, $\alpha = .05$ so we reject if

    $$\bar{x} < \mu_0 - z_{1-.05}\frac{\sigma}{\sqrt{n}} = 3300 - 1.645\frac{800}{\sqrt{78}} = 3151.01 = \text{dashed line}$$

    and on the right $\alpha = .001$, so we reject if

    $$\bar{x} < \mu_0 - z_{1-.001}\frac{\sigma}{\sqrt{n}} = 3300 - 3.090\frac{800}{\sqrt{78}} = 3020.08 = \text{dashed line}$$

- If the true population mean is $\mu = 2850$, then the bell-shaped curve in the pictures is the true p.d.f. of $\bar{x}$.
  - The area under that curve to the right of the dashed line is β, the probability of getting a value of $\bar{x}$ that would lead us to fail to reject H_0 even though it is false.
- So, as α decreases, β increases, and hence the power decreases too.

Example — Determining Power for A Proposed Study

- A new drug is proposed for people with high intraocular pressure (IOP), to prevent the development of glaucoma. A pilot study was conducted with the drug among 10 patients and their mean IOP decreased by 5 mm Hg with a SD of 10 mm Hg after 1 month of using the drug. The investigators propose to study $n = 50$ patients in the main study. What would the power of such a study be to detect a reduction of 5 mm Hg after 1 month of use of the drug?
- For now, we will assume that the true population SD is known to be 10 as obtained in the pilot study.
- We will also assume that the test to be used will be an $\alpha = .05$-level z test of $H_0 : \mu = \mu_0$ with a one-sided alternative $H_A : \mu < \mu_0$.
  - Here μ_0 is the population mean IOP among untreated subjects. Of course, this null value is known.

The power is given by

$$\begin{aligned}
\text{power} &= P(\text{reject } H_0 \text{ given that } H_0 \text{ is false and } \mu - \mu_0 = -5) \\
&= P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{1-\alpha} \middle| \mu_0 = \mu + 5\right) \\
&= P\left(\frac{\bar{x} - \mu - 5}{\sigma/\sqrt{n}} < -z_{1-\alpha}\right) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} - \frac{5}{\sigma/\sqrt{n}} < -z_{1-\alpha}\right) \\
&= P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < -z_{1-\alpha} + \frac{5}{\sigma/\sqrt{n}}\right) \\
&= P\left(Z < -z_{1-\alpha} + \frac{5\sqrt{n}}{\sigma}\right) = P\left(Z < -1.645 + \frac{5\sqrt{50}}{10}\right) \\
&= P(Z < 1.89) = 1 - P(Z \geq 1.89) = 1 - .029 = .971
\end{aligned}$$

What if we had used a two-sided alternative?

- In the IOP example suppose instead that we wished to test

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_A : \mu \neq \mu_0$$

In this case, we would reject H_0 if

$$\left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{1-\alpha/2}$$

or, equivalently, if

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{1-\alpha/2} \quad \text{or if} \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha/2}$$

Thus, the power is given by

$$\begin{aligned}
\text{power} &= P(\text{reject } H_0 \text{ given that } H_0 \text{ is false and } \mu - \mu_0 = -5) \\
&= P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{1-\alpha/2} \middle| \mu_0 = \mu + 5 \right) + P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha/2} \middle| \mu_0 = \mu + 5 \right) \\
&= P\left(\frac{\bar{x} - \mu - 5}{\sigma/\sqrt{n}} < -z_{1-\alpha/2} \right) + P\left(\frac{\bar{x} - \mu - 5}{\sigma/\sqrt{n}} > z_{1-\alpha/2} \right) \\
&= P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} - \frac{5}{\sigma/\sqrt{n}} < -z_{1-\alpha/2} \right) + P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} - \frac{5}{\sigma/\sqrt{n}} > z_{1-\alpha/2} \right) \\
&= P\left(Z < -z_{1-\alpha/2} + \frac{5\sqrt{n}}{\sigma} \right) + P\left(Z > z_{1-\alpha/2} + \frac{5\sqrt{n}}{\sigma} \right) \\
&= P\left(Z < -z_{1-\alpha/2} + \frac{5\sqrt{n}}{\sigma} \right) + P\left(Z < -z_{1-\alpha/2} - \frac{5\sqrt{n}}{\sigma} \right) \\
&= P\left(Z < -1.96 + \frac{5}{10/\sqrt{50}} \right) + P\left(Z < -1.96 - \frac{5}{10/\sqrt{50}} \right) \\
&= P(Z < 1.58) + P(Z < -5.50) \\
&= 1 - P(Z \geq 1.58) + P(Z > 5.50) = 1 - .057 + 0.000 = .943
\end{aligned}$$

- Note that the test with a one-sided alternative is more powerful than the test with a two-sided alternative.

General Result for Power of One-Sample z Test:

The power of an α-level one-sample z test of $H_0 : \mu = \mu_0$ (normal population, known population variance σ) is given by

$$\text{power} \begin{cases} = P\left(Z < -z_{1-\alpha} + \frac{|\delta|\sqrt{n}}{\sigma}\right) & \text{for a one-sided alternative} \\ = P\left(Z < -z_{1-\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) & \\ \quad + P\left(Z < -z_{1-\alpha/2} + \frac{\delta\sqrt{n}}{\sigma}\right) & \text{for a two-sided alternative} \end{cases}$$

where $\delta = \mu - \mu_0$, the difference between the true population mean μ and the null value μ_0.

- δ here is the effect we want to detect. In the example it was $\delta = -5$, a reduction of 5 mm Hg in IOP.

Sample Size:

Typically, at the design stage we fix power at a desired level and compute the sample size necessary to achieve that power rather than the other way around.

- One way to determine sample size for a given power is to use the methods we've just outlined to figure out the power for each of a range of values for n. Then select the smallest n that gives a power $\geq$ to the power we want.

- E.g., suppose we want to determine the minimum sample size necessary to ensure at least 90% power for the IOP example using a one-sided alternative and a z test.

