MATH 3339
Statistics for the Sciences
Live Lecture Help

James West
jdwest@uh.edu

University of Houston

Session 5

Office Hours: see schedule in the "Office Hours" channel on Teams
Course webpage: www.casa.uh.edu

Email policy

When you email me you **MUST** include the following

- MATH 3339 Section 20024 and a description of your issue in the **Subject Line**
- Your name and ID# in the **Body**
- Complete sentences, punctuation, and paragraph breaks
- Email messages to the class will be sent to your Exchange account (user@cougarnet.uh.edu)

Updates

- Test 1 scheduling opens at midnight tonight!

Using R and R-Studio

1. Download R from https://cran.r-project.org/
2. Download R-Studio from https://www.rstudio.com/

Outline

Least-Squares Regression

- The **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: $Y = \beta_0 + \beta_1 x + \varepsilon$
  - Y is 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 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, σ is the standard deviation of the deviations of the errors, ϵ_i in the regression model.
- We will gather information from a sample so we will have the least squares estimates model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x.$

Least-Squares Regression

Formulas:

$$\hat{Y} = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$b_1 = \hat{\beta}_1 = cor(x, y) \cdot \frac{s_y}{s_x}$

$b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Correlation Coefficient

Correlation Coefficient, r:

The correlation measures the strength and direction of the linear relationship between two quantitative variables.

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

The correlation, r is an average of the products of the standardized values of x and y for a sample of n data pieces.

Correlation Coefficient

Facts about Correlation:

1. Positive r indicates positive association and negative r indicates negative association between variables.
2. r is always between -1 and 1
3. Correlation is strongly influenced by outliers.

The R command for the correlation is cor(x,y)

Coefficient of Determination

The **coefficient of determination** is a measure that allows us to determine how certain one can be in making predictions with the line of best fit. It measures the proportion of the variability in the dependent variable that is explained by the regression model through the independent variable.

- The coefficient of determination is obtained by squaring the value of the correlation coefficient.
- The symbol used is r^2
- Note that $0 \leq r^2 \leq 1$
- r^2 values close to 1 would imply that the model is explaining most of the variation in the dependent variable and may be a very useful model.
- r^2 values close to 0 would imply that the model is explaining little of the variation in the dependent variable and may not be a useful model.

Sum of Squares

The Total Sum of Squares tells you how much variation there is in the dependent variable.

$$SS(tot) = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The Residual Sum of Squares (SS(resid)), is found by:

$$SS(resid) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

And,

$$r^2 = 1 - \frac{SS(resid)}{SS(tot)}$$

Because of this, r^2 is interpreted as the proportion of observed y variation that can be explained by the simple linear regression model.

Random Variables

Suppose an experiment is conducted. A **random variable** is a function whose domain is the sample space Ω of the random experiment.

Example: A coin is flipped resulting in either heads (H) or tails (T) and the random variable X is defined by

$$X(H) = 1 \qquad X(T) = 0$$

We say that the random variable X indicates heads.

$P(X = 1) = \frac{1}{2}$

$P(X = 0) = \frac{1}{2}$

Discrete Random Variables

Def: The **probability mass function** (pmf) of a discrete rv is defined for every number x_i by $f(x_i) = P(X = x_i)$.

Properties of f:

1. $f(x) \geq 0$ for all $x \in \mathbb{R}$
2. $\sum_x f(x) = 1$
3. $P(X \in A) = \sum_{x \in A} f(x)$, where $A \subset \mathbb{R}$ is a discrete set.

Discrete Random Variables

The Cummulative Distribution Function:
Def: The **cumulative distribution function** (cdf) $F(x)$ of a discrete rv X with pmf $f(x)$ is defined for every number x by $F(x) = P(X \leq x)$.

For any number x, $F(x)$ is the probability that the observed value of X will be at most x.

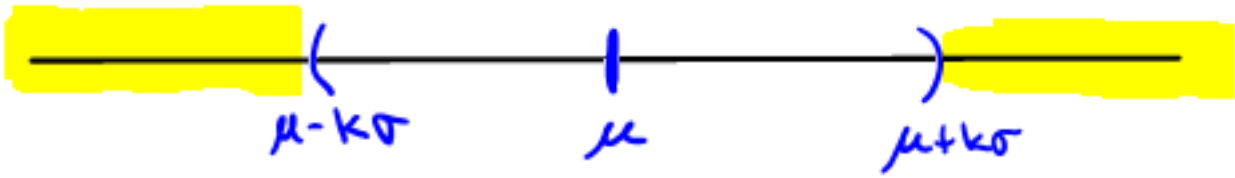
Chebyshev's Inequality

- Chebyshev's inequality, places a universal restriction on the probabilities of deviations of random variables from their means.

