Assignment 3

Probability, Statistics, and Discrete Mathematics
23.4.2018
Salla Vesterinen
Helsinki Metropolia University of Applied Sciences (https://www.metropolia.fi/en)

In [ ]:
```
# Import libraries needed
%pylab inline
import pandas as pd

# Read the example data into Python
file = "https://www.openintro.org/stat/data/bdims.csv"
sep = ","
data = pd.read_csv(file, sep)
data.head()
```

Problem 1

Using the body girth and skeletal diameter measurements data (see the code given above) compare the distributions of the chest diameter (`che.di`) for men and women.

- Create a graph where you have histograms of the chest diameter for both groups. Pay special attention for the bins settings.

Calculate the descriptive statistics for both groups. Compare the statistical values between the groups.

- What is the range of the values for each group?
- How much do the mean values differ between men and women?
- How much do the standard deviation values differ between men and women?

Histogram of chest diameter for males and females: (two ways)

In [ ]:
```
x=data["che.di"][data["sex"] == 1]
y=data["che.di"][data["sex"] == 0]

plt.hist([x,y], bins=range(20,40), label=['m','f'])
xlabel('Diameter')
ylabel('Density')
plt.legend(loc='upper left')
plt.show()
```

In [ ]:
```
data["che.di"][data["sex"] == 1].hist(bins = arange(20,40+0.5,0.5))
data["che.di"][data["sex"] == 0].hist(bins = arange(20,40+0.5,0.5), alpha=0.5)
xlabel('Diameter')
ylabel('Density')
```

Chest diameter males:

In [ ]:
```
data["che.di"][data["sex"] == 1].describe()
```

Chest diameter females:

In [ ]:
```
data["che.di"][data["sex"] == 0].describe()
```

Stats Male vs. Female

What is the range of the values for each group?
Male: 24.7 to 35.6 Female: 22.2 to 33.2

How much do the mean values differ between men and women?
29.9 and 26.1, difference of 3.8

How much do the standard deviation values differ between men and women?
2.08 and 1.81, difference of 0.27

Problem 2

Study the ankle girth data (ank.gi).

- Create a histogram of the values. Pay special attention to the bins settings.
- Find out the mean and standard deviation values for the data.
- Find out the range of values in which there are at least 95% of all values.
- Verify by calculating that there are 5% of values outside of this interval.

In [ ]:
```
# Your code
data["ank.gi"].hist(bins = arange(15,30+0.5,0.5), edgecolor='black')
xlabel('Diameter')
ylabel('Density')
```

a) Find out the range of values in which there are at least 95% of all values.
-Two standard deviations from the mean
b) Verify by calculating that there are 5% of values outside of this interval.
-Count number of items outside of range and compare to 5% (507*0.05=25.35)

In [ ]:
```
x = data["ank.gi"]
mn = x.mean()
print("Mean:",mn)
sd = x.std()
print("Standard deviation:",sd)
se = sd/sqrt(x.count())
print("Standard error:",se)
ci = 1.96*se
print("Confidence interval:",ci)

l=mn-2*sd
print("Low:",l)
h=mn+2*sd
print("High:",h)
```

In [ ]:
```
a = data["ank.gi"]
print("Total:",a.count())
c=507*0.05
print("5%:",c)
x=(a<18.432).sum()+(a>25.882).sum()
print("Count:",x)
```

Problem 3

Study the elbow diameter data (elb.di).

- Create a **density** histogram of the values both for men and women. Pay special attention to the bins settings.
- Calculate the mean and standard deviation values for both groups.
- Using the ´scipy.stats´ statistics normal distribution overlay the theoretical normal distribution function over the density histograms.
- How much do the density histogram and the theoretical normal distribution differ? Explain why.

Density histogram of elbow diameter in men:

In [ ]:
```
# Elbow diameter data for men
x = data["elb.di"][data["sex"] == 1]

# Calculate mean and standard deviation
m = x.mean()
sd = x.std()
print("Mean: ", around(m, 1))
print("Standard deviation:", around(sd, 1))
rv = norm(loc = m, scale = sd)
xrnd = rv.rvs(size = 260)

weights = np.ones_like(x)/float(len(x))
hist(x, bins = arange(8, 18, 0.5), alpha = 0.5, density = True, label = "data", weights=weights)

x = arange(8, 18, 0.1)
plot(x, rv.pdf(x), lw = 2, label = "theoretical")

xlabel('Diameter')
ylabel('Probability')
legend()
grid()
data["elb.di"][data["sex"]==1].skew()
```

Density histogram of elbow diameter in women:

In [ ]:
```
# Elbow diameter data for men
x = data["elb.di"][data["sex"] == 0]

# Calculate mean and standard deviation
m = x.mean()
sd = x.std()
print("Mean: ", around(m, 1))
print("Standard deviation:", around(sd, 1))
rv = norm(loc = m, scale = sd)
xrnd = rv.rvs(size = 260)

weights = np.ones_like(x)/float(len(x))
hist(x, bins = arange(8, 18, 0.5), alpha = 0.5, density = True, label = "data", weights=weights)

x = arange(8, 18, 0.1)
plot(x, rv.pdf(x), lw = 2, label = "theoretical")

xlabel('Diameter')
ylabel('Probability')
legend()
grid()
data["elb.di"][data["sex"]==0].skew()
```

How much do the density histogram and the theoretical normal distribution differ? Explain why.

The vague shape of the data is in a bell shape, but not the perfectly formed one like te theoretical value curve is.
This is because
a) nothing in the real world is so accurate/they differ greatly due to so many different variables and
b) the data set might be small in comparison, or it might have strange deviations from the norm which skew the calculations and
c) my graph for the data has bins and isn't portioned and curved like the theoretical layer on top

In [ ]: