Stats Exam 1 Review Slides Notes

Links: Math 3339

(based on slides 15 and 16 on the review powerpoint)

Possible Multiple Choice Questions

Detecting Outliers

- individual values that falls outside the overall pattern
- mean is affected by outliers, median is resistant to outliers in the dataset

Use the Interquartile Range (IQR) to detect outliers:
any point outside this interval is an outlier:
$Q_1 - 1.5(IQR)$ (which is the lower limit) and $Q_3 + 1.5(IQR)$ (which is the upper limit)

Remember IQR is calculated by: $IQR = Q_3 - Q_1$
In R you can get the IQR, Q1, and Q3 using the five number summary:

```
fivenum(dataset) # outputs the following: min, Q1, Q2, Q3, max
```

Looking at the graphs and determine the shape

Shape

- symmetric: right and left sides of the graph are approximateky mirror images of each other
- skewed to the right: the right side (higher values) extends much farther out than the left side (longer right tail)
- skewed to the left: if the left side (lower values) extends much father out than the right side (longer left tail)

- uniform: the graph is the same height (frequency) from lowest to highest value of the variable

Graphs

- Grpahs for Categorical Variables:
  - Bar Graphs: Each individual bar represents a category and the height of each of the bars are either represented by the count or percent.
    - In R: `barplot(table(dataset$variableName))`
  - Pie charts: Helps us see what part of the whole each group forms.
    - In R: `pie(table(dataset$variableName))`
- Graphs for quantitative variables:
  - Dot Plots: putting dots above the values listed on a number line.
    - In R: `dotchart(dataset$variableName)`
  - Stem and leaf plot: Separate each observation into a stem consisting of all but the final rightmost digit and a leaf, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.
    - In R: `stem(dataset$variableName)`
  - Histogram: The width of the bar represents an interval of values (range of numbers) for that variable. The height of the bar represents the number of cases within that range of values.
    - In R: `hist(dataset$variableName)`
  - Boxplots: a central box spans the quartiles, a line inside the box indicates the median, lines extend from the box out to the smallest and largest observations, asterisks represents any values that are considered to be outliers
    - In R: `boxplot(dataset$variableName)`
  - Scatterplot: values of one variable are on x (horizontal) axis and the other on the y (vertical) axis scatterplots can show if there is some kind of association between the two quantitative variables
    - In R: `plot(explanatory, response)`
    - response variable (or dependent variable): measures outcome of study (typically the *y-variable*)

    - explanatory variable (or independent varuable): explains or influences changes in a response variable (typically the *x-variable*)

Know the difference between the types of variables

- **Categorical variables** place a case into one of several groups or categories.
- **Quantitative Variables** take numerical values for which arithmetic operations such as adding and averaging make sense.
    - Discrete quantitative variables - a countable set of values.
    - Continuous quantitative variables - data that can take on any values within some interval.

Using Probability Rules

Rules

- Complement Rule: $P(E^C) = 1 - P(E)$
- Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Multiplication Rule: $P(A \cap B) = P(A) * P(B|A)$
- Other:
    - $P(A \cap B^C) = P(A) - P(A \cap B)$
    - $P(A^C \cap B) = P(B) - P(A \cap B)$
- Conditional Probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Baye's Rule: $P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$

When to add and when to multiply

- add when finding chance of A *or* B *or* Both happening (use Addition Rule)
- multiply when finding chance of bott A *and* B happeneing (use multiplication rule)

Probabilities from Discrete Distribution Table

Know how to find probabilities from a discrete distribution table (& binomial/poisson/hypergeometric distribution)

The **probability mass function** (pmf) of a discrete random variable (r.v.) is defined for every number x_i by $f(x_i) = P(X = x_i)$
Properties of the pmf function:

- $f(x) \geq 0$ for all $x \in \mathbb{R}$ (the pmf is greater than or equal to 0)
- $\Sigma_i f(x_i) = 1$ (adding all the pmf's together equals 1)

pmf for binomial, poisson, and hypergeometric distributions are on formula sheet.
R commands:

- Binomial:
  - $P(X = x)$: `dbinom(x,n,p)`
  - $P(X \leq x)$: `pbinom(x ,n, p)`
  - $P(X > x) = 1 - P(X \leq x)$=`1-pbinom(x,n,p)`
- Hypergeometric:
  - single value calculation $P(X = x)$: `dhyper(x,m,n,k)`
  - $P(X \leq x)$: `phyper(x,m,n,k)`
- Poisson:
  - $P(X = x)$: `dpois(x, mu)`
  - $P(X \leq x)$: `ppois(x, mu)`

