14 Power and Sample Size

Consider the simple problem from last semester where we sample from a normal population with a known σ and want to test a hypothesis of the form $H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \mu_0$. The rule for the test is to reject H_0 if $|Z| > z_{critical}$ where $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$. $z_{critical}$ is chosen to satisfy $P(|Z| > z_{critical}|H_0 \text{ is true}) = \alpha$, where α is the probability of a Type I Error, or the significance level. A Type I Error is rejecting H_0 when H_0 is actually true, i.e. claiming something important is going on when actually nothing important is going on. Usually we take $\alpha = .05$ which forces $z_{critical}$=1.96.

The significance level is the chance of rejecting H_0 when we should not do so. If H_A is true then of course we *should* reject H_0. The probability that we correctly reject H_0 when H_A is true is defined as the **Power** of the test. Power is a lot more complicated than α, though.

The extra complication comes from two sources. First, H_A is not *simple* like H_0 is, so if H_A is true there are many possible values of μ other than μ_0. If the actual μ is a long way from μ_0 then it should be fairly easy to tell that H_0 is not true and the power should be high. If the actual μ is close to μ_0, though, it should be pretty hard to tell that we are not sampling from the situation described by H_0, and the power may be low. Second, the sample size has a lot to do with the power. If n is large then we have a lot of information and it should be easy to tell if H_A is true, but if n is small it may be very difficult to tell that H_0 is not true. By contrast, the test is structured so that α does not depend upon the sample size – it is always the fixed number we choose (usually .05).

We can actually tell exactly how the Z-statistic above behaves if we specify exactly which of the values of μ is true. $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is standard normal if we used the right μ, and we are using $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ (which is standard normal if we used the right μ, i.e. if H_0 is true). If H_A is true write

$$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} = \frac{\overline{X}-\mu+(\mu-\mu_0)}{\sigma/\sqrt{n}} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} + \frac{\mu-\mu_0}{\sigma/\sqrt{n}}$$

which tells us that Z is a normal random variable with standard deviation 1 and mean $\sqrt{n}(\frac{\mu-\mu_0}{\sigma})$ if H_A is true.

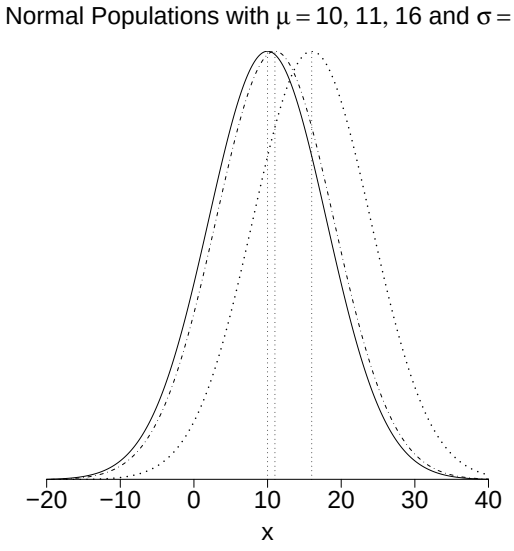

Figure 1: Three normal populations with $\sigma = 8$.

Example: Let $\sigma = 8$ and test $H_0 : \mu = 10$ vs. $H_A : \mu \neq 10$. There are infinitely many possible values of μ under H_A, but for purposes of illustration let us consider only $\mu = 11$ and $\mu = 16$. Figure 1 graphs all three populations. Clearly it is going to be fairly hard to tell whether we sampled from the population with $\mu = 11$ or the population with $\mu = 10$, since they are so little different, but it should be quite a bit easier to tell if we sampled from the population with $\mu = 16$ or the population with $\mu = 10$ (although that is not trivial either).

Consider a fixed sample size of $n = 16$. $P(|Z| > 1.96|\mu = 10) = \alpha = .05$. The power when $\mu = 11$ is $P(|Z| > 1.96|\mu = 11)$ while the power when $\mu = 16$ is $P(|Z| > 1.96|\mu = 16)$. Figure 2 (a) shows the distribution of the Z-statistic for sampling from each of these populations, (b) shows the calculation of α, (c) shows the power (.08) for $\mu = 11$ and (d) shows the power (.85) for $\mu = 16$. As expected, we have a good chance with this sample size of telling $\mu = 16$ from $\mu = 10$ but a slim chance of telling $\mu = 11$ from $\mu = 10$.

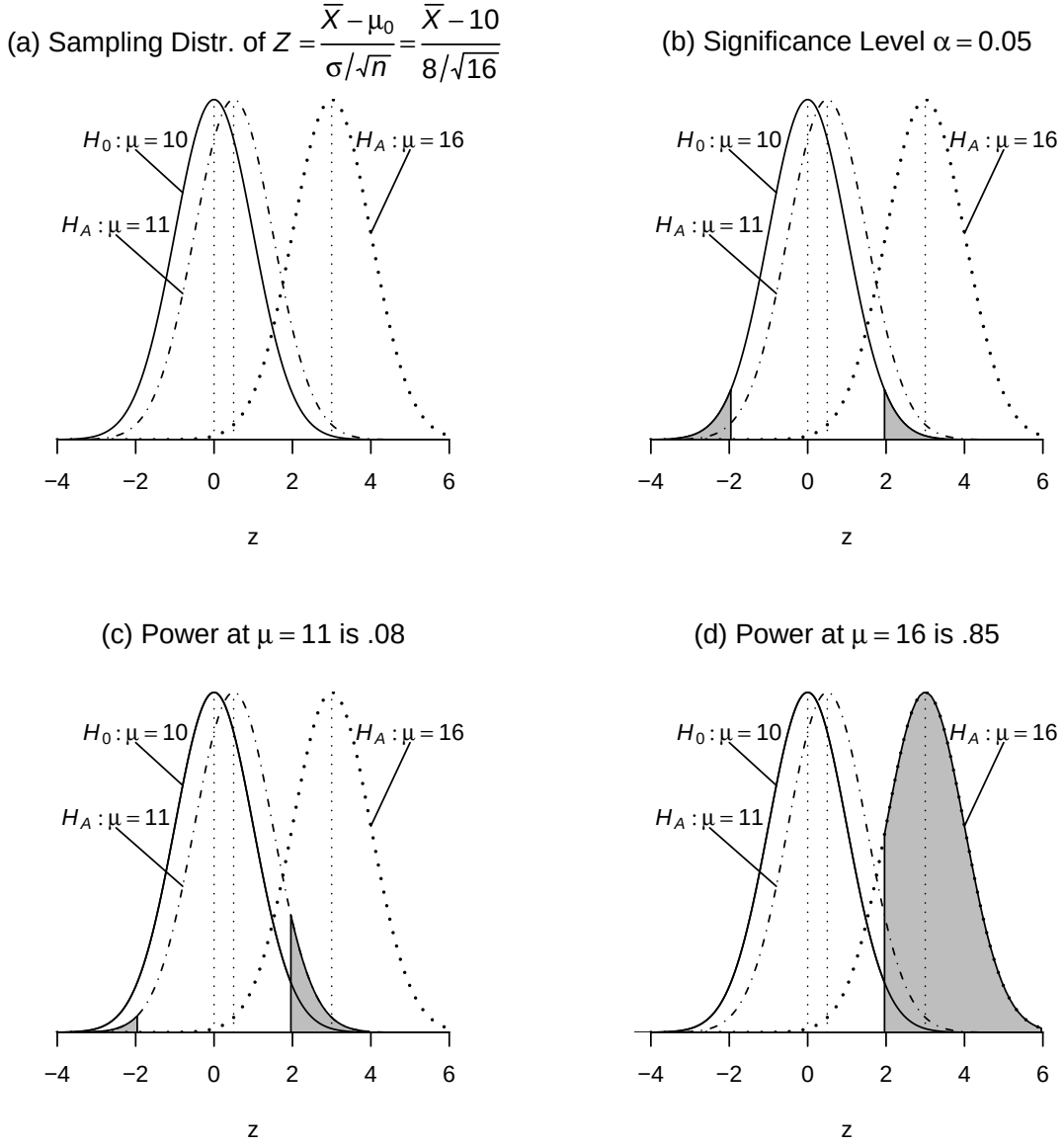

Figure 2: Power for a test of $H_0 : \mu = 10$ vs. $H_A : \mu \neq 10$ for a random sample of 16 from a normal population with $\sigma = 8$.

For a given alternative μ, the power also increases with n. Figure 3 demonstrates this behavior for the example using $\mu = 16$ and sample sizes of 5, 15, and 25.

Stata will do calculations like these for you. Follow the menu path

```
Summaries, tables, & tests
     → Classical tests of hypotheses
          → Sample size and power determination
```

and fill in the boxes. Here we are doing "One-sample comparison of mean to hypothesized value" so check that, give hypothesized value of 10, Std. deviation one of 8, and Postulated mean of 16. Next click on the Options box, ask to compute power, specify significance level of .05 and a two sided test with sample size 15. **Stata** returns the following:

```
. sampsi 10 16, alpha(.05) n1(15) sd1(8) onesample
Estimated power for one-sample comparison of mean
  to hypothesized value
Test Ho: m =     10, where m is the mean in the population
Assumptions:
         alpha =   0.0500  (two-sided)
 alternative m =       16
            sd =        8
 sample size n =       15
Estimated power:
         power =   0.8276
```

This is exactly what you find on Figure 3 for n = 15.

Much more common is the inverse of this problem, where we specify the power and ask for the sample size. Conceptually this is not different from what we have been doing, but there are some guidelines. Generally we specify some reasonable alternative, make a good guess based on published literature or preliminary data of the standard deviation, specify a two-tailed procedure at $\alpha = .05$, and target power of .8. The goal is to find n to yield that power. If we do that for the earlier problem with $\mu = 11$ and $\sigma = 8$ where $\mu_0 = 10$, we get

```
. sampsi 10 11, alpha(.05) power(.80) sd1(8) onesample
Estimated sample size for one-sample comparison of mean
  to hypothesized value
Test Ho: m =     10, where m is the mean in the population
Assumptions:
         alpha =   0.0500  (two-sided)
         power =   0.8000
 alternative m =       11
            sd =        8
Estimated required sample size:
             n =      503
```

which says we need over 500 observations to have an 80% chance of telling populations as close as the two closest in Figure 1 apart.