- If X is a random variable with mean μ and standard deviation σ and if k is a positive constant, then

$$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}$$

- That is: For any random variable, if we are k standard deviations away from the mean, then no more than $\frac{1}{k^2} * 100\%$ is beyond k standard deviations from the mean. Or at least $(1 - \frac{1}{k^2}) * 100\%$ is within k standard deviations from the mean.



Expected Values

The **expected value** or mean of the distribution of a random variable X is given by:

$$E[X] = \mu = \sum_{x} x \cdot f(x) = \sum_{i=1}^{n} x_i \cdot p_i$$

Variance

The variance of a rv X is

$$\sigma^2 = Var[X] = E[(X-\mu)^2] = E[X^2] - E[X]^2$$

Expected Values and Variance

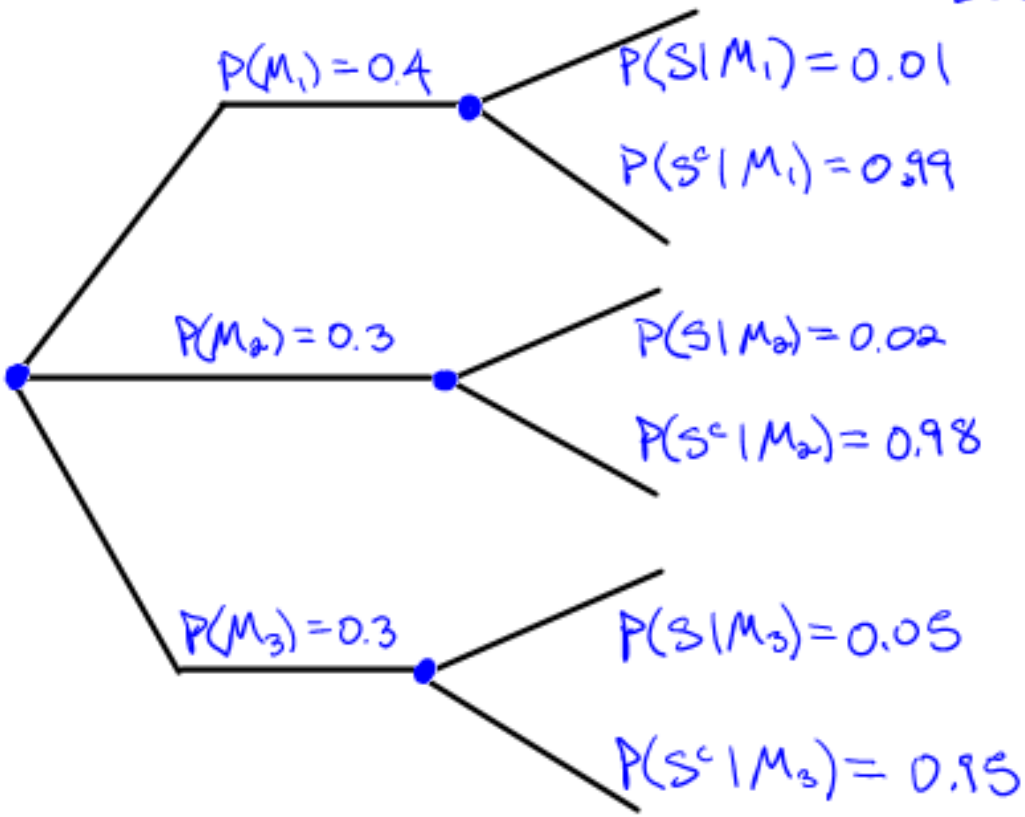
Properties of Expected values and Variance

1. $E[c] = c$ for any constant $c \in \mathbb{R}$
2. $E[aX + b] = aE[X] + b$
3. $E[aX + bY] = aE[X] + bE[Y]$
4. $E[h(X)] = \sum_x h(x) \cdot f(x)$
5. $Var[aX + b] = a^2 Var[X]$
6. $Var[aX + bY] = a^2 Var[X] + b^2 Var[Y]$

A party is held where everyone is offered and eats exactly one meal option. Of those in attendance 40% prefer the first meal (tacos), 30% prefer the second (pizza), and everyone else prefers the third (hot dog). Of the people who ate tacos, 1% got sick. Of the people who ate pizza, 2% got sick. 5% of the people who ate a hot dog got sick.