  Then repeating the calculations of p.165 for several n values we get:

n	Power
10	.4746
15	.6147
20	.7228
25	.8038
30	.8630
35	.9054
40	.9354

  Narrowing our search we find:

n	Power
30	.8630
31	.8727
32	.8817
33	.8902
34	.8981
35	.9054

  So that we need sample size of $n = 35$ to achieve power of at least .90 (90%) given our set of assumptions.

Alternatively, we can reason as follows to solve the problem more directly (rather than by trial and error):

For a z-test with a one-sided alternative, we determined that

$$\text{power} = P\left(Z < \underbrace{-z_{1-\alpha} + \frac{|\delta|\sqrt{n}}{\sigma}}_{(*)}\right)$$

If we want power equal to p, say, then this implies that $(*)$ should be the $100p^{\text{th}}$ percentile of the Z distribution. That is,

$$z_p = -z_{1-\alpha} + \frac{|\delta|\sqrt{n}}{\sigma}$$

Solving for n we have

$$\begin{aligned} z_p + z_{1-\alpha} = \frac{|\delta|\sqrt{n}}{\sigma} \quad &\Rightarrow \quad \sqrt{n} = \frac{\sigma(z_p + z_{1-\alpha})}{|\delta|} \\ &\Rightarrow \quad n = \frac{\sigma^2(z_p + z_{1-\alpha})^2}{\delta^2}. \end{aligned}$$

- E.g., in the IOP example if we want power of $p = .90$ and if we set $\alpha = .05$, $\sigma = 10$,

$$n = \frac{10^2(z_{.90} + z_{1-.05})^2}{5^2} = \frac{100(1.2816 + 1.645)^2}{25} = 34.36 \approx 35$$

<u>General Result for Sample Size for a One-Sample z Test:</u>

The sample size necessary to achieve power equal to p for an α-level one-sample z test of $H_0 : \mu = \mu_0$ (normal population, known population variance σ) is given by

$$n = \begin{cases} \frac{\sigma^2(z_p + z_{1-\alpha})^2}{\delta^2} & \text{for a one-sided alternative} \\ \frac{\sigma^2(z_p + z_{1-\alpha/2})^2}{\delta^2} & \text{for a two-sided alternative} \end{cases}$$

where $\delta = \mu - \mu_0$, the difference between the true population mean μ and the null value μ_0.

Comparison of Two Means*

In the last two chapters, we studied how to do inference on a single population mean μ based upon a single sample of data from that population.

We now take up the problem of inference on two means μ_1 and μ_2 based upon two samples of data.

When considering inference based upon two samples, it is important to distinguish between two scenarios for which different methodologies are appropriate: Paired Samples vs. Independent Samples.

In either case, we have data that we will represent as follows:

Sample 1	Sample 2
x_{11}	x_{12}
x_{21}	x_{22}
$\vdots$	$\vdots$
$x_{n_1 1}$	$x_{n_2 2}$

1. Paired Samples.

- For paired data, the sample size is the same in each sample. That is, $n_1 = n_2 = n$.
- In addition, the first observation in sample 1 corresponds to the first observation in sample 2, the second observation in sample 1 corresponds to the second observation in sample 2, etc.
  - That is, the ith observation in samples 1 and 2 are paired, in some sense. By "paired" we mean that they are connected in such a way so that it is not reasonable to consider them to be independent random variables.

* Read Ch.11 of our text.

- Pairing can occur in many different ways. E.g.,
  - Variables x_{i1} and x_{i2} might be pretest and posttest measurements on the same patients (study involves n paitients, indexed by $i = 1, \ldots, n$).
  - Variables x_{i1} and x_{i2} might be measurements or observations taken on the same unit (e.g., x-ray) by two different observers (e.g., radiologists), or taken with two different measuring devices.
  - Variables x_{i1} and x_{i2} might be measurements of the same response variable on the same subjects at two different time points (blood pressure at time 1, time 2), or two different locations (intraocular pressure (IOP) in the right eye and left eye).
  - Variables x_{i1} and x_{i2} might be measurements of the same variable from two different family members (e.g., husband and wife, in a study involving n married couples).
- In all of these situations, we would expect that x_{i1} and x_{i2}, the measurements taken on the i^{th} subject (or pair) might be similar to one another, or statistically dependent, because of common characteristics of the subject or pair.
  - It would be reasonable to assume that observations from subject to subject (pair to pair) are independent, but that two observations from the same subject (or pair) would be dependent.

2. Independent Samples.

- Alternatively, the two samples might not be paired, and therefore, the data would be independent both within samples and between samples.
- In this situation, x_{i1} and x_{i2} are not paired in any sense (don't come from a common source), and we can have samples of different sizes. That is, n_1 is not necessarily equal to n_2.

- Independent samples are common as well. Examples include:
  - n_1 subjects randomly assigned to group 1 (e.g., they receive an active treatment) and n_2 other subjects randomly assigned to group 2 (e.g., a placebo, or control, group), and then the same response measured on each subject.
  - n subjects in the study, but n_1 subjects (selected at random) are measured at time 1 and the remaining $n_2 = n - n_1$ subjects measured at time 2.
  - Same as before, but n_1 subjects could have IOP measured in their left eye, n_2 could have IOP measured in their right eye.
  - n_1 husbands measured, n_2 wives measured from $n = n_1 + n_2$ married couples (no one in the sample married to each other).

Paired Samples:

The paired sample problem is the easier of the two because it can be handled by the methods we have already studied.

For paired data, what is typically of interest is the difference

$$\delta = \mu_1 - \mu_2$$

where μ_1 is the population mean corresponding to sample 1, and μ_2 is the population mean corresponding to sample 2.

- Notice that δ, the difference in the population means, can also be thought of as the population mean of the differences.

In a paired situation, instead of thinking about having two samples, its really more appropriate to say that we have a single sample of differences whose population mean is $\delta = \mu_1 - \mu_2$.

