Hypothesis Testing

- basic ingredients of a hypothesis test are
  1. the *null hypothesis*, denoted H_o
  2. the *alternative hypothesis*, denoted H_a
  3. the *test statistic*
  4. the *the data*
  5. the *conclusion*
- the hypotheses are usually statements about the values of one or more unknown parameters, denoted θ here
- the null hypothesis is usually a more restrictive statement than the alternative hypothesis, *e.g.* $H_o: \ \theta = \theta_o, \ \ H_a : \theta \neq \theta_o$
- the burden of proof is on the alternative hypothesis
- we will continue to believe in the null hypothesis unless there is very strong evidence in the data to refute it
- the test statistic measures agreement of the data with the null hypothesis
  - a reasonable combination of the data and the hypothesized value of the parameter
  - gets bigger when the data agrees less with the null hypothesis
- when $\hat{\theta}$ is an estimator for θ with standard error $s_{\hat{\theta}}$, a common test statistic has the form
$$z = \frac{\hat{\theta} - \theta_o}{s_{\hat{\theta}}}$$
- when the data agrees perfectly with the null hypothesis, $z = 0$
- when the estimated and hypothesized values for θ become farther apart, z increases in magnitude
- there are two closely related approaches to testing
  1. one weighs the evidence against H_o
  2. the other ends in a decision to reject, or not to reject H_o.
- the first uses the significance probability or P-value
  - the probability of obtaining a value of the test statistic as or more extreme than the value actually observed, assuming that H_o is true
  - this requires knowledge of the distribution of the test statistic under the assumption that H_o is true, the *null distribution*

- for the two-sided alternative and test statistic mentioned above, the P-value is
$$P = 2Pr(|z| \geq |z_{observed}|)$$
- the factor 2 is required because *a priori* the sign of $z_{observed}$ is not known, and large (in magnitude) negative and positive values of z give evidence against H_o
- occasionally we use a one-sided alternative, $H_a : \theta > \theta_o$ or $H_a : \theta < \theta_o$
- in these cases
$$P = Pr(z \geq z_{observed})$$
and
$$P = Pr(z \leq z_{observed})$$
respectively
- the strength of the evidence against H_o is determined by the size of the P-value
  - a smaller value for P gives stronger evidence
- the logic is that if H_o is true, extreme values for the test statistic are unlikely, and therefore a possible indication that H_o is not true
- by convention we draw the following conclusions

P value	Strength of evidence against H_o
$> .10$	none
(.05, .10]	weak
(.01, .05]	strong
$< .01$	very strong

- when $P < .01$, for example, we could say that 'the results are statistically significant at the .01 level'
- the second approach to hypothesis testing requires a decision be made whether or not to reject H_o
- one way to do this is to compare the P value to a small cut-off called the significance level α and to reject H_o if $P \leq \alpha$
- another approach is to choose a *rejection region* and to reject H_o if the test statistic falls in this region
- two types of error are possible with this approach
  1. a *type I error* occurs if H_o is rejected when it is true
  2. a *type II error* occurs if H_o is not rejected when it is false

- the type I error is considered to be much more important than the type II error
- a common analogy is with a court of law
  - in murder cases the presumption of innocence (H_o) is rejected only when the jury is convinced "beyond a shadow of a doubt" by very strong evidence (an extreme value for the test statistic)
  - the type I error would be to convict and hang the accused (reject H_o) when he is innocent (H_o is true)
  - the type II error, considered less serious, would be to let a guilty man go free (don't reject H_o when it is false)
- recognizing the seriousness of the type I error, the rejection region is chosen so that the probability of rejecting H_o when it is true is a small value α
- for example, the test statistic z discussed above frequently has an approximate normal distribution. For the two-sided alternative, with $\alpha = .05$, the rejection region consists of the values $|z| \geq z_{\alpha/2} = 1.96$.
- when the data is assumed to be normally distributed and the variance is unknown and estimated by a sample variance, we use the t distribution
- finally, the data is collected and the test statistic is computed
- if the test statistic falls in the rejection region we *reject* H_o at *level* α.
- otherwise we *do not reject* H_o at level α
- remember that
  - a rejected H_o may in fact be true
  - an H_o which is not rejected is probably not true either (This is why I *never* say 'H_o is accepted').
  - a result which is statistically significant (*i.e.* we have rejected H_o) may have no practical significance. With a very large sample size almost any H_o will be rejected.