STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2019

Dr. Shivon Sue-Chee

January 8, 2019

About your instructor

- **Dr. Shivon Sue-Chee**
  - Office: Stewart Building, 149 College, EP 104
  - Office hours:
    - T 12-1pm in EP 104, from Jan 15
    - R 11-12noon in BA1160, from Jan 17
    - (or by appointment)
  - Email: shivon.sue.chee@utoronto.ca
- CLTA-Contract Teaching Professor
- Ph.D. Stats (under Fang Yao at UofT, 2014, "*Partially Functional Quantile Regression*")
- Interests: statistics education, statistical consulting, high-dimensional data, robust regression, survey design

Class Schedule

- **Section L0101**
  - Tues. 10-12noon, Thurs. 10-11am
  - **BA 1160** (Bahen Centre, 40 Saint George St.)

- **Section L0201**
  - Tues. 3-5pm, **KP 108** (Koffler House, 569 Spadina Av.)
  - Thurs. 12-1pm, **MP 202** (Physics Lab, 255 Huron) 11-12

- except Reading Week (Feb. 18-22)
- Building Map: Link
- **TA Office Hours** will be scheduled before tests and assignments due dates.

What is this course about?

Objective: extends the theory and practice of linear regression to indicator variables, and cases where the Gauss-Markov assumptions may not apply:

- Non-linear methods- t-tests, Pearson's chisquare test
- Logistic, Poisson regression and log-linear models (GLM)
- Longitudinal data/ repeated measurements and mixed effect models (GLMM)
- Non-parametric regression

Teaching Approach:

- *Emphasis on application and interpretation*
- *Data driven with case studies*
- course participation highly encouraged

Required prerequisite knowledge

- Basic probability and random variables (expectation and variance)
- Normal, t, F and χ^2_n distributions and properties
- Point estimation (LS, MLE, unbiasedness, MVUE, consistency, BLUE)
- Statistical inference for regression parameters
- Simple linear regression (SLR) in scalar form
- Multiple linear regression (MLR) in matrix form, including all standard results from STA302
- First and second year calculus, linear algebra

Q: Did you take STA 221/ 248/ 255 and/or 261?

Who can take this course?

- Undergrads and Grads with STA302/1001 or equivalent preparation
- Notes:
  - **Pre-requisites are strictly enforced by the department**
  - Instructor does not handle registration
  - Email Gillis (gillis.aning@utoronto.ca) if you deferred the STA302 exam or have a transfer credit, to make sure you won't be removed.

Recommended Textbooks

- *Categorical Data Analysis, 3rd edition* by A. Agresti (Wiley)
  Chapters 2, 4, 5, and 6. (On reserve at the Math Library)

- *Applied Linear Regression Models, 4th edition* by Kutner, Nachtsheim, and Neter (Mc-Graw Hill).
  Chapters 8, 11, 13 and 14. (On reserve at the Math Library)

For Egs

- *A Modern Approach to Regression with R* by S. J. Sheather (Springer)
  Chapters 8, 9 and 10. Available as an e-resource via UT library.

- *Applied linear regression, 4th edition* by S. Weisberg (Wiley).
  Third edition available as an e-resource via UT library.

For Egs

- *TSS: The Statistical Sleuth, 3rd edition* by Ramsey and Schafer (Brooks/Cole). Datasets are available in R.

How will you be evaluated?

	Sch. 1*	Sch. 2*	Date	Time
Quizzes*	0%	8%	from Jan. 23	
Assignment 1	4.5%	4.5%	F, Jan. 25	due by 10pm
Assignment 2	7.5%	7.5%	F, Feb. 15	due by 10pm
Term Test*	33%	25%	R, Feb. 28	10:10-11:40 (L01) 11:10-12:40 (L02)
Assignment 3	10%	10%	R, Mar. 21	due by 10pm
Final Exam	45%	45%	Btw Apr. 6-30	(3 hours)

*Your final grade will be your better performance of the two schemes.

Grad students will be evaluated based on a slightly different scheme.