Example — Systolic Blood Pressure and Oral Contraceptives

- A study of the effects of taking oral contraceptives (OCs) on systolic blood pressure (SBP) was conducted in which a random sample of $n = 10$ women had their SBP measured before starting to use OCs (i.e., at baseline) and after having taken OCs for 6 months.
- The data are as follows:

Subject Number i	Sample 1 x_{i1}=Baseline SBP	Sample 2 x_{i2}=SBP using OCs	Difference d_i
1	115	128	-13
2	112	115	-3
3	107	106	1
4	119	128	-9
5	115	122	-7
6	138	145	-7
7	126	132	-6
8	105	109	-4
9	104	102	2
10	115	117	-2

- The data are paired here because samples 1 and 2 correspond to 2 measurements on the same women.
  - If a woman has high SBP at baseline, she's more likely to have relatively high SBP at the second measurement occasion, too. Therefore, these measurements are dependent.

Let

μ_1 = population mean SBP when not taking OCs,

μ_2 = population mean SBP when taking OCs,

$\delta = \mu_1 - \mu_2$

There are two types of inferences taht we might be interested in concerning δ:

- Hypothesis test: we may want to test

$$\begin{aligned} H_0 : \delta = 0 \quad &\text{versus} \quad H_A : \delta \neq 0 \\ \text{or, perhaps,} \quad &\text{versus} \quad H_A : \delta < 0 \end{aligned}$$

- Confidence interval: we may instead prefer to estimate δ and form a $100(1-\alpha)\%$ (e.g., 95%) CI for δ.

Both of these problems are ones which we already know how to handle, if we just notice that we can think of this as a one sample problem.

- Here we have a single sample of differences: $d_1, \ldots, d_n$, where

$$d_i = x_{i1} - x_{i2}, \qquad i = 1, \ldots, n$$

- We assume that the d_i's are independent, each with distribution

$$d_i \sim N(\delta = \mu_1 - \mu_2, \sigma_d^2)$$

- We estimate the population mean δ and population sd σ_d with the corresponding sample quantities:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{10}(-13 + (-3) + \cdots + (-2)) = -4.80$$

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2} = \sqrt{\frac{1}{n-1}\{(\sum_{i=1}^{n} d_i^2) - n\bar{d}^2\}}$$

$$= \sqrt{\frac{1}{9}\{(-13)^2 + \cdot + (-2)^2 - 10(-4.80)^2\}} = 4.566$$

Therefore, inference for δ can be done with the one sample methods we've already learned.

E.g., assuming that σ_d is unknown, and for a two-tailed alternative H_A : $\delta \neq 0$, we have the following t-test of $H_0 : \delta = \delta_0$:

Test statistic:

$$t = \frac{\bar{d} - \delta_0}{s_d/\sqrt{n}} = \frac{-4.80 - 0}{4.566/\sqrt{10}} = -3.32$$

Two-sided $p-$value:

$$p = 2P(t(n-1) > |t|) = 2P(t(9) > 3.32) = 2\{1 - P(t(9) < 3.32)\}$$
$$= 2\{1 - .9956\} = .0089$$

- So, at level $\alpha = .05$, we reject H_0 and conclude that there is a significant difference between the mean SBP with and without OC use. The mean SBP when using OCs is higher.

A 95% two-sided CI for δ would be given by

$$\bar{d} \pm t_{1-\alpha/2}(n-1)\frac{s_d}{\sqrt{n}} = -4.80 \pm \underbrace{t_{.975}(9)}_{=2.2622}\frac{4.566}{\sqrt{10}}$$
$$= (-8.066, -1.534)$$

- We are 95% confident that the true mean difference between the SBP at baseline and the SBP when using OCs lies between -8.066 and -1.534. A negative difference here means that the SBP at baseline is lower.
- If σ_d, the population sd of the difference between the measurements in the two samples had been known, we would have used a z-test and z-based confidence interval rather than the t-based inferences illustrated here.

Independent Samples:

In the independent samples case, we can't reduce the problem to one which we already know how to solve. Instead, we're going to need some new methodology.

We consider testing first.

As in the one-sample problem, we will assume that we have samples from normally distributed populations. If not, then our results will not hold exactly, but will be approximately valid if the sample size is reasonably large by the CLT.

In particular, we assume that for sample 1

$$x_{11}, x_{21}, \ldots, x_{n_1 1} \quad \text{are independent, with} \quad x_{i1} \sim N(\mu_1, \sigma_1^2)$$

and for sample 2

$$x_{12}, x_{22}, \ldots, x_{n_1 2} \quad \text{are independent, with} \quad x_{i2} \sim N(\mu_2, \sigma_2^2)$$

and we assume that samples 1 and 2 are independent of each other.

- That is, we have two normal samples with population means μ_1 and μ_2 and population SDs σ_1 and σ_2.
- The steps we take in conducting a hypothesis test in this setting are the same as always:

1. State the research question in terms of the null and alternative hypothesis.
   - The null hypothesis that we are interested in will be $H_0 : \mu_1 = \mu_2$, or equivalently,

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_A : \mu_1 - \mu_2 \neq 0 \quad \text{(two-sided)}$$
$$\text{or, perhaps,} \quad \text{versus} \quad H_A : \mu_1 - \mu_2 < 0 (> 0) \quad \text{(one-sided)}$$

2. Specify a significance level.
   - E.g., $\alpha = .05$.

3. Select an appropriate test statistic.

   - Since we are interested in whether $\mu_1 - \mu_2 = 0$, it is natural to examine how far $\bar{x}_1 - \bar{x}_2$ is from 0. Here $\bar{x}_1$ is the sample mean from sample 1, $\bar{x}_2$ is the sample mean from sample 2.

   - Similar to the one-sample problem, we judge how far $\bar{x}_1 - \bar{x}_2$ is from its null value, 0, relative to its standard error. That is, our test statistic is going to be of the general form:

$$\frac{\bar{x}_1 - \bar{x}_2 - 0}{\text{s.e.}(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\hat{\text{var}}(\bar{x}_1 - \bar{x}_2)}}$$

   - (Recall that the standard error of a statistic is its estimated standard deviation; i.e., the square root of its estimated variance.)

   - The exact form of this test statistic depends upon what we assume about the population SDs, σ_1 and σ_2.

   - Specifically, the standard error in the denominator of our test statistic depends upon whether σ_1 and σ_2 are assumed (i) known or unknown, and assumed (ii) equal or unequal.

