# Mock midterm

(Week 03 lecture notes)

Wei Q. Deng

Department of Statistical Sciences

July 18th 2018

# Mock Midterm

- Find teammates if you would like to work in a group and discuss with others
- Format is similar to the midterm
- Pay attention to each step you are required show to receive full credits

# Some mathematics

- finding the maximum of a continuous function
- if and only if statements

# Topics covered so far

- Lecture 1: Graphical display and summary statistics
- Lecture 2: Sampling distribution and (point estimator of) parameter
- Lecture 3: Point estimates and interval estimates
- Lecture 4: Confidence Intervals and statistical testing

# From lecture 1:

You should know

- the difference between a sample and a population
- concepts such as parameter, statistic, estimator
- the types of data
- what appropriately numerical and graphical summary to use for each type of data
- when to use a boxplot and a histogram
- shapes of data (centre, spread, skewness, shape of tail and outliers)
- quantile range equivalence to a multiple of standard deviation for a normal random variable
- how to detect non-normality in data (in terms of shape) through boxplot, histogram and quantile-quantile plots

# Lecture 1 related questions:

The following statements are true or false?

1. A finite collection of objects cannot be a statistical population.
2. A statistical population is specific to the underlying question of interest.
3. A parameter is a constant while a statistic is a random variable.
4. An estimator is an algorithm for calculating an estimate of a parameter based on observed data.
5. There are three types of discrete/categorical data, nominal, ordinal and integer.

# Lecture 1 related questions:

Answer the following:

1. If the data based on the boxplot seem right skewed, what is an appropriate measure of central tendency (or centre)?
2. For a fat-tailed distribution, such as t-distribution, what would the quantile-quantile plot against a normal quantile look like? Remember the quantile-quantile plots the sample quantile (of the data distribution) against the theoretical quantile of a normal distribution. For example, the sample quantiles of the data can be obtained by ordering the data by size ($100/n, 200/n, \ldots, 100 - 100/n\%$ quantile), the matching theoretical quantiles of a standard normal are $\Phi^{-1}(100/n), \Phi^{-1}(200/n), \ldots, \Phi^{-1}(100 - 100/n)$.
3. If a data distribution is bimodal but symmetric, how does the histogram look like and how does the boxplot look like?
4. If the data distribution is skewed, what is the relationship between the mean and median (discuss the two skewed cases)?

# Lecture 1 related questions (cont'd):

Answer the following:

5. Suppose we add one outlier to the data, how would the mean and median change? Now suppose an additional outlier has been added to the data but on the opposite tail, how would the mean and median change as compared to the original data?
6. Let $X$ be a normal random variable with mean 0 and variance $\sigma^2$, the relationship between its quantile and multiples of standard deviation can be summarized by the definition of cumulative distribution function $\Phi(x)$

$$\mathrm{P}(\frac{X}{\sigma} < \frac{x}{\sigma}) = \Phi(\frac{x}{\sigma})$$

Let $Y = 2X$, in other words, $Y$ is a normal random variable with mean 0 and variance $4\sigma^2$. How does the IQR of $X$ compare to the IQR of $Y$, what about their inter-95% quantile range?

# From lecture 2:

You should know

- definition of a random sample and its properties
- how to check if a sample is roughly random given that it was sampled with replacement and sampled without replacement
- parameters of normal, t- and binomial/bernoulli distributions
- definition of a statistic
- (large sample) sampling distributions of common statistic such as mean, variance, and proportion
- when the sampling distribution is exact and when to invoke the central limit theorem
- what the central limit theorem is and how to use it
- for a normal random sample,
  - $\bar{X}$ and $S^2$ are independent,
  - distribution of $\bar{X}$ and $(n-1)S^2/\sigma^2$
  - distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$
- how to find estimators using MoM and MLE (at least for normal)
- difference between an estimator and an estimate
- properties of estimators, unbiasedness, sufficiency and consistency
- how to show an estimator is unbiased ($\bar{X}$ and $S^2$ are unbiased)

# Lecture 2 related questions:

The following statements are true or false?

1. A collection of random variables must be independent and have the same distribution if they were sampled with replacement from the same population.
2. The student's t random variable has only one parameter.
3. If the marginal distribution of a variable from a random sample is given, then the joint distribution is the product of marginal probability functions over all random variables in the sample.
4. It is impossible to obtain a near random sample from a finite population.
5. The mean estimator is a function of the sample, and does not depend on the population mean.
6. For unbiased estimators, the smaller the variance the better.
7. For a given set of data, at a fixed parameter, the likelihood is equal to the joint probability.