Expected Value and Variance from Discrete Distribution Table

Know how to find expected value and variance from a discrete distribution table and binomial/Poisson/hypergeometric distribution and from a linear expression

The **expected value** (or mean of the distribution) of a random variable *X* is given by:
$E[X] = \mu = \sum_x x * f(x) = \sum_{i=1}^{n} x_i * p_i$ (*p* is probability, remember f(x) is pmf)
The **variance** of a random variable *X* is
$\sigma^2 = Var[X] = E[(X - \mu)^2] = E[X^2] - (E[X])^2$
(σ^2 represents the variance)

the expected values and variance for binomial, poisson, and hypergeometric distributions are on formula sheet.

Possible Free Response Questions

Determining outliers, shape, center and spread from descriptive numerical values

(see the multiple choice section for detecting outliers)

four main characteristics to describe a distribution

- Shape
- Center
- Spread
- Outliers

Shape

- symmetric: right and left sides of the graph are approximateky mirror images of each other
- skewed to the right: the right side (higher values) extends much farther out than the left side
- skewed to the left: if the left side (lower values) extends much father out than the right side
- uniform: the graph is the same height (frequency) from lowest to highest value of the variable

Center
the values with roughly half the observations taking smaller values and half taking larger values

Spread
from the graphs we describe the spread of a distribution by giving *smallest and largest values.* (range=max-min)

Outliers

individual values that falls outside the overall pattern

Probability rules, including the Baye’s rule

(see the multiple choice section above)

Know when two events are independent

The probability of one does not affect the probability of the other

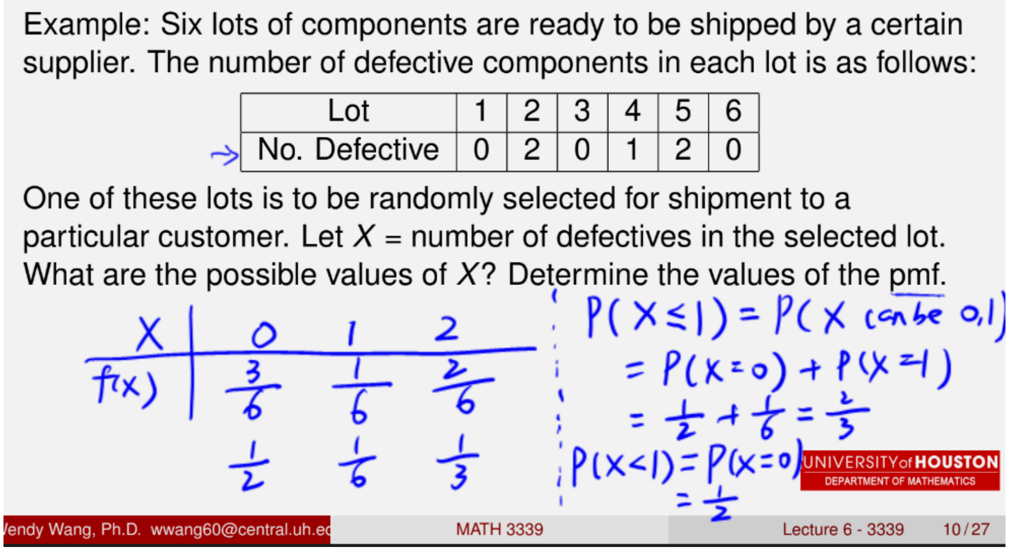
$Pr(A \cap B) = Pr(A) * Pr(B)$

$Pr(A|B) = Pr(A)$

Know how to find probabilities from a discrete distribution table

(see the multiple choice section above)

(simple) Example from lecture:



Example: Six lots of components are ready to be shipped by a certain supplier. The number of defective components in each lot is as follows:

Lot	1	2	3	4	5	6
→ No. Defective	0	2	0	1	2	0

One of these lots is to be randomly selected for shipment to a particular customer. Let X = number of defectives in the selected lot. What are the possible values of X? Determine the values of the pmf.