4. Collect the data and compute the test statistic.

5. Calculate the p-value and make conclusion.

   - The computation of the p-value depends upon which test statistic is appropriate given our assumptions regarding σ_1 and σ_2 (step 3). Different test statistics have different distributions under H_0, which affects the p-value or critical value.

In general, under the assumptions of independent samples such that

$$\begin{array}{ll} x_{11}, x_{21}, \ldots, x_{n_1 1} & \text{are independent, with} \quad x_{i1} \sim N(\mu_1, \sigma_1^2) \\ x_{12}, x_{22}, \ldots, x_{n_2 2} & \text{are independent, with} \quad x_{i2} \sim N(\mu_2, \sigma_2^2) \end{array}$$

then

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \qquad (*)$$

i.e., $\text{var}(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

If we standardize (convert to z scores), then (*) becomes

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1) \qquad (**)$$

Under $H_0 : \mu_1 - \mu_2 = 0$, (**) becomes

$$\underbrace{\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}_{\text{test statistic}} \sim N(0, 1) \qquad (\dagger)$$

Cases:

Case 1: σ_1^2, σ_2^2 both known (may or may not be equal).

In this case, the standard error in the denominator of our test statistic above is

$$\text{s.e.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

which can be computed directly. Therefore, our test statistic and its distribution are given by (†).

Example — SBP and OC Use, Two-Sample Experiment

- Suppose that instead of the paired design described before in which each woman was measured twice, once when not using OCs and once when using OCs, the following design was used:

- A random sample of $n_1 = 8$ 35 to 39-year-old nonpregnant, premenopausal OC users and a random sample of $n_2 = 21$ 35 to 39-year-old nonpregnant, premenopausal non-OC users were obtained.

- The OC users were found to have a mean SBP of $\bar{x}_1 = 132.86$ mm Hg, and the non-OC user's were found to have a mean SBP of $\bar{x}_2 = 127.44$ mm Hg.

- σ_1, the population SD of SBP among OC users and σ_2, the population mean SD among non-OC users are assumed to be the same, equal to the common value $\sigma_1 = \sigma_2 = \sigma = 16.0$ mm Hg.

Then our test statistic is

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{132.86 - 127.44}{\sqrt{\frac{16.0^2}{8} + \frac{16.0^2}{21}}} = 0.815$$

Since our test statistic z is distributed as $N(0,1)$, the p-value for a two-sided test is

$$p = 2P(Z > .815) = 2(.207) = .414$$

and we would fail to reject $H_0 : \mu_1 = \mu_2$ based on an $\alpha = .05$ level test.

- There is insufficient evidence to conclude that the mean SBP is different for the OC users than for the non-OC users.

General Rule under Case 1:

One-sided alternative: reject H_0 if $|z| > z_{1-\alpha}$ where $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.

Equivalently, reject H_0 if $p < \alpha$ where $p = P(Z > |z|)$.

Two-sided alternative: reject H_0 if $|z| > z_{1-\alpha/2}$. Equivalently, reject H_0 if $p < \alpha$ where $p = 2P(Z > |z|)$.

Case 2: σ_1^2, σ_2^2 unknown, but assumed equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$, say).

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the test statistic in (†) becomes

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

which would still be $N(0, 1)$ if we know σ^2.

However, we don't know σ^2. Obvious thing to do: replace σ^2 by a sample estimate.

Two possible estimators come to mind:

$$s_1^2 = \text{sample variance from } 1^{\text{st}} \text{ sample}$$
$$s_2^2 = \text{sample variance from } 2^{\text{nd}} \text{ sample}$$

Under the assumption that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, both are estimators of the same quantity, σ^2, each based on only a portion of the total number of relevant observations available.

Better idea: combine these two estimators by taking their (weighted) average:

$$\hat{\sigma}^2 = s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\Rightarrow \qquad \text{s.e.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{s_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\Rightarrow \qquad \text{test stat.} = t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
$$\sim t(n_1 + n_2 - 2)$$

- Note that replacing σ^2 by an estimate s_P^2 (which is known as the **pooled estimate of** σ^2) changes the distribution of our test statistic from $N(0, 1)$ to $t(n_1 + n_2 - 2)$.

Example — SBP and OC Use, Two-Sample Experiment

- In the same set-up as before, now assume that σ_1, the population SD of SBP among OC users, and σ_2, the population SD of SBP among OC non-users, are assumed to be equal, but their common value $\sigma = \sigma_1 = \sigma_2$ is unknown.
- Suppose also that the sample SD among OC users was $s_1 = 15.34$ mm Hg, and the sample SD among OC non-users was $s_2 = 18.23$ mm Hg.

The pooled estimate of σ^2, the common variance in the two populations, is

$$s_P^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(8-1)15.34^2 + (21-1)18.23^2}{8+21-2} = 307.18$$

Therefore, our test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_P^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{132.86-127.44}{\sqrt{307.18\left(\frac{1}{8}+\frac{1}{12}\right)}} = 0.74$$

which we compare to the $t(n_1+n_2-2) = t(8+21-2) = t(27)$ distribution, the distribution of this test statistic under the null hypothesis.

For a two-sided alternative hypothesis, the p-value would be

$$\begin{aligned} p &= 2P(t(n_1+n_2-2) > |t|) = 2P(t(27) > .74) \\ &= 2\{1 - P(t(27) < .74)\} = 2(1-.7684) = .4632 \end{aligned}$$

and the critical value for a .05-level test is $t_{1-\alpha/2}(n_1+n_2-2) = t_{.975}(27) = 2.052$.

- Since $p = .4632 > \alpha = .05$ (or, equivalently, since $|t| = .74 < t_{.975}(27) = 2.052$) we fail to reject H_0. There is insufficient evidence to conclude that the mean SBP for OC users is different from that of OC non-users.

General Rule under Case 2:

One-sided alternative: reject H_0 if $|t| > t_{1-\alpha}(n_1 + n_2 - 2)$ where

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Equivalently, reject H_0 if $p < \alpha$ where $p = P(t(n_1 + n_2 - 2) > |t|)$.

Two-sided alternative: reject H_0 if $|t| > t_{1-\alpha/2}(n_1 + n_2 - 2)$. Equivalently, reject H_0 if $p < \alpha$ where $p = 2P(t(n_1 + n_2 - 2) > |t|)$.

