Calculating the Risk of Type II Error (beta risk)

When a test on the mean of a population is carried out, there are four possible results: two possible conditions in reality (Null is true or alternative is true) times two possible test results (reject the null or fail to reject the null).

		Decision(Action)	
		Accept H_0	Reject H_0
State of Nature	H_0 True	Correct Decision	Type I error
	H_0 False	Type II Error	Correct Decision

Our *level of significance* in a test is the acceptable level of risk for type I error. The p-value is the risk of type I error with samples of the type taken. The risk of type II error is harder to work with, and enters into the design of tests and the planning of studies more than into direct decisions once the data are collected.

Both these risks are conditional probabilities. They do not give "the probability of a correct/incorrect decision".
Risk of Type I error = $P(\text{reject } H_0 | H_0 \text{ is true})$ [To actually find $P(\text{making type I error})$ we would need $P(\text{reject } H_0 \text{ and } H_0 \text{ is true})$ — the multiplication rule tells us this would be $P(\text{reject } H_0 | H_0 \text{ is true}) \times P(H_0 \text{ is true})$ — and we don't know the second of these)
Risk of Type II error = $P(\text{fail to reject } H_0 | H_0 \text{ is not true})$ [not same as $P(\text{fail to reject } H_0 \text{ and } H_0 \text{is not true})$].

We can calculate the risk of type I error — thats the p-value —- because assuming H_0 is true gives us a value for μ, so we can convert to t or Z.

For type II error, we dont have a value for μ - saying "the mean isnt k" doesnt give a value for computations.

For each possible value a of μ we get a value for β -risk: $P(\text{fail to reject } H_0 | \mu = a)$ If we select a set of values, starting near k (the value used in H_0) we get a table showing how the risk changes as we consider different possibilities for μ .

To calculate β-risk, we need to have:

1. a null-hypothesis value k
2. the form $(>, <, \neq)$ of the alternative hypothesis
3. A specific value of α, the level of significance for the test.
4. An assumed *population* standard deviation σ so we can work with Z

Procedure:

1. Based on the test, find the rejection criterion ("Reject H_0 if sample Z ...") and convert to the "nonrejection criterion" ("Fail to reject H_0 if sample Z ...") based on the critical values (or values)
2. Convert this Z inequality to an inequality for $\bar{x}$, based on the formula $\bar{x} = \mu + Z\sigma_{\bar{x}}$ (based on the assumption that $\mu = k$ and remember that $\sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}}$)
3. Calculate the probability of this condition (stated in 2) based on the assumption that $\mu = a$ (this involves converting to Z but based on a *different value* of μ).

an example
Suppose Rachel Researcher is carrying out a test to determine whether the mean of a population is different from 45, at level $\alpha = .05$, using a sample size 30, if the population standard deviation is known to be 25.

Let us find the β-risk if the population mean really is 47 [This is the probability that Rachel will fail to find a difference, even though the real mean is 47 - which clearly *is* different from 45]

Similarly, let us find the risk if the mean really is 49.

What if it is 51?

Power
The complement of Beta risk is the probability that the hypothesis test really *will* catch the difference. It is called the

power of the test. That is:
Power of a test is a function — for each possible real value of the parameter (μ in our case)is the probability that the test will detect the difference (and allow us to reject H_0)

Sample size
In designing studies and experiments, the researchers take both type I and type II risks into account of course type II risk has to be the risk of making an error "If the *real* mean is ..."
(This should start to sound a lot like the questions that come up in acceptance sampling - balancing two kinds of risk, having to pay attention to the difference between what we want and what actually happens...)

1. A limit on the risk of type I error (α)
2. A limit on the risk of type II error (β)for a specified alternative mean (μ_α)

We need to know the null hypothesis value (μ_0) and population standard deviation (σ).

We will derive and use this formula [Its on p. 378 in your text]. Note it is the same for both "greater" and "less" tests, but we must make the usual '$\alpha/2$" adjustment for "not equal" tests.
That is: To limit the risk of type I error to no more than α and the risk of type II error for a true mean μ_a to no more than β when we test with alternative $\mu > \mu_0$ or $\mu < \mu_0$, when the standarad deviation of the variable is σ, the sample size n must satisfy $n \geq \dfrac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_a)^2}$
With the same situation but an alternative of $\mu \neq \mu_0$ (a two-tailed alternative) the sample size must be at least $\dfrac{\left(z_{\alpha/2} + z_\beta\right)^2 \sigma^2}{(\mu_0 - \mu_a)^2}$

Examples of calculations of β-risk

1. a "<" example: We are testing to see if the mean amount of soda put in a bottle by a machine is less than 1000 ml. We know the standard deviation of the amount put in is 24 ml. We want to test at the $\alpha = .05$ level, and want to know the risk of missing the difference (believing the machine is working correctly) if the machine is underfilling by 10ml. and we use a sample of 40 bottles filled by the machine
Variable is X=amount of soda put in a bottle
$H_0 : \mu = 1000$
$H_a : \mu < 1000$
We want to find β if the real mean is 990 ml.