PARTICIPATION QUIZZES (8%*)

- Via Class Quizzes or Online Surveys
- To promote in-class engagement and provide formative feedback on your understanding
- Roughly 1 quiz/survey per week
- *Participation is OPTIONAL!
- Do not need to answer correctly; your attempt counts.
- Participation starts to count from week of Jan. 22
- No makeups

PARTICIPATION QUIZZES (8%*)

- Via Class Quizzes or Online Surveys
- To promote in-class engagement and provide formative feedback on your understanding
- Roughly 1 quiz/survey per week
- *Participation is OPTIONAL!
- Do not need to answer correctly; your attempt counts.
- Participation starts to count from week of Jan. 22
- No makeups

PARTICIPATION QUIZZES (8%*)

- Via Class Quizzes or Online Surveys
- To promote in-class engagement and provide formative feedback on your understanding
- Roughly 1 quiz/survey per week
- *Participation is OPTIONAL!
- Do not need to answer correctly; your attempt counts.
- Participation starts to count from week of Jan. 22
- No makeups

ASSIGNMENTS (22%)

- Data analysis projects for practical experience
- Need to use `R` (and RStudio)
- PDF compilation preferred
- Submitted online into Crowdmark by due times.
- Late assignments will be subject to a 20% penalty per day late.
- Expect Assignment 1 by the end of next week

MIDTERM TEST AND FINAL EXAM

- Locations: TBA

- Closed book and closed notes
- Relevant values, formulas and tables provided
- Need a non-programmable scientific calculator for class/quizzes/ test/ exam
- Walk with Photo ID

- No makeup test
- All final exam matters are governed by FAS
- Accessibility accommodations available at: https://www.studentlife.utoronto.ca/as

Where to get help?

- **Don't spin your wheels, ask for help!!**
- Do practice problems.
- Try posting on the discussion forum.
- Visit the instructor or TAs during office hours
- Email the instructor in cases of emergencies or personal matters

Some variables of interest

1. Favourite season
2. Height in inches
3. Hair color
4. Area of interest
5. Sex /Gender
6. Weight in kg
7. Expected final grade
8. Eye color
9. Weekly food expenditure

x

y

x

Other X's

Age

Sleep hrs

Body fat

Exercise

- Compare an experiment to an observational study.
- Which random variables are categorical?
- What can we use a two-sample t-test for?
- How can we establish a two-way contingency table?
- What plots can be used to describe the data?

Week 1 Topics

REVIEW

-Data summary: Five-number summary, Boxplots

-Large-sample distribution theory: derived from Normal

-Statistical inference: confidence interval, hypothesis tests, errors, power

-Normality Test, Equal variance test

T-TESTS

-One-sample t-test

-Paired t-test

-Two-sample t-test

-Non-parametric alternatives

Parameters and Statistics

θ $\hat{\theta}$

What is the difference between a parameter and a statistic?

- A parameter is a population quantity and a statistic is a quantity based on a sample drawn from the population.

Example: The population of all adult (18+ years old) males in Toronto, Canada.

- Suppose that there are N adult males and the quantity of interest, y, is age.
- A sample of size n is drawn from this population.
- The population mean is $\mu = \sum_{i=1}^{N} y_i / N$.
- The sample mean is $\bar{y} = \sum_{i=1}^{n} y_i / n$.

The Normal Distribution

The density function of the normal distribution with mean μ and standard deviation σ is:

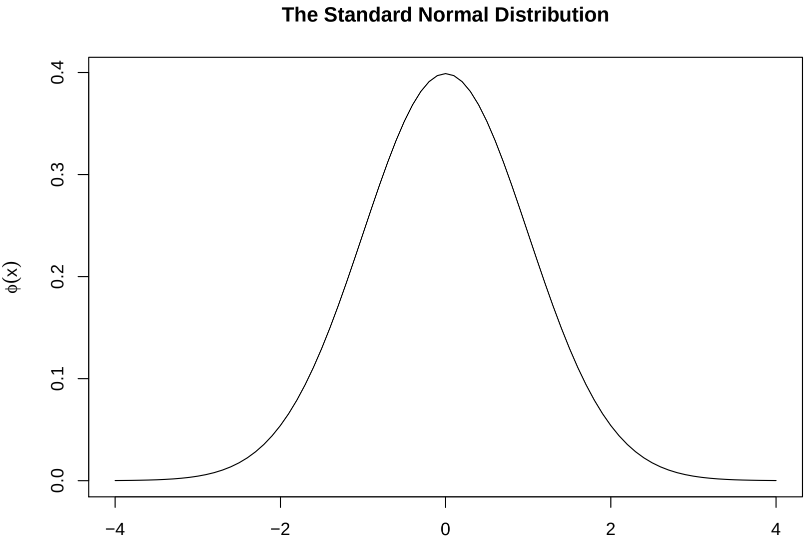
$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