Case 3: σ_1^2, σ_2^2 both unknown but assumed different.

In this case, the test statistic in (†):

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is not available because we don't know σ_1^2 and σ_2^2.

Obvious solution: replace σ_1^2 by s_1^2, the sample SD from the first sample, and replace σ_2^2 by s_2^2, the sample SD from the second sample.

The resulting test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Problem: even though this test statistic makes good sense, its distribution under H_0 is difficult to derive mathematically.

However, it can be shown that this test statistic has a null distribution which is well approximated by a t distribution with degrees of freedom that can be approximated from the data. That is,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \dot{\sim} t(\nu)$$

where

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1 - 1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2 - 1)}.$$

- Note that this quantity should be rounded down to the nearest integer to give an approximate degrees of freedom for the distribution of t under H_0.

- The approximation to the distribution of t under H_0 given above is based on what is known as **Satterthwaite's approximation.**

Example — SBP and OC Use, Two-Sample Experiment

- In the same set-up as before, now assume that σ_1, the population SD of SBP among OC users, and σ_2, the population SD of SBP among OC non-users, are unknown and we are not willing to assume that they are equal.

- Suppose again that the sample SD among OC users was $s_1 = 15.34$ mm Hg, and the sample SD among OC non-users was $s_2 = 18.23$ mm Hg.

In this situation, our test statistic becomes

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{132.86 - 127.44}{\sqrt{\frac{15.34^2}{8} + \frac{18.23^2}{21}}} = .8058$$

Using Sattherthwaite's approximation, this test statistic is approximately distributed as $t(\nu)$ under H_0 where

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1-1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2-1)}$$
$$= \frac{\left(\frac{15.34^2}{8} + \frac{18.23^2}{21}\right)^2}{\left(\frac{15.34^2}{8}\right)^2/(8-1) + \left(\frac{18.23^2}{21}\right)^2/(21-1)} = 15.04$$

which we round down to $\nu = 15$.

Therefore, our p-value is

$$p = 2P(t(\nu) > |t|) = 2P(t(15) > .8058)$$
$$= 2\{1 - P(t(15) < .8058)\} = 2(1 - .7835) = .433$$

and our .05-level critical value is

$$t_{1-\alpha/2}(\nu) = t_{.975}(15) = 2.131$$

Therefore, since $p = .433 > \alpha = .05$ (or, equivalently, because $|t| = .8058 < t_{.975}(15) = 2.131$) we fail to reject H_0.

- There is insufficient evidence here to conclude that the mean SBP of OC users differs from that of OC non-users.

General Rule under Case 3:

One-sided alternative: reject H_0 if $|t| > t_{1-\alpha}(\nu)$ where

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad \nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

Equivalently, reject H_0 if $p < \alpha$ where $p = P(t(\nu) > |t|)$.

Two-sided alternative: reject H_0 if $|t| > t_{1-\alpha/2}(\nu)$. Equivalently, reject H_0 if $p < \alpha$ where $p = 2P(t(\nu) > |t|)$.

Confidence Intervals for $\mu_1 - \mu_2$

As we've learned, the acceptance region of an α level test forms a $110(1-\alpha)\%$ confidence interval.

Therefore, the tests we have just derived for the two independent samples problem can all be *inverted* to form confidence intervals.

General Rule for Confidence Limits under Case 1:

One-sided limits: a $100(1-\alpha)\%$ upper confidence bound on $\mu_1 - \mu_2$ under case 1 is given by

$$(\bar{x}_1 - \bar{x}_2) + z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

A $100(1-\alpha)\%$ lower confidence bound on $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) - z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Two-sided limits: a $100(1-\alpha)\%$ confidence interval on $\mu_1 - \mu_2$ under case 1 is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

General Rule for Confidence Limits under Case 2:

One-sided limits: a $100(1-\alpha)\%$ upper confidence bound on $\mu_1 - \mu_2$ under case 2 is given by

$$(\bar{x}_1 - \bar{x}_2) + t_{1-\alpha}(n_1 + n_2 - 2)\sqrt{s_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

A $100(1-\alpha)\%$ lower confidence bound on $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) - t_{1-\alpha}(n_1 + n_2 - 2)\sqrt{s_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Two-sided limits: a $100(1-\alpha)\%$ confidence interval on $\mu_1 - \mu_2$ under case 2 is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{1-\alpha/2}(n_1 + n_2 - 2)\sqrt{s_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

General Rule for Confidence Limits under Case 3:

One-sided limits: a $100(1-\alpha)\%$ upper confidence bound on $\mu_1 - \mu_2$ under case 3 is given by

$$(\bar{x}_1 - \bar{x}_2) + t_{1-\alpha}(\nu)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

A $100(1-\alpha)\%$ lower confidence bound on $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) - t_{1-\alpha}(\nu)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Two-sided limits: a $100(1-\alpha)\%$ confidence interval on $\mu_1 - \mu_2$ under case 3 is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{1-\alpha/2}(\nu)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Notice that all of these confidence intervals are of the same general form: $\bar{x}_1 - \bar{x}_2$ plus or minus t_{crit} or z_{crit} standard errors of $\bar{x}_1 - \bar{x}_2$.

Example — Blood Glucose Level and Stenosis

- A study was performed concerning risk factors for carotid artery stenosis (narrowing) among 464 men born in 1914 and residing in the city of Malmö, Sweden. The following data were reported for blood-glucose level (mmol/L):

Stenosis Status	n	Sample Mean	Sample SD
No Stenosis	356	5.3	1.4
Stenosis	108	5.1	0.8

Using an appropriate procedure, test whether there is a significant difference between the mean blood-glucose levels of men with and without stenosis. Use $\alpha = .01$. In addition, form a 99% confidence interval for the difference in the population mean blood-glucose levels of those with and without stenosis.

Let μ_1 =population mean blood-glucose of men with stenosis, and μ_2 be the corresponding mean for those without stenosis. We are interested in testing

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_A : \mu_1 - \mu_2 \neq 0$$

and forming a 99% CI for $\mu_1 - \mu_2$.