1. We reject H_0 if sample $Z < -Z_{.05} = 1.645$, so we *do not reject* if sample $Z \geq -1.645$.
2. Since sample $Z = \frac{\bar{x}-1000}{24/\sqrt{40}}$, we fail to reject if $\bar{x} \geq 1000 - 1.645(24/\sqrt{40}) = 993.76$
3. The probability of the result stated in 2 is $P(\bar{x} \geq 993.76|\mu = 990) = P(Z \geq \frac{993.76-990}{24/\sqrt{40}}) = P(Z \geq .99) = 1 - .8389 =$ .1611.

There is a 16% chance of failing to detect the difference if the real mean is 990 ml.

If we wanted to reduce the chance of missing a 10-ml shortfall in the mean (risk of missing the difference when mean really is 990 ml) to 10%, keeping the α-risk (risk of *claiming* the mean is short when it really isn't) at 5% we 'd need a larger sample. Minimum size: $n \geq \dfrac{(1.546 + 1.282)^2 24^2}{(990 - 1000)^2} = 49.3$. We 'd need to sample at least 50 bottles.

2. Effect of a change in the significance level α. [compare to #1]
If we change the significance level of the test to $\alpha = .01$ (decrease the risk of type I error, otherwise the same test)

1. We reject H_0 if sample $Z < -Z_{.01} = -2.326$, so we *do not reject* if sample $Z \geq -2.326$.
2. Since sample $Z = \frac{\bar{x}-1000}{24/\sqrt{40}}$, we fail to reject if $\bar{x} \geq 1000 - 2.326(24/\sqrt{40} = 991.17$
3. . The probability of the result stated in 2 is $P(\bar{x} \geq 991.17|\mu = 990) = P(Z \geq \frac{991.17-990}{24/\sqrt{40}}) = P(Z \geq .31) =$ $1 - .6217 = 3783$.

There is a 38% chance of failing to detect the difference if the real mean is 990 ml.
Decreasing the α risk caused an increase in the β risk. If the same test (sample size, hypotheses) is used, reducing α will always produce an increase in β.

If we want to retain this significance level ($\alpha = .01$) and reduce our β-risk a for a mean of 990 ml to 10%, we would again need a larger sample: $n \geq \dfrac{(2.326 + 21.282)^2 24^2}{(990 - 1000)^2} = 74.95$. To achieve the stronger significance level with this β we'd need to sample at least 75 bottles.

3. A "$\neq$" example.
We wish to determine whether the mean year-end bonus at a securities firm is different from the reported industry average \$125,500 (standard deviation \$30,000). We will use a sample of 40 bonuses from this firm and test at the .05 level. Our test is : X = bonus for one person at this firm
$H_0 : \mu = \$125,500$
$H_a : \mu \neq \$125,500$
We would like to know the risk of concluding there is no difference (failing to reject H_0) if the actual mean bonus at this firm is \$118,000.

1. We reject H_0 if sample $Z > Z_{.025} = 1.960$ or if sample $Z < -Z_{.025} = -1.960$, so we do not reject (and could make a type II error) if $-1.960 \leq Z \leq 1.960$
2. Since sample $Z = \frac{\bar{x} - 125500}{30000/\sqrt{40}}$, we fail to reject if $125500 - 1.96(30000/\sqrt{40}) \leq \bar{x} \leq 125500 + 1.96(30000/\sqrt{40})$ - that is, if $116202 \leq \bar{x} \leq 134797$
3. The probability of this result is $P(116202 \leq \bar{x} \leq 134797 | \mu = 118000) = P\left(\frac{116202 - 118000}{30000/\sqrt{40}} \leq Z \leq \frac{134797 - 118000}{30000/\sqrt{40}}\right) = P(-.38 \leq Z \leq 3.54) = P(Z \leq 3.54) - P(Z < -.38) \approx 1.0000 - .3520 = .6480$

There is a 65% chance this test will fail to detect a difference if the real mean bonus is \$118,000. Notice that $P(Z < 3.54)$ is about 1.00 — we see this because 3.54 is off the right-hand side of our table. This sort of thing (one of the Z-values off the table) will often, but not always, occur in the β calculation for a two-sided test.

If we want to keep this α but reduce the β for a true mean of \$118,000 (difference \$7500) to 50%, sample size would need to be $n \geq \dfrac{(1.960 + 0)^2 30000^2}{7500^2} = 61.46$ — minimum realistic sample size is 62.[Notice the 1.960 which is $z_.025$)
To reduce β for a difference of \$7500 to 10%, need $n \geq \dfrac{1.960 + 1.282)^2 30000^3}{7500^2} = 168.17$ – we'd need at least 169 bonuses.
[Reality check: At this point we'd almost certainly be taking a large proportion of bonuses - should/could adjust our calculations using the finite population correction, so the sample size wouldn't need to be this large.]