Stata does power analysis like the preceding on two-sample tests for means and both one- and two-sample tests for proportions. There is not much conceptual difference, but you will get a chance to experiment a bit in lab. There are specialized packages for more complex/complete calculation (I usually use PASS in NCSS because I have it), but there is free software as well. Check out UCLA's nice little calculator (`Power Calculator` at `http://calculators.stat.ucla.edu/`) for a wider variety of procedures.

The paper by Cohen on the web site is standard reading on this topic. His approach can be very useful.

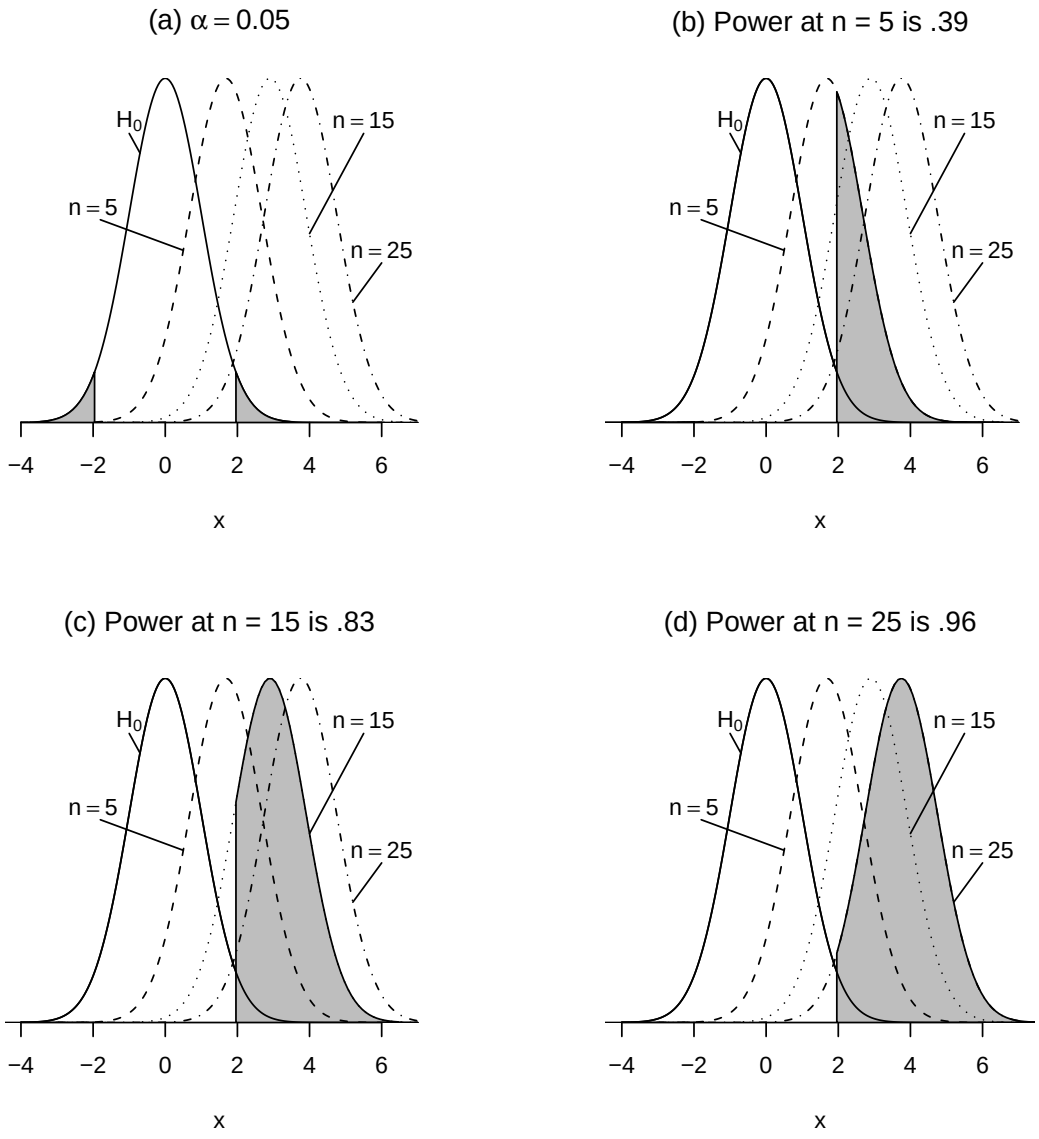

Figure 3: Power for a test of $H_0 : \mu = 10$ vs. $H_A : \mu = 16$ for random samples of 5, 15, 25 from a normal population with $\sigma = 8$.