1. Draw a tree diagram for this problem.

Let M_i be event ate meal i
Let S be event got sick



A party is held where everyone is offered and eats exactly one meal option. Of those in attendance 40% prefer the first meal (tacos), 30% prefer the second (pizza), and everyone else prefers the third (hot dog). Of the people who ate tacos, 1% got sick. Of the people who ate pizza, 2% got sick. 5% of the people who ate a hot dog got sick.

2. If a guest is randomly selected, what is the probability that the guest ate pizza and did not get sick?

Want $P(M_2 \cap S^c) = P(M_2) \cdot P(S^c | M_2)$

$= 0.3 \cdot 0.98 = 0.294$

A party is held where everyone is offered and eats exactly one meal option. Of those in attendance 40% prefer the first meal (tacos), 30% prefer the second (pizza), and everyone else prefers the third (hot dog). Of the people who ate tacos, 1% got sick. Of the people who ate pizza, 2% got sick. 5% of the people who ate a hot dog got sick.

3. Given that a guest got sick, what is the probability that the guest ate hot dogs?

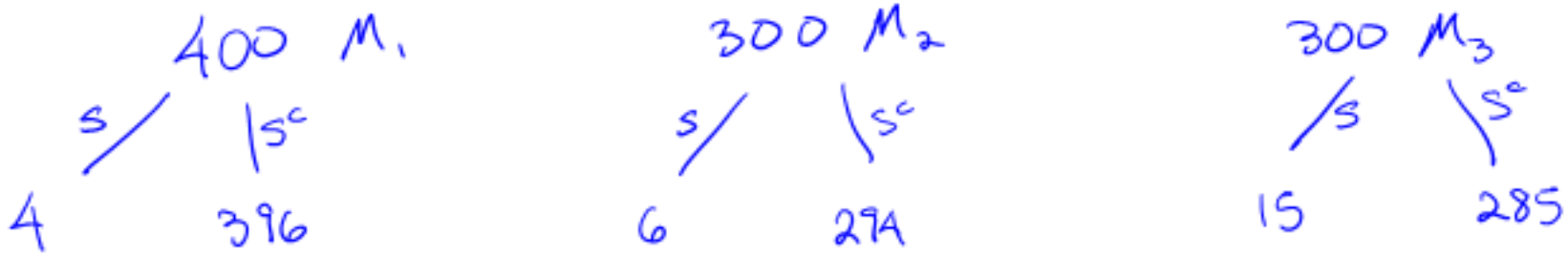
Want $P(M_3|S) = \dfrac{P(M_3 \cap S)}{P(S)}$

$P(M_3 \cap S) = P(M_3) \cdot P(S|M_3)$

$P(S) = P(S \cap M_1) + P(S \cap M_2) + P(S \cap M_3)$

$$P(M_3|S) = \frac{0.3 \cdot (0.05)}{0.4 \cdot (0.01) + 0.3 \cdot (0.02) + 0.3 \cdot (0.05)} = 0.6$$

Suppose 1000 people attend



A certain company sends 50% of its overnight mail parcels via express mail service E1. Of these parcels, 1% arrive after the guaranteed delivery time (denote the event late delivery by L). Suppose that 10% of the overnight parcels are sent via express mail service E2 and the remaining 40% are sent via E3. Of those sent via E2 only 2% arrive late, whereas 5% of the parcels handled by E3 arrive late.

a. Draw a tree diagram for this problem.
b. If a record of an overnight mailing is randomly selected from the company's file, what is the probability that the parcel went via E2 and was late?
c. What is the probability that a randomly selected parcel arrived on time?
d. If a randomly selected parcel has arrived late, what is the probability that is was not sent via E1?

(d) Want $P(E1^c \mid L) = \dfrac{P(E1^c \cap L)}{P(L)}$

$$= \frac{P(E2 \cap L) + P(E3 \cap L)}{P(E1 \cap L) + P(E2 \cap L) + P(E3 \cap L)}$$

$$= \frac{P(E2) \cdot P(L \mid E2) + P(E3) \cdot P(L \mid E3)}{P(E1) \cdot P(L \mid E1) + P(E2) \cdot P(L \mid E2) + P(E3) \cdot P(L \mid E3)}$$

$$P(|X-\mu| > k\sigma) \leq \frac{1}{k^2}$$

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$$

$|X-\mu| > k\sigma$

$X-\mu > k\sigma$ or $-(X-\mu) > k\sigma$

$X > \mu + k\sigma$ $\qquad$ $X-\mu < -k\sigma$

$X < \mu - k\sigma$

So, $P(X < \mu - k\sigma) + P(X > \mu + k\sigma) \leq \frac{1}{k^2}$

Using R and R-Studio

1. Download R from https://cran.r-project.org/
2. Download R-Studio from https://www.rstudio.com/