# Lecture 2 related questions (cont'd):

Answer the following:

1. State the central limit theorem (the classic version was introduced in this class - with finite variance).
2. What is the sampling distribution of $\bar{X}$ calculated from a random sample from $\mathcal{N}(\mu, \sigma^2)$?
3. What is the sampling distribution of $(n-1)S^2/\sigma^2$ given a random sample from $\mathcal{N}(\mu, \sigma^2)$ given $\sigma^2$ is known?

$$S^2 = \frac{1}{n-1}\Big[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\Big] = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

4. What is the sampling distribution of $\bar{X}$ calculated from a random sample from $\mathcal{N}(\mu, \sigma^2)$ given that $\sigma^2$ is unknown?
5. Suppose the mean squared error of an estimator approaches 0 as sample size $n \to \infty$, is this estimator consistent if it is also unbiased?
6. Express the MSE of an estimator to show that it captures both precision and accuracy of an estimator.

# Lecture 2 related questions (cont'd):

## Sample mean of random variables

Let $X_1, \ldots, X_n$ be a random sample from a population with mean $\mu$ and variance $\sigma^2 < \infty$. Define $T(X_1, \ldots, X_n) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $T(X_1, \ldots, X_n) = S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$. Show

1. $\mathrm{E}(\bar{X}) = \mu$
2. $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$
3. $\mathrm{E}(S^2) = \sigma^2$

# Proof of a)

$$\mathsf{E}(\bar{X}) = \mathsf{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) \tag{1}$$

$$= \frac{1}{n}\mathsf{E}(\sum_{i=1}^{n} X_i) \tag{2}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}(X_i) \tag{3}$$

$$= \frac{1}{n} n\mathsf{E}(X_i) \tag{4}$$

$$= \mu \quad \square \tag{5}$$

# Proof of b)

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \tag{6}$$

$$= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \tag{7}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \tag{8}$$

$$= \frac{1}{n}\mathrm{Var}(X_i) \tag{9}$$

$$= \frac{\sigma^2}{n} \quad \square \tag{10}$$

# Before c), remember the following from STA247

Here we used the following relationship between the mean and variance:

$$\mathrm{Var}(X_i) = \mathrm{E}[(X_i - \mu)^2] = \mathrm{E}[X_i^2 - 2\mu X_i + \mu^2] = \mathrm{E}(X_i^2) - \mu^2$$

similarly

$$\mathrm{Var}(\bar{X}) = \mathrm{E}(\bar{X}^2) - E(\bar{X})^2 = \mathrm{E}(\bar{X}^2) - \mu^2$$

# Proof of c)

$$\mathrm{E}(S^2) = \mathrm{E}\Big(\frac{1}{n-1}\Big[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\Big]\Big) \tag{11}$$

$$= \frac{1}{n-1}\mathrm{E}\Big[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\Big] \tag{12}$$

$$= \frac{1}{n-1}\Big[n\mathrm{E}(X_i^2) - n\mathrm{E}(\bar{X}^2)\Big] \tag{13}$$

$$= \frac{n}{n-1}\Big[\mathrm{E}(X_i^2) - \mathrm{E}(\bar{X}^2)\Big] \tag{14}$$

$$= \frac{n}{n-1}\Big[\mathrm{Var}(X_i) + \mathrm{E}(X_i)^2 - \mathrm{Var}(\bar{X}) - E(\bar{X})^2\Big] \tag{15}$$

$$= \frac{n}{n-1}\Big[\sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2\Big] \tag{16}$$

$$= \frac{n}{n-1}\Big[\sigma^2 - \frac{\sigma^2}{n}\Big] \tag{17}$$

$$= \sigma^2 \quad \square \tag{18}$$

# Lecture 2 related questions (cont'd):

## Calculate the mean squared error for estimators of normal parameters

For iid $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, show the following given that for $Y \sim \chi^2(n-1)$, $\mathrm{E}(Y) = n-1$, $\mathrm{Var}(Y) = 2(n-1)$.

$$\mathrm{E}(\bar{X} - \mu)^2 = \frac{\sigma^2}{n}$$

and

$$\mathrm{E}(S^2 - \sigma^2)^2 = \frac{2\sigma^4}{n-1}$$

# Proof of this uses the fact that $\bar{X}$ and $S^2$ have sampling distributions

- (Slide page 31) $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
- and for $Y \sim \chi^2(n-1)$, $\mathrm{E}(Y) = n-1$, $\mathrm{Var}(Y) = 2(n-1)$

# Lecture 2 related questions (cont'd):

## Method of moments example

Let $X_1, \ldots, X_n$ be a random sample from the gamma distribution with parameter $\alpha$ and $\theta$, the p.d.f of a gamma is

$$f_X(x) = \frac{1}{\Gamma\alpha\theta^{\alpha}} x^{\alpha-1} e^{-x/\theta}, \text{ for all } x > 0$$

Given that the first two central moments of gamma are $\mathsf{E}(X) = \alpha\theta$ and $\mathsf{Var}(X) = \alpha\theta^2$. Find estimators for $\theta$ and $\alpha$ using method of moments.