We do not know the SDs for the two populations here, so we know that we are going to use a t test here rather than a z test. However, are we in case 2 (equal population SDs) or in case 3 (unequal population SDs)?

- To answer this question, we can choose between cases 2 and 3 based upon looking at whether the sample SDs are close to each other and by using our medical knowledge/judgement as to whether its reasonable to assume equal variability in blood glucose level in these two groups.

Alternatively, we can do a formal hypothesis test of

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad H_A : \sigma_1^2 \neq \sigma_2^2 \qquad (\ddagger)$$

There exists a statistical test of this hypothesis for data from two independent normally distributed samples. It is called the **F test for equal variances**, and it is performed as follows:

The test statistic for H_0 is given by

$$F = \begin{cases} s_1^2/s_2^2 & \text{if } s_1^2 \geq s_2^2 \\ s_2^2/s_1^2 & \text{if} s_1^2 < s_2^2 \end{cases}$$

Under H_0, this statistic follows the F **distribution**. The F distribution has two parameters, called the **numerator degrees of freedom** which is equal to one less than the sample size associated with the variance in the numerator, and the **denominator degrees of freedom** which is one less than the sample size associated with the variance in the denominator of F.

- We will denote this distribution as $F(\text{num df}, \text{denom df})$ and the $100p$th percentile by $F_p(\text{num df}, \text{denom df})$.

We reject H_0 at level α, if $F > F_{\text{crit}}$ where F_{crit} is given by

$$F_{\text{crit}} = \begin{cases} F_{1-\alpha}(n_1 - 1, n_2 - 1) & \text{if } s_1^2 \geq s_2^2 \\ F_{1-\alpha}(n_2 - 1, n_1 - 1) & \text{if } s_1^2 < s_2^2 \end{cases}$$

- Critical values of the F distribution are given in table A.5 in the back of our book.

Equivalently, we reject H_0 at level α if $p < \alpha$ where

$$p = \begin{cases} 2P(F(n_1 - 1, n_2 - 1) > F) & \text{if } s_1^2 \geq s_2^2 \\ 2P(F(n_2 - 1, n_1 - 1) > F) & \text{if } s_1^2 < s_2^2 \end{cases}$$

- Probabilitites associated with the F distribution can be computed with computer programs such as Minitab.

Back to the example:

We will conduct the t test of $H_0 : \mu_1 - \mu_2 = 0$ versus a two-sided alternative under both cases 2 and 3, but then we will do the F test for equal variances to see which case is more appropriate for these data.

Under case 2, our test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$\begin{aligned} s_P^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(356 - 1)1.4^2 + (108 - 1)0.8^2}{356 + 108 - 2} = 1.654 \end{aligned}$$

So,

$$t = \frac{5.3 - 5.1}{\sqrt{1.654\left(\frac{1}{356} + \frac{1}{108}\right)}} = \frac{.2}{.1413} = 1.416$$

and our p-value and critical value are

$$\begin{aligned} p &= 2P(t(n_1 + n_2 - 2) > |t|) = 2P(t(462) > 1.416) = 2\{1 - P(t(462) < 1.416)\} \\ &= 2(1 - .9212) = .158 \end{aligned}$$

and

$$t_{1-\alpha/2}(n_1 + n_2 - 2) = t_{.995}(462) = 2.587$$

- So, we fail to reject H_0 at level $\alpha = .01$ because $p = .1413 > \alpha = .01$ (equivalently, because $|t| = 1.416 < t_{\text{crit}} = 2.587$).

Under case 2, a 99% CI for $\mu_1 - \mu_2$ would be

$$(\bar{x}_1 - \bar{x}_2) \pm t_{1-\alpha/2}(n_1 + n_2 - 2)\sqrt{s_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = .2 \pm 2.587(.1413) = (-.165, .565)$$

Under case 3, our test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{.2}{\sqrt{\frac{1.4^2}{356} + \frac{0.8^2}{108}}} = \frac{.2}{.1069} = 1.871$$

The approximate degrees of freedom for Satterthwaite's approximation are

$$\begin{aligned}\nu &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1 - 1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2 - 1)} \\ &= \frac{\left(\frac{1.4^2}{356} + \frac{.8^2}{108}\right)^2}{\left(\frac{1.4^2}{356}\right)^2/(356 - 1) + \left(\frac{.8^2}{108}\right)^2/(108 - 1)} = 315.97\end{aligned}$$

which we round down to $\nu = 315$. Therefore, our p-value and critical value are

$$p = 2P(t(\nu) > 1.871) = 2\{1 - P(t(315) < 1.871)\} = 2\{1 - .9689\} = .062$$

and

$$t_{1-\alpha/2}(\nu) = t_{.995}(315) = 2.592$$

- So, again, we fail to reject H_0 at $\alpha = .01$ (although our p-value is now considerably smaller than in case 2).

Under case 3, a 99% CI for $\mu_1 - \mu_2$ would be

$$(\bar{x}_1 - \bar{x}_2) \pm t_{1-\alpha/2}(\nu)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = .2 \pm 2.592(.1069) = (-.077, .477)$$

Now to choose between cases 2 and 3 by conducting an F test of $H_0 : \sigma_1^2 = \sigma_2^2$.

Note that $s_1 = 1.4 > .8 = s_2$, so we compute

$$F = \frac{s_1^2}{s_2^2} = \frac{1.4^2}{.8^2} = 3.06$$

The p-value and critical value here are

$$\begin{aligned} p &= 2P(F(n_1 - 1, n_2 - 1) > F) = 2P(F(355, 107) > 3.06) \\ &= 2\{1 - P(F(355, 107) < 3.06)\} = 2\{1 - 1.000\} = 0.000 \end{aligned}$$

and

$$F_{\text{crit}} = F_{1-\alpha}(n_1 - 1, n_2 - 1) = F_{.99}(355, 107) = 1.46$$

- So, because $p = 0.000 < \alpha = .01$ (or, equivalently, because $F = 3.06 > F_{\text{crit}}$), we reject H_0, and conclude that the population variances are different here, so that the case 3 analysis was more appropriate.

Inference for Proportions*

So far we have confined our discussion of inference to means of continuous random variables. However, **dichotomous (also known as binary, or 0-1, or Bernoulli)** variables are also very common in the health sciences.

- Examples of dichotomous random variables:
  - Disease status (0=disease free, 1=diseased)
  - Mortality (0=dead, 1=alive)
  - Pregnancy (0=not pregnant, 1=pregnant)
  - Adherence to a protocol (0=no, 1=yes)
  - Gender (0=male, 1=female)
- Note that these are all essentially qualitative variables, but we assign the numbers 0 and 1 to make them numeric to allow analysis.
- Note also that the sample mean of a 0-1 variable is the proportion of the sample members who fall in the "1" category.
- A population mean of a 0-1 variable is the corresponding population proportion in the "1" category, which also has the interpretation as the probability of being in the "1" category.
- As always we can express proportions and probabilities as percentages by multiplying by 100%.

Given that a proportion is a mean, and given that the CLT says that means of even non-normally distributed random variables are approximately normal, for large sample sizes, it should be no suprise that the normal-theory inference that we have just been studied can be extended to proportions and justified as approximately valid for large sample sizes.

* Read Ch.14 of our text.

Normal Approximation to the Binomial

Recall the binomial distribution gives the probability function for a random variable X defined as the number of successes that occur out of n trials, where the trials are independent, identically distributed with constant success probability p.

- We write this as $X \sim Bin(n, p)$.
  - Recall that

$$\mathrm{E}(X) = np \qquad \mathrm{var}(X) = np(1-p)$$

Recall also from pp.111-113 of these notes that the CLT implies that the normal distribution can approximate the binomial distribution well when the sample size is large.

- *Which normal distribution?* The one with the same mean and variance as the binomial distribution that we are trying to approximate.

That is, if $np \geq 5$ and $n(1-p) \geq 5$, then for $X \sim Bin(n, p)$,

$$X \dot{\sim} N(np, np(1-p))$$

What does this have to do with inference for a proportion?

Notice that if X =the number of successes out of n trials, then the proportion of successes out of n trials is just

$$\hat{p} = X/n$$

Since $X \sim Bin(n, p) \dot{\sim} N(np, np(1-p))$ then

$$\hat{p} = \frac{X}{n} \sim \frac{1}{n} Bin(n, p) \dot{\sim} \frac{1}{n} N(np, np(1-p)) = N(p, p(1-p)/n)$$

- So, we have that

$$\hat{p} \dot{\sim} N(p, p(1-p)/n) \qquad (*)$$

  which says that a sample proportion $\hat{p}$ is approximately normally distributed with mean p, the corresponding population proportion, and variance $p(1-p)/n$.

One-Sample Confidence Intervals for p

Based on the distributional result (*), we can standardize (convert to z scores) to get the following result:

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \dot{\sim} N(0,1) \qquad (**)$$

- Therefore, for example, $z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ should fall between -1.96 and 1.96 approximately 95% of the time.
- $z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ should fall between -1.645 and 1.645 approximately 90% of the time.
- In general, $z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ should fall between $-z_{1-\alpha/2}$ and $z_{1-\alpha/2}$ approximately $100(1-\alpha)\%$ of the time.

That is, we can make the probability statement:

$$P(-1.96 \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq 1.96) \approx .95$$

If we rearrange the left-hand side so that p falls in the middle of the inequality, we get

$$P(\hat{p} - 1.96\sqrt{p(1-p)/n} \leq p \leq \hat{p} + 1.96\sqrt{p(1-p)/n}) \approx .95$$

- Therefore, $(\hat{p} - 1.96\sqrt{p(1-p)/n}, \hat{p} + 1.96\sqrt{p(1-p)/n})$ is an approximate 95% CI for p.
- Note that the endpoints of this interval depend upon p, the true value of the population proportion, which is of course unknown.
- Therefore, we replace p by $\hat{p}$, leading to

  $$(\hat{p} - 1.96\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + 1.96\sqrt{\hat{p}(1-\hat{p})/n})$$

  as an approximate 95% CI for p.

More generally, for $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$, an approximate $100(1-\alpha)\%$ CI for p is given by

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

- Notice that this interval is of the usual form: estimator plus or minus some number of standard errors.
  - Here, the standard error of $\hat{p}$ is $\sqrt{\hat{p}(1-\hat{p})/n}$ and the multiplier is the upper $\alpha/2$th critical value of a z (standard normal) distribution.

Example — Prevalence of Breast Cancer

- Suppose we are interested in estimating the prevalence (population proportion with a condition or characteristic) of breast cancer among 50–54-year old women whose mothers have had breast cancer.
- Suppose that in a random sample of 1,000 such women, 40 are found to have had breast cancer at some point in their lives.
- Obtain a point estimate and 99% confidence interval for the prevalence of breast cancer in this population.

The best point estimate of p is the sample proportion, $\hat{p} = x/n$ where x =the number with breast cancer, and n is the sample size. So, our estimate of p is

$$\hat{p} = \frac{40}{1000} = .040$$

or 4%.

To check whether the sample size is large enough in this problem to justify our normal theory confidence interval, we notice that

$$n\hat{p} = 1000(.040) = 40 \geq 5 \quad \text{and} \quad n(1-p) = 1000(1-.040) = 960 \geq 5,$$

so we should be OK.

For a 99% CI, $100(1-\alpha) = 99$ so $\alpha = .01$. Therefore,

$$z_{1-\alpha/2} = z_{.995} = 2.576 \quad \text{(back of the book)}$$

The standard error of $\hat{p}$ is

$$\text{s.e.}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{.040(1-.040)/1000} = .00620$$

so that our approximate 99% CI for p is

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} = .040 \pm 2.576(.00620) = (.024, .056)$$

- Thus, we are 99% confident that the true prevalence of breast cancer among 50–54-year-old women whose mothers had breast cancer lies between 2.4% and 5.6%.