The cumulative distribution function (CDF) of a $N(0,1)$ distribution,

$$\Phi(x) = P(X < x) = \int_{-\infty}^{x} \phi(x)dx$$

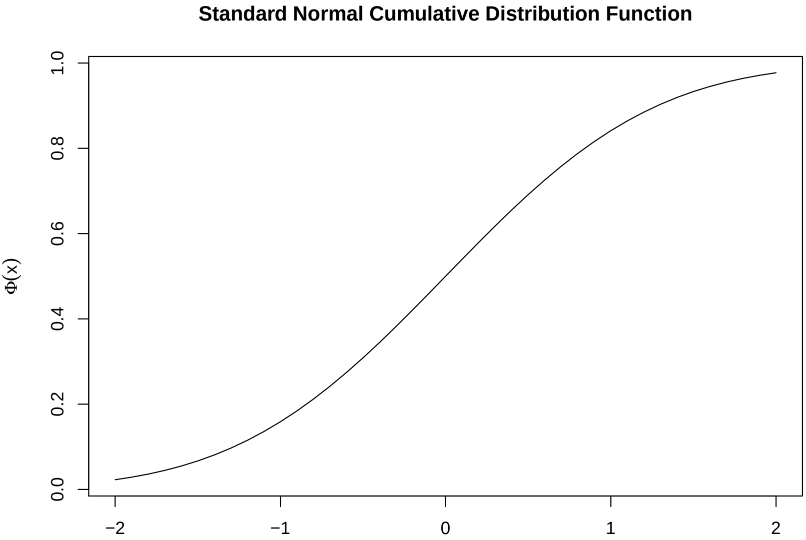
The Standard Normal Distribution

```
x <- seq(-4,4,by=0.1)
plot(x,dnorm(x),type="l",main = "The Standard Normal Distribution",
     ylab=expression(paste(phi(x))))
```



The Standard Normal CDF

```
plot(x <- seq(-2,2,by=0.1),pnorm(x),type="l",
     xlab="x",ylab=expression(paste(Phi(x))),
     main = "Standard Normal Cumulative Distribution Function")
```



The Normal and Standard Normal Distributions

A random variable X that follows a normal distribution with mean μ and variance σ^2 will be denoted by

$$X \sim N\left(\mu, \sigma^2\right).$$

If $X \sim N\left(\mu, \sigma^2\right)$ then

$$Z \sim N(0,1),$$

where

$$Z = \frac{X - \mu}{\sigma}.$$

The Normal Distribution

$X \sim N(0, 1)$. Use R to find $P(-2 < X < 2)$.

```
pnorm(2,mean = 0,sd = sqrt(1))-pnorm(-2,mean = 0,sd = sqrt(1))
```

```
## [1] 0.9544997
```

Normal Quantile-Quantile Plots

-used to visually assess Normality of a sample of measurements

-in R, use `qqnorm()` for the normal qq plot and `qqline()` to add the straight line.

Linear combination of independent Normals

If $X_i \sim N(\mu_i, \sigma_i^2)$ independently, then

$$V = a + \sum_1^n b_i X_i \sim N(a + \sum_1^n b_i \mu_i, \sum_1^n b_i^2 \sigma_i^2)$$

Chi-Square Distribution

Let $X_1, X_2, ..., X_n$ be independent and identically distributed random variables that have a $N(0, 1)$ distribution. The distribution of

Z

$$\sum_{i=1}^{n} X_i^2,$$

$Z^2 \sim \chi^2_1$

$\sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$

has a chi-square distribution on n degrees of freedom or χ^2_n.

The mean of a χ^2_n is n with variance $2n$.

Chi-Square Distribution

Let $X_1, X_2, ..., X_n$ be independent with a $N(\mu, \sigma^2)$ distribution. What is the distribution of the sample variance $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$?

$$\frac{(n-1)\,S^2}{\sigma^2} \sim \chi^2_{n-1}$$

t Distribution

If $X \sim N(0, 1)$ and $W \sim \chi_n^2$ then the distribution of $\frac{X}{\sqrt{W/n}}$ has a t distribution on n degrees of freedom or $\frac{X}{\sqrt{W/n}} \sim t_n$.