# Lecture 2 related questions (cont'd):

## Method of moments example - hints

$$\mathsf{E}(X) = \alpha\theta = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

$$\mathsf{Var}(X) = \alpha\theta^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Hint: $+$ solve for $\alpha$ from the first equation and then substitute back in the second.

- $\hat{\theta} = \frac{1}{n\bar{X}}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- $\hat{\alpha} = \frac{\bar{X}}{\theta} = \frac{n\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

# Lecture 2 related questions (cont'd):

## Maximum likelihood estimator

Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. The p.d.f of a normal random variable $X$ with mean $\mu$ and variance $\sigma^2$: $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

- Find the MLE of $\mu$ and $\sigma^2$
- Confirm that the estimator $\hat{\mu}$ and $\hat{\sigma}^2$ are the maximum by checking the second partial derivatives are less than 0.

# Lecture 2 related questions (cont'd):

## Maximum likelihood estimator

Recall: the maximum likelihood estimator (MLE) $\hat{\theta}$ is defined to be:

$$\hat{\theta}(\mathbf{X}) = \text{argmax}_\theta L(\theta|\mathbf{X})$$

The likelihood is:

$$L(\mu, \sigma^2|\mathbf{x}) = \Pi_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{19}$$

$$= \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}} \tag{20}$$

The log-likelihood is:

$$\log[L(\mu, \sigma^2|\mathbf{x})] = n \log[\frac{1}{(\sqrt{2\pi}\sigma)}] - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}$$

# An example of a normal random sample (cont'd)

$$\frac{\partial}{\partial \mu} \log[L(\mu, \sigma^2|\mathbf{x})] = 2 \sum_{i=1}^{n} \frac{(x_i - \mu)}{2\sigma^2} = 0$$

implies that

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

# An example of a normal random sample (cont'd)

On the other hand,

$$\frac{\partial}{\partial \sigma^2} L(\mu, \sigma^2 | \mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^4} = 0$$

and plugging in $\hat{\mu}$ we have

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n} = \frac{n-1}{n} S^2$$

Check for yourself that the two second partial derivatives are negative.

# Definition of unbiasedness

An estimator $\hat{\theta}$ is unbiased for the parameter $\theta$ if

$$\mathsf{E}_{x|\theta}(\hat{\theta}) = \theta$$

otherwise the difference between the two

$$\mathsf{E}_{x|\theta}(\hat{\theta}) - \theta$$

is called the bias of $\hat{\theta}$ relative to $\theta$.

# From lecture 3:

You should know

- how to calculate the MSE by calculating the bias and variance of an estimator
- the two computational approaches to estimate the variance of an estimator
- understand why we want to look at interval estimator on top of point estimators
- how to obtain an interval estimator for $\bar{X}$
  - when $\sigma^2$ is known
  - when $\sigma^2$ is unknown
- how to obtain an interval estimator for the population proportion $p$
- whether the CI becomes wide/narrow as you increase/decrease sample size $n$, confidence level $\alpha$
- how to read the Z-table and t-table for quantile values

# Lecture 3 related questions (cont'd):

## Compute the jackknife estimate of the standard error for $\hat{\theta} = \bar{x}$

Hints:

$$\hat{\theta}_{(-i)} = (n\bar{x} - x_i)/(n-1)$$

$$\hat{\theta}_{(.)} = \bar{x}$$

simplify everything you get

$$\Big[\frac{1}{(n-1)n}\sum_{i=1}^{n}\Big(x_i - \bar{x}\Big)^2\Big]^{1/2}$$

# Lecture 3 related questions (cont'd):

## Derive $1-\alpha$ Confidence interval for $\bar{X}$ **when** $\sigma^2$ **is given**

Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Find $l$ and $u$ such that $\Pr_{\bar{X}}(l < \mu < u) = 0.95$.

# Lecture 3 related questions (cont'd):

## Derive $1-\alpha$ Confidence interval for $\bar{X}$ **when** $\sigma^2$ **is not given**

Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Find $l$ and $u$ such that $\Pr_{\bar{X}}(l < \mu < u) = 0.95$.