X	0	1	2
$f(x)$	$\frac{3}{6}$	$\frac{1}{6}$	$\frac{2}{6}$
	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{3}$

$P(X \leq 1) = P(X \text{ can be } 0,1)$

$= P(X=0) + P(X=1)$

$= \frac{1}{2} + \frac{1}{6} = \frac{2}{3}$

$P(X<1) = P(X=0)$

$= \frac{1}{2}$

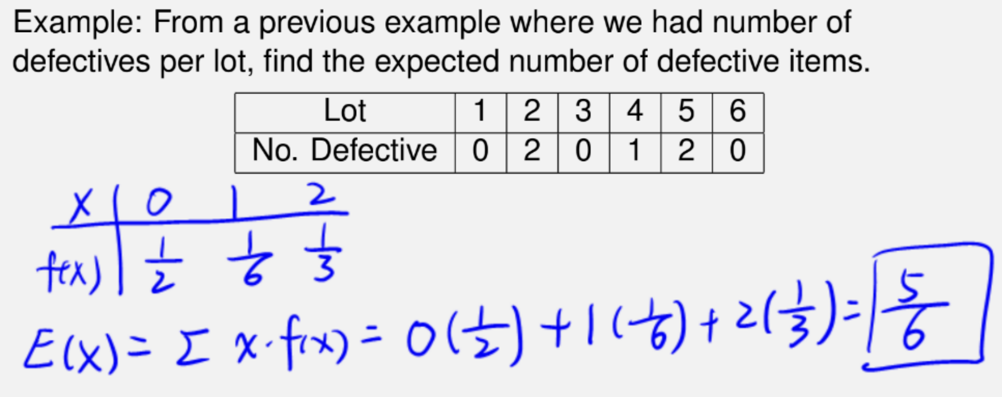
UNIVERSITY of HOUSTON DEPARTMENT OF MATHEMATICS

Vendy Wang, Ph.D. wwang60@central.uh.ec MATH 3339 Lecture 6 - 3339 10 / 27

Know how to find expected value from a discrete distribution table

(see the multiple choice section above)

(simple) Example from lecture:



Example: From a previous example where we had number of defectives per lot, find the expected number of defective items.

Lot	1	2	3	4	5	6
No. Defective	0	2	0	1	2	0

x	0	1	2
$f(x)$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{3}$

$$E(x) = \Sigma\, x \cdot f(x) = 0\left(\tfrac{1}{2}\right) + 1\left(\tfrac{1}{6}\right) + 2\left(\tfrac{1}{3}\right) = \boxed{\tfrac{5}{6}}$$

Creating a least-squares linear equation from the data, scatterplot, correlation, coefficient of determination, and residual

(see multiple choice section above for scatterplot)

least squares regression line (LSRL)

- of *Y* on *X* is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.
- the linear regression model is: $Y = \beta_0 + \beta_1 x + \epsilon$
  - *Y* is the dependent variable (response)
  - *x* is the independent variable (explanatory)
  - β_0 is the population intercept of the line
  - β_1 is the population slope of the line
  - ϵ is the error term which is assumed to have a mean value 0. This is a random variable that incorporates all variation in the dependent variable due to factors other than *x*
  - The variability: σ of the response *y* about this line. More precisely the standard deviation of the deviations of the errors, ϵ_i in the regression model
- gather information from a sample so we will have the least squares estimates model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

How to use R to get this estimate:

```
lm(dataset$responseVariable ~ dataset$explanatoryVariable)
# i.e. lm(y ~ x)
# you can also use to those c(thing1, thing2, etc.) array things for x & y
```

in the output under the "Coefficients" section, look under "Estimate", the number next to the "(Intercept)" is the intercept, and the number next to (or below) it is the slope.

coefficient of determination

- r^2 (coefficient of determination, remember just r is correlation coefficient) is the percent (fraction) of variability in the response variable (*Y*) that is explained by the least squares regression with the explanatory variable.
- measure of how succesful the regression equation was in predicting the response variable.
- the closer r^2 is to one (100%) the better our equation is at predicting the response variable.
- In the R output is its the *Multiple R-squared* value

residuals
A **residual** is the difference between an observed value of the repsonse variable and the value predicted by the regression line (another way to look at if the model is good or not).
$residual = observerd\, y - predicted\, y$ (aka $y - \hat{y}$)

- we can determine residuals for each observation.
- the closer the residuals are to zero, the better we are at predicting the resposne variable
- we can plot the residuals for each observation, these are called the residual plots

in R:

```
# input the data
name.lm = lm(y~x)
plot(x, resid(name.lm))
abline(0,0) # graphs another line where intercept and slope is zero
# make that line so you can compare the residual plot to it
```

examining the residual plot

- curved pattern: relationship is not linear
- increasing spread about the zero line as x increases indicates the prediction of y will be less accurate for larger x. Decreasing spread about the zero line as x increases indicates the prediction of y to be more accurate for larger x.
- individual poitns with larger residuals are considered outliers in the vertical (y) direction.
- individual points that are extreme in the x direction are considered outliers for the x-variable.
- if original plot shows non-linar relationship, you can do data transformations (like logs, etc.)