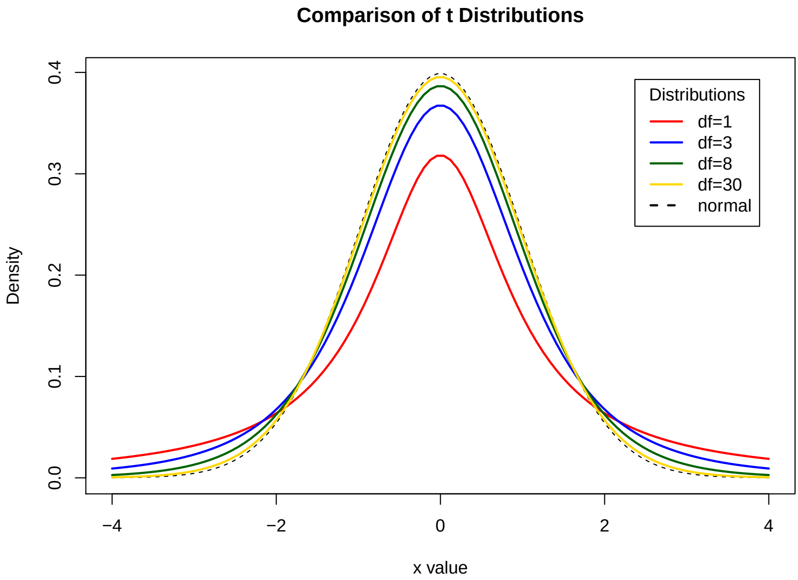
t Distribution

Let $X_1, X_2, ...$ is an independent sequence of identically distributed random variables that have a $N(0,1)$ distribution. What is the distribution of

$$\boxed{\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}} \sim t_{n-1}$$

where $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$?

t Distribution



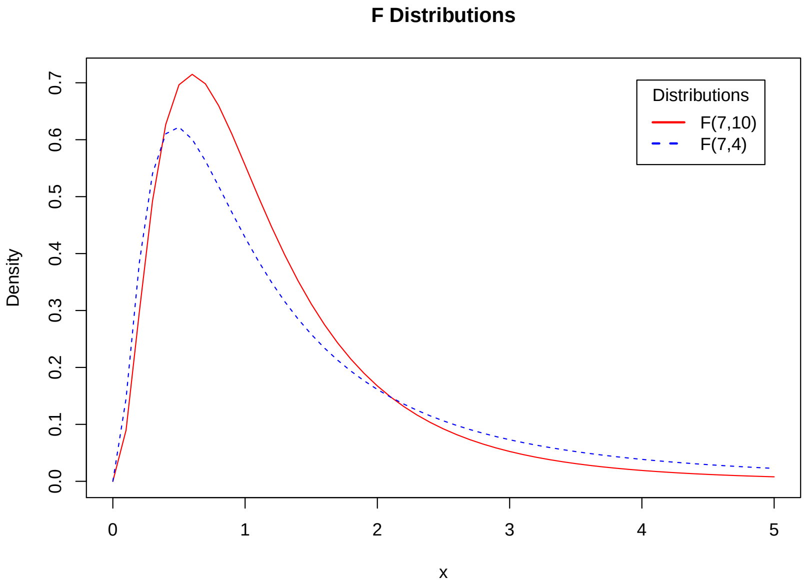
F Distribution

Let $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ be independent. The distribution of

$$W = \frac{X/m}{Y/n} \sim F_{m,n},$$

where $F_{m,n}$ denotes the F distribution on m, n degrees of freedom. The F distribution is right skewed (see graph below). For $n > 2, E(W) = n/(n-2)$. It also follows that the square of a t_n random variable follows an $F_{1,n}$.

F Distribution



The Sample Mean

If $X_1, \ldots, X_n \sim_{iid} N(\mu, \sigma^2)$ then

- $\bar{X} \sim N(\mu, \sigma^2/n)$
- $S^2 = \sum (X - \bar{X})^2/(n-1)$ and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

- $\bar{X} \perp S^2$ and
-

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \boxed{\frac{\bar{X}-\mu}{S/\sqrt{n}}} \sim t_{n-1}$$

Simple Linear Regression

A simple linear regression model is obtained by estimating the intercept and slope in the equation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, ..., n$$

where $\epsilon_i \sim N(0, \sigma^2)$. The values of β_0, β_1 that minimize the sum of squares

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2,$$

are called the least squares estimators. They are given by:

- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- $\hat{\beta}_1 = r \frac{S_y}{S_x}$

r is the correlation between y and x, and S_x, S_y are the sample standard deviations of x and y respectively.

Case Study 1: The Spock Conspiracy Trial

- Boston, 1968
- Dr. Benjamin Spock (paediatrician and author) on trial for conspiring to violate the Selective Service Act.
- Accused of encouraging people to dodge military draft by his books that adviced on how mothers should raise children.
- Spock's jury had NO women.

Q: Is there evidence of gender bias in the jury selection for Spock's trial?