# Lecture 3 related questions (cont'd):

## Derive $1-\alpha$ Confidence interval for $\bar{X}$ **when** $\sigma^2$ **is given**

Let $X_1, \ldots, X_n$ be **a random sample** with mean $\mu$, a known $\sigma^2$ and $n > 30$. Find $l$ and $u$ such that $\Pr_{\bar{X}}(l < \mu < u) = 0.95$.

- You may skip this one if you have the answer right away.

# Lecture 3 related questions (cont'd):

Consider the following sample of fat content (in percentage) of randomly selected hot dogs:

`25.2 21.3 22.8 17.0 29.8 21.0 25.5 16.0 20.9 19.5`

- $\bar{x} = 21.90$, $s = 4.134$
- $t_{0.025,9} = 2.262$, $t_{0.025,10} = 2.228$, $t_{0.05,9} = 1.833$, $t_{0.05,10} = 1.812$
- Assuming that these were selected from a normal population distribution, a 95% CI for (interval estimate of) the population mean fat content is?

# Lecture 3 related questions (cont'd):

From the t-table for a random variable with degrees of freedom of 20, the areas to the right of the values 0.687, 0.860, and 1.064 are .25, .20, and .15, respectively.

- Which of the three intervals would you recommend be used, and why?

1. $(\bar{x} - 0.687s/\sqrt{21}, \bar{x} + 1.725s/\sqrt{21})$
2. $(\bar{x} - 0.860s/\sqrt{21}, \bar{x} + 1.325s/\sqrt{21})$
3. $(\bar{x} - 1.064s/\sqrt{21}, \bar{x} + 1.064s/\sqrt{21})$

- What is the confidence level of the chosen interval?
- What can be done to decrease the width of the CI?

# From lecture 4:

You should know

- Structure of a hypothesis test
- how to select a test statistic
- how to calculate the rejection region and p-value given the null and alternative
- definition of a p-value
- statistical tests of $\mu$ (normal, t, and binomial)
- Statistical tests to compare two samples
  - test for equality of means (three versions of two-sample t-test)
  - assumptions for different t-tests
  - test for equality of variance (F-test)

# Lecture 4 related questions (cont'd):

## CI for $p$ the population proportion

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with $B(1, p)$. Denote $Y = \sum_{i=1}^{n} X_i \sim B(n, p)$.

- Find an estimator of $p$ using method of moments and maximum likelihood.
- Use CLT to conclude the sampling distribution of $p$ assuming the variance can be replaced by its sample value.
- Find the sampling distribution involving $\hat{p}$ and $p$
- Find the large sample 95% CI for $p$

# Lecture 4 related questions (cont'd):

## Hypothesis test about $p$ the population proportion

Let $p$ be the proportion of defective circuit boards among all circuit boards produced by a certain manufacturer. Ideally, the proportion should be lower than 0.05. For a randomly sample 500 items, we found the defective rate in this random sample is $\hat{p} = 0.02$. Conduct a formal statistical test and conclude whether you should believe the proportion of defective items is lower than 0.05.

# Lecture 4 related questions (cont'd):

## Hypothesis test - Step by Step

- Write out the random sample: what each variable stand for and the total number of samples.
- Identify the parameter of interest
- Write out the null and alternative hypothesis
- Identify an estimator for the parameter and then construct a test statistic based on it
- Write out the sampling distribution of the test statistic
- Calculate the observed test statitic assuming the null hypothesis is true
- Compare the observed test statistic to the sampling distribution to find the p-value.
  - Alternativey, you might be given the significance threshold to find the rejection region
- Conclude whether the finding was significant or not based on a significance threshold $\alpha$.
  - Alternatively, you can compare the observed test statistic to the rejection region and conclude whether the null should be rejected.

# Lecture 4 related questions (cont'd):

## Hypothesis test about $\mu$ the population mean

Let $X_1, \ldots, X_n$ denote a random sample from a normal population distribution with $\sigma^2$ known.

- If we test the null hypothesis $\mu = \mu_o$ against $H_1 : \mu < \mu_o$ using the test statistic $\bar{X}$, show the rejection region of $\bar{X} < \mu_o + \Phi^{-1}(\alpha)\sigma/\sqrt{n}$ has significance level $\alpha$.

