4: MORE HYPOTHESIS TESTING

The Logic Behind Hypothesis Testing

For simplicity, consider testing $H_0 : \mu = \mu_0$ against the two-sided alternative $H_1 : \mu \neq \mu_0$. 

Even if $H_0$ is true (so that the expectation of $\bar{X}$ is $\mu_0$), $\bar{X}$ will probably not equal $\mu_0$ exactly.

Instead, we need to decide if the observed difference between $\bar{x}$ and $\mu_0$ can plausibly be accounted for by chance (i.e., by the natural variability of $\bar{X}$) or should be attributed to a systematic difference between the true and hypothesized means, $\mu$ and $\mu_0$. 
If $H_0$ is true, then $Z$ is approximately standard normal, and will very rarely lie outside the interval $(-z_{\alpha/2}, z_{\alpha/2})$.

But if $\mu \neq \mu_0$ then the distribution of $Z$ will have a nonzero mean, with the same sign as $\mu - \mu_0$, and it would not be so unusual to find $Z$ in the rejection region.
So if for our given data we find that $z$ is in the rejection region, there are only two possibilities:

- EITHER $H_0$ is true, in which case the observed value of $z$ must be just a "fluke", or rare event, due simply to the natural variability of $\bar{X}$; (This "false alarm" scenario is not impossible, although it is somewhat implausible, especially if $\alpha$ is small),

- OR ELSE $H_0$ must be false.
Here, a reasonable person would conclude that there is sufficient evidence to reject $H_0$.

The situation is analogous to having an alarm which almost never goes off falsely, but which is now ringing.

It is more plausible that the largeness of $|z|$ is caused by some systematic effect (i.e., that $\mu \neq \mu_0$), rather than by the natural variability of a standard normal. Thus, we reject $H_0$. 
Statistical Significance And The Meaning Of $\alpha$

- If $H_0$ is rejected, we say that the results are statistically significant at level $\alpha$. In this case, we have proven that $H_1$ is true, beyond a reasonable doubt (but not beyond all doubt).

Note that $\alpha$ is not the probability that $H_0$ is true, since there is nothing random about $H_0$.

Instead, $\alpha$ represents the false alarm rate (Type I error rate) of the test, i.e., the proportion of the time that a test of this kind would reject $H_0$ if $H_0$ were in fact true.
A finding of statistical significance does not provide absolute proof that $H_0$ is false.

We may be committing a Type I error (i.e., we may have a false alarm).

To make matters worse, we may never find out whether we made a mistake by rejecting $H_0$.

We do know, however, that if $H_0$ were true, then false alarms would be unlikely to occur: they would have probability $\alpha$. 
If $H_0$ is not rejected, then we say that the results are not statistically significant at level $\alpha$.

The terminology often used here is that $H_0$ is "accepted", but this should be avoided, since our inability to find sufficient evidence to reject $H_0$ does not in any way demonstrate that $H_0$ is true. (By analogy, the acquittal of a defendant on murder charges obviously does not constitute proof of innocence.)
Tests For $\mu$ When $\sigma$ Is Unknown

When $\sigma$ is unknown, we estimate it by the sample standard deviation, $S_x$.

The test statistic to use in this case is

$$t = \frac{\bar{X} - \mu_0}{S_x / \sqrt{n}}.$$

If the population is normal and $H_0$ is true, then $t$ has a Student’s $t$ distribution with $n-1$ degrees of freedom.
The criteria for a level $\alpha$ test are:

\[ H_1 \quad \text{Rejection Region} \]

\[
\begin{align*}
\mu \neq \mu_0 & \quad |t| > t_{\alpha/2} \\
\mu < \mu_0 & \quad t < -t_\alpha \\
\mu > \mu_0 & \quad t > t_\alpha
\end{align*}
\]

This test is commonly referred to as the $t$-test.

Values of $t_\alpha$ can be found in Table 2, using $\nu = n - 1$.

As $\nu$ gets larger, $t_\alpha$ becomes smaller.

For $\nu \geq 30$, $t_\alpha$ and $z_\alpha$ are virtually identical.
Before applying the $t$-test, it is wise to check a histogram of the data for approximate normality. Although it is safe to apply the $t$-test even if the data contain outliers, the actual level (false alarm rate) of the test will be somewhat smaller than $\alpha$ in this case.

A more serious problem is that the probability of a Type II error will be larger, so the test has a harder time detecting that $H_1$ is true, than in the normal case.