Occasionally, we want a one-sided interval (lower or upper bound). Here is the general result:

For $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$, an approximate $100(1-\alpha)\%$ lower bound on p is given by

$$\hat{p} - z_{1-\alpha}\sqrt{\hat{p}(1-\hat{p})/n}$$

An approximate $100(1-\alpha)\%$ upper bound on p is given by

$$\hat{p} + z_{1-\alpha}\sqrt{\hat{p}(1-\hat{p})/n}$$

One-Sample Hypothesis Tests for p

- Suppose that in the breast cancer example, it is known that the population prevalence of breast cancer among women with no family history of breast cancer is 2%.
- Then to determine whether a family history of breast cancer is a risk factor for this disease, we may be interested in testing the hypothesis

$$H_0 : p = p_0 \quad \text{versus} \quad H_A : p > p_0, \quad \text{where } p_0 = .02.$$

How can we test such a hypothesis?

Recall from (**) that

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \dot{\sim} N(0,1)$$

where p is the true population proportion.

Under the null hypothesis, $p = p_0$ so this result becomes

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \dot{\sim} N(0,1)$$

- Since z compares the sample proportion $\hat{p}$ to the null value p_0 (relative to the standard error of $\hat{p}$), and since the distribution of z is known, z is the natural test statistic for testing $H_0 : p = p_0$.

General method for an approximate α-level test of $H_0 : p = p_0$ versus a one- or two-sided alternative:

Critical value approach: reject H_0 if $\hat{p} - p_0$ is consistent with the alternative hypothesis and if

$$|z| = \left| \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} \right| > \begin{cases} z_{1-\alpha} & \text{for a one-sided alternative} \\ z_{1-\alpha/2} & \text{for a two-sided alternative.} \end{cases}$$

Otherwise, we fail to reject.

p-value approach: reject H_0 if $p < \alpha$. The p-value is computed as

$$p = \begin{cases} P(Z < z), & \text{if the alternative is } H_A : p < p_0, \\ P(Z > z), & \text{if the alternative is } H_A : p > p_0, \\ 2P(Z > |z|) & \text{if the alternative is } H_A : p \neq p_0 \end{cases}$$

Here, Z denotes a $N(0,1)$ random variable, and z is the value of our test statistic.

- This normal-theory test can be justified by the CLT, and should work well provided that $np_0 \geq 5$ and $n(1-p_0) \geq 5$.

Example — Breast Cancer Prevalence

Suppose we wish to conduct an $\alpha = .01$-level test of

$$H_0 : p = p_0 \quad \text{versus} \quad H_A : p > p_0, \quad \text{where } p_0 = .02.$$

- Our test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{.040 - .02}{\sqrt{.02(1-.02)/1000}} = 4.52$$

- Since $\hat{p} = .040 > p_0 = .02$, the sample results provide evidence in favor of $H_A : p > .02$.

- Our critical value here is $z_{1-.01} = z_{.99} = 2.327$, so since $|z| = 4.52 > z_{.99} = 2.327$, we reject H_0 in favor of $H_A : p > p_0$.

- The conclusion is that there is a significantly higher prevalence (at level .01) for women whose mothers had breast cancer.

- The p-value for our test would be

$$p = P(Z > z) = P(Z > 4.52) = .0000031 \quad \text{(from Minitab)}$$

Power and Sample Size for Testing a Proportion

We have already studied power and sample size calculation methods for one-sample z tests.

Therefore, when using normal-approximation methods (z tests) for inference on p, a population proportion, the power and sample size methods we've already learned apply with little modification.

Example — Breast Cancer Prevalence

- Suppose we wish to investigate whether women who's sisters have a history of breast cancer are at higher risk for breast cancer themselves.
- Suppose we assume that the prevalence of breast cancer is 2% among 50–54 year-old US women with no famiuly history, whereas it is 5% among those women who's sisters have had breast cancer.
- We propose to interview 500 50-54 year-old women with a sister history of the disease.
- Assuming that we conduct a one-sided test at $\alpha = .05$, what would be the power of such a study?

Here, we are going to test

$$H_0 : p = p_0 \quad \text{versus} \quad H_A : p > p_0, \quad \text{where } p_0 = .02$$

This hypothesis would be rejected if our test statistic exceeds the appropriate critical value. That is, if

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > z_{1-\alpha} = z_{.95} = 1.645$$

We have assumed that the null hypothesis is really false and that the true prevalence is $p = p_1$ where $p_1 = .05$. Therefore, the power is the probability that the test statistic z exceeds the critical value $z_{1-\alpha} = 1.645$

given that $p = p_1 = .05$. That is,

$$\begin{aligned}
\text{power} &= P\left(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > z_{1-\alpha} \middle| p = p_1\right) \\
&= P\left(\hat{p} > p_0 + z_{1-\alpha}\sqrt{p_0(1-p_0)/n} \middle| p = p_1\right) \\
&= P\left(\frac{\hat{p} - p_1}{\sqrt{p_1(1-p_1)/n}} > \frac{p_0 + z_{1-\alpha}\sqrt{p_0(1-p_0)/n} - p_1}{\sqrt{p_1(1-p_1)/n}} \middle| p = p_1\right) \\
&= P\left(Z > z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} + \frac{p_0 - p_1}{\sqrt{p_1(1-p_1)/n}}\right)
\end{aligned}$$

So, in this example, the power is

$$\begin{aligned}
\text{power} &= P\left(Z > z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} + \frac{p_0 - p_1}{\sqrt{p_1(1-p_1)/n}}\right) \\
&= P\left(Z > 1.645\sqrt{\frac{.02(1-.02)}{.05(1-.05)}} + \frac{.02 - .05}{\sqrt{.05(1-.05)/500}}\right) \\
&= P(Z > -2.02) = 1 - P(Z > 2.02) = 1 - .022 = .978
\end{aligned}$$

General result for the power of a one-sample z test for p:

$$\text{power} = P(Z > \tilde{z}) = P(Z < -\tilde{z})$$

where

$$\tilde{z} = \begin{cases} z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} - \frac{|p_0 - p_1|}{\sqrt{p_1(1-p_1)/n}} & \text{if the alternative is one-sided} \\ z_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} - \frac{|p_0 - p_1|}{\sqrt{p_1(1-p_1)/n}} & \text{if the alternative is two-sided} \end{cases}$$

- This result holds provided that the sample size is large enough to justify using the normal approximation (z test). That is, provided that $np_0 \geq 5$ and $n(1-p_0) \geq 5$.