D = disease

1. $P(D) = .02$

   $P(\text{No } D) = .98$

$P(\text{Pos} \mid D) = .98$

$P(\text{Pos} \mid \text{No } D) = .02$

Bayes Theorem

we want $P(D \mid \text{Pos})$?

$$P(D \mid \text{Pos}) = \frac{P(\text{Pos} \mid D) \cdot P(D)}{P(\text{Pos})}$$, where $P(\text{Pos})$ is...

$$\begin{aligned} P(\text{Pos}) &= P(\text{Pos} \mid D) \cdot P(D) + P(\text{Pos} \mid \text{No } D) \cdot P(\text{No } D) \\ &= (.98)(.02) + (.05)(.98) \\ &= 0.0196 + 0.049 \\ &= 0.0686 \end{aligned}$$

$$P(D \mid \text{Pos}) = \frac{P(\text{Pos} \mid D) \cdot P(D)}{P(\text{Pos})} = \frac{(.98)(.02)}{.0686}$$

$$= .285714 \approx 28.6\%$$

There is 28.6% probability that they actually have disease if they test positive

② $n = 15$ students — using $\alpha = 0.05$
Sample mean is 76

Test claim that the average test score is 80 with a standard deviation of 8.

Hypothesis

a) Null $\mu = 80$ — Average test score of the pop is 80

b) alternative $\mu \neq 80$ — Average test score is not 80

Test statistics

$$Z = \frac{\text{sample mean} - \text{pop. mean}}{\underbrace{\text{pop std} / \sqrt{\text{sample size}}}_{\text{standard error on the mean}}} = \frac{76 - 80}{8/\sqrt{15}} \approx -1.936$$

Critical value

This is two-tailed b/c alt. hypothesis is checking all differences

$\alpha = 0.05/2 = 0.025$

The critical values are therefore ± 1.96

Decision

Because our Z-score (−1.93) is between (−1.96, 1.96), we fail to reject our null hypothesis!

We don't have evidence to say average is different from 80.

③ Traffic control office records # of cars passing through intersection

$\lambda = 5$ cars pass through per minute

This is a Poisson Distribution $\quad P(x=k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$

a) Prob. of 7 cars pass thru in a given minute

$$P(X=7) = \frac{5^7 e^{-5}}{7!} = \frac{78125 \times 0.0067}{5040} = \frac{526.402}{5040} = .10444486$$

The prob. that exactly 7 cars pass through in a given minute is $\boxed{10.4\%}$.

b) Prob. of at most 2 cars pass through

$$P(X=0) + P(X=1) + P(X=2) = P(X \le 2)$$

$$P(X=0) = \frac{5^0 e^{-5}}{0!} = .00674$$

$$P(X=1) = \frac{5^1 e^{-5}}{1!} = \frac{5 \times .00674}{1} = .0336897$$

$$P(X=2) = \frac{5^2 \times e^{-5}}{2!} = \frac{25 \cdot e^{-5}}{2} = .08422433$$

$$P(X \le 2) = .00674 + .0336897 + .08422433 = 0.1246$$

The prob. that @ most 2 cars pass through in a given minute is $\boxed{12.5\%}$.

4 "correlation implies causation" is incorrect...

1 correlations between an X and Y can be **spurious**, meaning they occurred by **chance**, and doesn't mean X caused Y.

2 A correlation may occur because of a third factor, a confounding variable, that is related to both variables (X, Y).

An example where correlation does **NOT** imply causation comes from lecture. There is a positive correlation between ice cream sales & drowning **incidents** in the summer months. In this example, ice cream sales are not causing drownings or drownings are not causing ice cream purchases. A third variable, the temperature, may actually be causing rises & falls in each variable. As the temperatures increase, more people buy ice cream & more people want to go swimming, which creates more opportunities for drowning.

Another example that represents a spurious correlation, comes from the Tyler Vigen website, and is the distance between Neptune & Earth with the Burglaries in Kansas. As the burgulary rate has decreased from 1985-2020, the planetary distance has also decreased.

5 Find adjusted R^2

a) Model A has $R^2 = 0.75$ with 5 predictors

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1-R^2)(n-1)}{n-k-1}\right) \qquad \begin{array}{l} k=5 \\ n=100 \end{array}$$

$$= 1 - \left(\frac{(1-0.75)(100-1)}{100-5-1}\right)$$

$$= 1 - \left(\frac{0.25 \times 99}{94}\right)$$

$$= 1 - .2633 = .7367$$

The adjusted R^2 of model A with $R^2 = 0.75$ and 5 predictors is $\boxed{.7367}$

b) Model B has $R^2 = 0.80$ with 10 predictors

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1-R^2)(n-1)}{n-k-1}\right) \qquad \begin{array}{l} k=10 \\ n=100 \end{array}$$

$$= 1 - \left(\frac{(1-.80)(100-1)}{100-10-1}\right)$$

$$= 1 - \left(\frac{0.20 \times 99}{89}\right)$$

$$= 1 - .22247 = .777528$$

The adjusted R^2 of Model B with $R^2 = 0.80$ and 10 predictors is $\boxed{.7775}$

⑥ Sometimes a model may have a high R^2 because it is too complex with **too many** predictors and this leads to **overfitting**. This means that the model is good at predicting training data, but then proves poorly with new data. In this case, a high R^2 is not a good indicator of model performance.

There are also issues of high R^2 in models where you add **spurious correlations**. These highly correlated variables may increase the R^2, but may not actually predict the dependent variable accurately.

7. Logistic Regression

$$\underbrace{\log\left(\frac{p}{1-p}\right)}_{\text{log odds}} = B_0 + B_1 x \text{, where } B_1 = 0.5$$

→ For every 1-unit increase in x, the **log-odds** of the event ($y=1$) increases by **0.5.**

→ If we convert the **log-odds to odds...**

$$\text{odds} = e^{0.5} = 1.65$$

For every 1-unit increase in x, the odds of the event ($y=1$) are **increased by 65%**.

$1.65 - 1 = .65$ or 65%