In [1]:
```
# DSC530-T302
# Stephen Smitshoek
# Week03
# Exercise 2-4
```

In [2]:
```
import sys
import numpy as np
import thinkstats2
import math
```

In [3]:
```
def ReadFemPreg(dct_file='2002FemPreg.dct',
                dat_file='2002FemPreg.dat.gz'):
    """Reads the NSFG pregnancy data.

    dct_file: string file name
    dat_file: string file name

    returns: DataFrame
    """
    dct = thinkstats2.ReadStataDct(dct_file)
    df = dct.ReadFixedWidth(dat_file, compression='gzip')
    CleanFemPreg(df)
    return df
```

In [4]:
```
def CleanFemPreg(df):
    """Recodes variables from the pregnancy frame.

    df: DataFrame
    """
    # mother's age is encoded in centiyears; convert to years
    df.agepreg /= 100.0

    # birthwgt_lb contains at least one bogus value (51 lbs)
    # replace with NaN
    df.loc[df.birthwgt_lb > 20, 'birthwgt_lb'] = np.nan

    # replace 'not ascertained', 'refused', 'don't know' with NaN
    na_vals = [97, 98, 99]
    df.birthwgt_lb.replace(na_vals, np.nan, inplace=True)
    df.birthwgt_oz.replace(na_vals, np.nan, inplace=True)
    df.hpagelb.replace(na_vals, np.nan, inplace=True)

    df.babysex.replace([7, 9], np.nan, inplace=True)
    df.nbrnaliv.replace([9], np.nan, inplace=True)

    # birthweight is stored in two columns, lbs and oz.
    # convert to a single column in lb
    # NOTE: creating a new column requires dictionary syntax,
    # not attribute assignment (like df.totalwgt_lb)
    df['totalwgt_lb'] = df.birthwgt_lb + df.birthwgt_oz / 16.0

    # due to a bug in ReadStataDct, the last variable gets clipped;
    # so for now set it to NaN
    df.cmintvw = np.nan
```

In [5]:
```
def data_split(preg_df):
    # Find all the live births and split them into first babies and other babies
```

```
    live = preg_df[preg_df.outcome==1]
    first = live[live.birthord == 1]
    other = live[live.birthord != 1]

    return first, other
```

In [6]:
```
def cohen_effect_size(group1, group2):
    diff = group1.mean() - group2.mean()

    var1 = group1.var()
    var2 = group2.var()

    n1, n2 = len(group1), len(group2)

    pooled_var = (n1 * var1 + n2 * var2) / (n1 + n2)

    d = diff / math.sqrt(pooled_var)
    return d
```

In [7]:
```
def main():
    preg_df = ReadFemPreg()
    CleanFemPreg(preg_df)
    first, other = data_split(preg_df)

    print('Summary of First Baby Weight vs Other Baby Weight')
    print('First Babies Mean: {} lbs'.format(round(first.totalwgt_lb.mean(), 1)))
    print('Other Babies Mean: {} lbs'.format(round(other.totalwgt_lb.mean(), 1)))
    print('Cohen Effect Size: {}'.format(round(cohen_effect_size(first.totalwgt_lb, ot
```

In [8]:
```
if __name__ == '__main__':
    main()
```

```
Summary of First Baby Weight vs Other Baby Weight
First Babies Mean: 7.2 lbs
Other Babies Mean: 7.3 lbs
Cohen Effect Size: -0.09
```

In [9]:
```
print('\nThe Cohen effect size for first baby weight vs other babies weight is 0.09.',
      '\nThe Cohen effect size for first pregnacy length vs other babies pregnacy leng
```

```
The Cohen effect size for first baby weight vs other babies weight is 0.09.
The Cohen effect size for first pregnacy length vs other babies pregnacy length  is
0.03.
```