# Lecture 4 related questions (cont'd):

- Step 1: give the sampling distribution of $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.
- Step 2: The rejection region is where the observed test statistic is more extreme than expected at the significance level $\alpha$

$$\mathrm{P}_X(Z < \Phi^{-1}(\alpha)) = \mathrm{P}(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \Phi^{-1}(\alpha)) = \alpha$$

Since under the null $\mu = \mu_o$, $\bar{X} < \mu + \Phi^{-1}(\alpha)\sigma/\sqrt{n}$ can be taken to be $\bar{X} < \mu_o + \Phi^{-1}(\alpha)\sigma/\sqrt{n}$.

# Lecture 4 related questions (cont'd):

## One sample z-tests

A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is 130°F. **A sample of** $n = 9$ **systems**, when tested, yields a sample average activation temperature of 131.08°F. If the distribution of activation times is normal with standard deviation 1.5°F, does the data contradict the manufacturer's claim at significance level $\alpha = 0.01$?

# Lecture 4 related questions (cont'd):

## One sample t-tests

A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is 130°F. **A sample of** $n = 40$ **systems**, when tested, yields a sample average activation temperature of 131.08°F. If the activation times has a sample standard deviation 1.5°F, does the data contradict the manufacturer's claim at significance level $\alpha = 0.01$?

# Lecture 4 related questions (cont'd)

## Two sample t-tests

House Insulation: Whiteside's Data: Mr Derek Whiteside of the UK Building Research Station recorded the weekly gas consumption and average external temperature at his own house in south-east England for two heating seasons, one of 26 weeks before, and one of 30 weeks after cavity-wall insulation was installed.

# Lecture 4 related questions (cont'd)

- Read the outputs from the next slides and answer the questions
  - Was there a difference in gas consumption prior and post installing the insulation wall?
  - If we can accept the population variance of gas consumption before and after to be the same, give a 95% Confidence interval for the mean difference in gas consumption ($t_{0.025,54} = 2.005,\ t_{0.025,55} = 2.004$).

$$T = \frac{\bar{X} - \bar{X}' - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t(n1 + n2 - 2)$$

  - Give a one-sided 99% Confidence interval for the decrease in mean gas consumption ($t_{0.01,54} = 2.397,\ t_{0.01,55} = 2.396$)
  - Suppose we ask was there a decrease in gas consumption after installing the wall, what is the alternative hypothesis and a p-value?

# Lecture 4 related questions - Two sample t-tests Cont'd

```
tapply(whiteside$Gas, whiteside$Insul, length)
```

```
## Before  After
##     26     30
```

```
tapply(whiteside$Gas, whiteside$Insul, mean)
```

```
##   Before    After
## 4.750000 3.483333
```

```
tapply(whiteside$Gas, whiteside$Insul, sd)
```

```
##    Before     After
## 1.1628413 0.8064752
```

```
var.test(whiteside$Gas~whiteside$Insul)
```

```
##
##  F test to compare two variances
##
## data:  whiteside$Gas by whiteside$Insul
## F = 2.079, num df = 25, denom df = 29, p-value = 0.05947
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9706482 4.5533314
## sample estimates:
## ratio of variances
##           2.079021
```

# Lecture 4 related questions - Two sample t-tests Cont'd

```
t.test(whiteside$Gas~whiteside$Insul)
```

```
##
##  Welch Two Sample t-test
##
## data:  whiteside$Gas by whiteside$Insul
## t = 4.6662, df = 43.649, p-value = 2.919e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7194608 1.8138725
## sample estimates:
## mean in group Before  mean in group After
##             4.750000             3.483333
```

```
t.test(whiteside$Gas~whiteside$Insul, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  whiteside$Gas by whiteside$Insul
## t = 4.7868, df = 54, p-value = 1.357e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7361415 1.7971918
## sample estimates:
## mean in group Before  mean in group After
##             4.750000             3.483333
```

# Lecture 4 related questions - Two sample t-tests Cont'd

```
import pandas
import numpy as np
from scipy import stats
whiteside = pandas.read_csv("whiteside.csv")
before = whiteside[whiteside['Insul'] == 'Before']['Gas']
after = whiteside[whiteside['Insul'] == 'After']['Gas']
print([len(before), len(after), before.mean(), after.mean(), before.std(), after.std()])
```

```
## [26, 30, 4.75, 3.4833333333333329, 1.1628413477340751, 0.80647523139311272]
```

```
print(stats.levene(before, after))
```

```
## (4.2867976248903235, 0.043204813060714366)
```

```
print(stats.ttest_ind(before, after))
```

```
## (array(4.786792407777229), 1.3569174842231742e-05)
```

```
print(stats.ttest_ind(before, after, equal_var=False))
```

```
## (array(4.666213847535168), 2.9191081268861838e-05)
```