The purpose of this exercise is to see how we can use actual data to calculate the power of alternative designs for an experiment. In the process you should learn how Stata stores the results of your regressions, and how you can use this information.

It's important to note that the data you use for power calculations don't have to come from a baseline study: on the contrary, once a baseline is in place, many of the key decisions for which power calculations are useful have already been fixed.

In this case we'll consider the design of a study that would issue certificates of customary or mailo occupancy to landholders in Uganda. Such interventions are currently being piloted by the Ministry of Lands. We'll UNHS data to design a baseline study, assuming an element of randomization has been introduced into implementation. We will look at the impact of receiving a certificate on yields, since economic theory predicts that security of property rights should improve investment.

1. Load, describe, and summarize the dataset.
   `use` "*pathname*/land.dta", `clear`
   `describe`
   `summarize`

   We'll focus on the outcome variable *yield*; you might want to get a feel for this with a histogram:

   `histogram` *yield*

2. First let's estimate the minimum detectable effect under the assumption that randomization is done at the *individual* level. This would imply that we could randomly give a certificate to one person and not to their neighbor—of course we can't. The MDE is then given by

$$\beta_{MDE} = (t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}. \tag{1}$$

   Assume that

   - We sample half 'treated' people and half 'control' people, so $P = 0.5$. There will be 1200 observations (this gives N).
   - We will eventually test the hypothesis at a 5% confidence level, and we want power of 80%. This means (for large N) that $t_{1-\kappa} + t_{\alpha/2}$ is approximately 1.96. Stata will calculate this for you if you enter
     `display invttail`(1200, .05/2) + `invttail`(1200, 1-.8)

   Stata will store these values if you define what is called a **local** variable. You can define a local variable by typing, for example,

`local` *M* 1.96
Then this value can be called by using the name of the local variable, enclosed in single quotes, for example,
`display` *'M'*

(a) Store the values of N, M, and P as local variables under these names, where $M = t_{1-\kappa} + t_{\alpha/2} = 1.96$.

(b) Now the only missing piece from our MDE calculation is the variance of the *yield* outcome variable. You can get this with the command `summarize`:
`summarize` *yield*
Now you can define a local variable (call this *sigmasq*) to store this value by hand. Alternatively, note that typing
`return list`
gives you a set of local variables returned by the summarize command. To automatically define *sigmasq* as the variance of *yield*, type
`local` *sigmasq* 'r(Var)'

(c) You are ready to have Stata calculate the MDE for you. Type
`display` 'M' * sqrt(1 / ('P'/(1-'P'))) * sqrt('sigmasq' / 'N')
Check this against the formula above. What is the value of the MDE that you get? Make note of this here.

3. Suppose we know from a pilot that 80% of people who are offered such a certificate (in fact, I don't have this information) take up the offer. Thus the fraction of *compliers*, c, is 0.8; let's also assume that 10% of people who are *not* offered a certificate by the government obtains one (so the fraction of *defiers*, d, is 0.1). The MDE is given by

$$MDE = \frac{1}{c-s}(t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}} \tag{2}$$

Can you define local variables to store the values for c and d, and modify the formula used in part 2c to get this result? What do you get? How does non-compliance affect the study's power?

4. Now let's explicitly allow for the fact that the randomization will occur at the parish level. The model is now

$$Y_{ij} = \beta T_j + u_j + e_{ij} \tag{3}$$

where u_j are unobserved parish characteristics, and e_{ij} are unobserved characteristics of individual i in parish j. The MDE is given by

$$\beta_{MDE} = \frac{t_{1-\kappa} + t_{\alpha/2}}{\sqrt{P(1-P)J}}\sqrt{\rho + \frac{1-\rho}{n}}\sqrt{\sigma_u^2 + \sigma_e^2} \tag{4}$$

where σ_u^2 is the variance of the u_j, σ_e^2 is the variance of the e_{ij}, and $\rho = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$.

(a) The first task is to recover the values for ρ, σ_u^2 and σ_e^2. We saw these on Day 1 when we used `xtreg` to estimate a fixed effects model. To see these, estimate a fixed effects regression where the 'groups' are defined by parishes, *with no explanatory variables*:
`xtreg` *dispute* , `fe i`(parish)
This gives us exactly what we want in the output automatically. Type `ereturn list` to see the local variables automatically stored by this command. Store these values as local variables: `local` sigma_u 'e(sigma_u)'
`local` sigma_e 'e(sigma_e)'
`local` rho 'e(rho)'

(b) Now we need to decide how our sample will be allocated across groups. To begin, assume $P = .5$ as before. Let's let $J = 30$ (30 parishes included), and $n = 40$ (40 households per parish). Note the total households has not changed—there are still 1,200. Define local variables for n, P, and J.

(c) Now tell Stata to display the MDE:
`display` ('M' / sqrt('P' * (1-'P')*'J')) *
sqrt('rho' + (1-'rho')/'n')
* sqrt('sigma_u'$\hat{2}$ + 'sigma_e'$\hat{2}$)
How much does the MDE change relative to previous estimates? Why?

(d) What would be the effect of concentrating households in fewer parishes? Recalculate the MDE for $J = 10$, $n = 120$.

(e) How sensitive are the results to changes in the proportion of treated households? Try values of $P = .4, .25, .10$.

5. Does including covariates improve (i.e., lower) the MDE?

Our estimate of the treatment effect may be more precise if we include covariates in the estimation. Power calculations are a straightforward modification: just redo (4.a) and (4.b), but with covariates in the `xtreg` estimation at the outset:
`xtreg` *dispute mailo customary female femaledivorced femalewidowed size_est*, fe i(parish)
Of course, these may be of independent interest.