INTRODUCTION TO PROBABILITY MODELS

Lecture 41

Qi Wang, Department of Statistics

Dec 5, 2018

REMINDERS

- The final exam will be from **10:30am to 12:30pm, Dec 12, 2018** in **CL50**

FINAL EXAM

- Cumulative, about 70% will be on the material covered after Exam2
- 2-Hour Exam, 125 points
- You are allowed the following aids
  - 2 one-page 8.5” x 11” HANDWRITTEN cheat sheets
  - Scientific (non-graphing) calculator (in accordance with the syllabus)
  - Pencils, pens, erasers

MATERIAL COVERED AFTER EXAM2

- Normal Distribution: definition, parameter, PDF, CDF, expected value, variance, standard Normal, Z-score, emprical rules, approximation to Binomial
- Five Number Summary and Boxplot
- Types of Data, summarizing Data and graphs
- Contigency Table and χ^2 test
- Scatterplot, correlation and linear regression

NORMAL RANDOM VARIABLE

- **Parameter:**
  - μ: the mean of the random variable, determines the center of the distribution
  - σ: the standard deviation of the random variable, determines the shape of the distribution
- The standard normal distribution is the normal distribution with $\mu = 0, \sigma = 1$, namely, $X \sim N(\mu = 0, \sigma = 1)$
- The CDF of standard normal distribution is denoted as $\Phi(x)$
- You convert $X \sim N(\mu, \sigma)$ to $Z \sim N(\mu = 0, \sigma = 1)$, where Z has the standard Normal distribution. Convert/standardize using:
  $$Z = \frac{X - \mu}{\sigma}$$
  This standardized value is called a Z-score
- Remember that your table gives you the probability $P(Z \leq z) = \Phi(z)$

- Steps to finding the sample score if you are given a probability and know $X \sim N(\mu, \sigma)$

  1. Set up your problem as follows $P(Z \leq z_0) = probability$(Note: adjust $>$to $\leq$ if necessary by using "1-probability".)
  2. Find the z-score by looking up the probability in

2. Find the z-score by looking up the probability in the body of normal table
3. If you have a two-sided probability, use
   $P(-z_0 < Z \leq z_0) = 2P(Z \leq z_0) - 1 = 2\Phi(z_0) - 1$
4. Convert the z-score to x using
   $$z = \frac{x - \mu}{\sigma}$$

BOXPLOT

Boxplot is a graphic depiction of the 5 number summary

1. Draw a horizontal or vertical axis that is evenly spaced and well-labeled(make sure it covers the full range of the data)
2. Locate Q_1 and Q_3. There are the "ends" of your box. Draw the box.
3. With the box, locate the Median and mark it
4. Locate and mark the Minimum and Maximum. Extend a line("whisker") from each end of the box to the Max or Min

To draw a modified boxplot, Step 1, 2, 3 are the same, BUT we indicate the outliers with a o or a $\star$. Then draw the line from the ends of the box ot the highest or lowest data point that is NOT an outlier. Most software generate boxplots are modified boxplots.

CONTINGENCY TABLE

- Describes the relationship between two categorical variables, represents a table of counts (can include percentages).
- Calculate joint, conditional marginal probability

Test if there is a relationship between two qualitative (categorical) variables via Chi-Square(χ^2) Hypothesis test

1. State the Null and Alternative hypothesis
2. Determine the confidence level and the significan level α
3. Find the test statistic
$$\chi^2 = \sum \frac{(observed\ cell\ count - expected\ cell\ count)}{expected\ cell\ count}$$
4. Determine the degrees of freedom needed to use the χ^2 table
5. Find the χ^2 critical value from the χ^2 table. Compare critical value from the table to the calculated χ^2 value.
6. State the conclusion in terms of the problem

LEAST-SQUARES REGRESSION

- Minimizes $\sum_i^n e_i^2$
- Equation of the line is: $\hat{y} = b_0 + b_1 x$
- **Slope** of the line is:b_1, where the slope measures the amount of change caused in the response variable when the explanatory variable is increased by one unit.
- **Intercept** of the line is:b_0, where the intercept is the value of the response variable when the explanatory variable = 0. (i.e. value where line intersects the y-axis)
- Used for Prediction: using the line to find y-values corresponding to x-values that are within the range of your data x-values
- Using values outside range of the collected data can lead to **extrapolation**
- Coefficient of Determination: Denoted by r^2, it gives the proportion of the variance of the response variable that is predicted by the explanatory variable. So when r^2 is high, close to 1 or 100%, you have explained most of the variability. Also, it equals to the equare of the correlation between x and y, $r^2 = r_{xy}^2$
- Residuals: the difference between the observed value of the response variable (y) and the predicted value ($\hat{y}$): residuals = observed y - predicted y, $e = y - \hat{y}$

MATERIAL COVERED BEFORE EXAM2

- Refer to <u>**Lecture 15**</u> for a summary of materials before Exam 1
- Refer to <u>**Lecture 21**</u> for a summary of discrete random variables
- Refer to <u>**Lecture 28**</u> for a summary of materials after Exam 1
- Discrete
    - Bernoulli
    - Binomial
    - Hypergeometric
    - Poisson
    - Geometric
    - Negative Binomial
- Continuous
    - Uniform
    - Exponential