1. The file `wers.csv` contains data from a 1998 workplace employee-relations survey, altered to protect confidentiality.

   The survey had 71 strata. Within each stratum, a simple random sample of firms was taken. As is common in business surveys, the sampling fraction was very different in different strata.

   The variables in the file are

   | | |
   |---|---|
   | "serno" | an id number |
   | "est_wt" | sampling weights re-scaled to have unit mean |
   | "female" | percentage of female employees |
   | "nempsize" | total number of employees |
   | "disabgrp" | % disabled employees |
   | "ethnic" | % employees from ethnic minorities |
   | "eo" | Is there a formal equal-opportunity policy? |
   | "grosswt" | Sampling weight |
   | "strata" | Stratum id |
   | "sampfrac" | Sampling fraction for that stratum |

   a. Define a survey design object in R to describe these data
   b. Estimate the mean and median percentage of female employees. Compare these estimates to the unweighted mean and median.
   c. Estimate the difference in mean % female employees between workplaces with and without a formal equal-opportunity policy, and test whether the data are consistent with no difference existing in the population (hint: `svyttest`)
   d. Is there evidence of an association between workplace size and existence of an equal-opportunity policy?
   e. Estimate the average proportion of female employees and ethnic-minority employees in each workplace size group.

2) Data from NHANES on blood pressure and diet are in `combined-data.csv`. The primary sampling unit variable is `SDMVPSU`, the stratum variable is `SDMVSTRA`, and the weight variable for this sample is `fouryearwt`.
   a. Create a survey design object with this data. You will need to remove the records with missing weights, and you will need the `nest=TRUE` option. Look at the values for `SDMVPSU` and the help page for svydesign to see why `nest=TRUE` is needed. The National Center for Health Statistics always does it this way.
   b. Draw graphs of systolic (`BPXSAR`) and diastolic (`BPXDAR`) blood pressure by age and sex.
   c. Build a model for systolic blood pressure using race/ethnicity (`RIDRETH1`), gender (`RIAGENDR`), age (`RIDAGEYR`), BMI (`BMXBMI`) sodium intake (`DR1TSODI`) and potassium intake (`DR1TPOTA`). Feel free to give the variables more sensible names. The function `svyglm()` fits linear models; there is an example in yesterday's slides.

d. Define a new variable for hypertension by (eg)
`nhanedes<-update(nhanesdes, htn=(BPXSAR>140) | (BPXDAR>90))`
and do a similar analysis using logistic regression, which you get by adding `family=quasibinomial()` to the call to `svyglm()`

3) With the data from question 1 Draw appropriate graphs to illustrate
   a. the relationship between presence of a formal equal-opportunity policy and % female employees (and compare to an unweighted graph)
   b. The relationship between total number of employees and %female
   c. The relationship between presence of a formal equal-opportunity policy and % female employees, conditioning on the total number of employees

4) The file `shs.csv` has data from the 2001 Scottish Household Survey (slightly fictionalized for confidentiality). The survey using stratified sampling of households in high-density areas and stratified cluster sampling in low-density areas. The primary sampling unit is identified by `PSU`, the stratum by `STRATUM`, and the post-stratified weight by `GROSSWT`.
   a. Examine how internet use (`INTUSE`) varies by `COUNCIL`
   b. Examine how internet use varies by age and gender and income, with regression and/or graphics

5) The file `scothealth.rda` contains data from the 1998 Scottish Health Survey, on smoking prevalence. This was actually a systematic PPS sample of postcode sectors ordered on deprivation index followed by random sampling of addresses, but we are analyzing it as if it was a stratified sample. The 'stratum' variable is `REGSTRAT`, the postcode sector sampling unit variable is `PSU` and the weight variable is `GROSSWT`.
   a. Compare the weighted and unweighted prevalences of smoking by region and by age
   b. Built a logistic regression model for smoking

6) The file `certsurvey.rda` contains data from a 1994 survey of the American Statistical Association, asking whether the members supported a professional certification proposal. The five variables `CERTIFY` (should ASA have professional certification?), `APPROVE` (do you approve of the specific proposal?), `SPECCERT` (should there be specific certification for sub-disciplines?), `WOULDYOU` (would you apply?), and `RECERT` (should regular re-certification be required?) are on a 1=yes, 5=no scale, with 0 for non-response.

    a. There were 18609 members of the ASA, and 5001 responses were obtained. From membership data, 55% have PhDs and 38% have Masters degrees; 29% work in industry, 34% in academia, and 11% in government. Use `rake()` or `calibrate()` to reweight the data to these population figures, and see how much the answers to the five questions change.

    b. Is there a relationship between `CERTIFY` and `WORKENV`(work environment) or `COLLEGE`(highest degree)? How about `WOULDYOU`? If you know about loglinear models you might like to use `svyloglin()` here, but two-way tables are enough. You might also look at what `plot()` does with tables

7) The file frs.csv has (confidentialized) data from the Scotland Family Resources Survey. The primary sampling units, indicated by the variable `PSU`, are postcode sectors. Weights are in the `GROSS2` variable. These published weights have been raked on `CTBAND` (council tax band) and `TENURE` (housing tenure).

    a. Because the published weights are already raked, you can compute the population totals for the raking variables from the sample. Do this.

    b. Create a raked survey design object using these population totals

    c. Compare the raked and unraked inference for mean income (`HHINC`) and its standard error

    d. Look at the relationship between mean income and number of dependent children, `DEPCHLDH`