DS 122 Midterm 1

Name:

BUID:

Problem A (10 points)

Part 1

(5 points) A box contains three coins: two regular coins and one fake two-headed coin ($P(H) = 1$). You pick a coin at random and toss it. What is the probability that it lands heads up?

Solution

This is another typical problem for which the law of total probability is useful. Let C_1 be the event that you choose a regular coin, and let C_2 be the event that you choose the two-headed coin. Note that C_1 and C_2 form a partition of the sample space. We know that $P(H|C_1) = 0.5$ and $P(H|C_2) = 1$.

Thus, we can use the law of total probability to write

$$P(H) = P(H|C_1) \cdot P(C_1) + P(H|C_2) \cdot P(C_2)$$

$$= 1/2 \cdot 2/3 + 1 \cdot 1/3$$

$$= 2/3$$

Part 2

(5 points) You pick a coin at random and toss it, and get heads. What is the probability that it is the two-headed coin?

Solution

For this problem, we are interested in $P(C_2|H)$. We use Bayes' rule

$$P(C_2|H) = \frac{P(H|C_2)P(C_2)}{P(H)}$$

$$= \frac{1 \cdot 1/3}{2/3} = 1/2$$

Problem B (20 points)

Let X be a discrete random variable having the following distribution:

$$p_X(x) = \begin{cases} 1/3 \text{ when } x = 1, 2, 3 \\ 0 \text{ otherwise} \end{cases}$$

Part 1

(5 points) Find $E[X]$.

Solution

$$E[X] = \sum_{-\infty}^{\infty} x \cdot p(x) = \sum_{1}^{3} x \cdot p(x)dx = 1/3(1 + 2 + 3) = 2$$

Part 2

(5 points) Find $Var(X)$.

Solution

Answer: First find $E[X^2]$:

$$E[X^2] = \sum_{-\infty}^{\infty} x^2 \cdot p(x) = \sum_{1}^{3} x^2 \cdot p(x) = 1/3(1^2 + 2^2 + 3^2) = 14/3$$

Then:

$$Var[X] = E[X^2] - (E[X])^2 = 14/3 - 2^2 = 2/3$$

Part 3

(10 points) Define the random variable Y to be $3X^2 - X$. Find the expected value of Y.

Can either use linearity of expectation w/ solutions from part 1 and part 2:

$$E[Y] = E[3X^2 - X] = 3E[X^2] - E[X] = 3 * (14/3) - 2 = 12$$

Problem C (10 points)

Show that for two independant random variables X and Y

$$Var(X - Y) = Var(X) + Var(Y)$$

Hint: Start with the fact that $Var(X) = E[X^2] - E[X]^2$

Solution

$$Var(X - Y) = E[(X - Y)^2] - E[X - Y]^2$$

$$= E[X^2 + Y^2 - 2XY] - E[X - Y]E[X - Y]$$

$$= E[X^2] + E[Y^2] - 2E[XY] - (E[X] - E[Y])(E[X] - E[Y])$$

$$= E[X^2] + E[Y^2] - 2E[XY] - (E[X]^2 + E[Y]^2 - 2E[X]E[Y])$$

$$= E[X^2] + E[Y^2] - 2E[XY] - E[X]^2 - E[Y]^2 + 2E[X]E[Y])$$

$$= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 - 2(E[XY] - E[X]E[Y])$$

$$= Var(X) + Var(Y) - 2Cov(X, Y)$$

X and Y are independant random variables, so $Cov(X, Y) = 0$. Thus:

$$Var(X - Y) = Var(X) + Var(Y)$$

Problem D (10 points)

You challenge a friend to a free throw shooting contest. For each round of this contest you each take a shot. If you both make it or neither of you make it, no one wins. If only only you score, your friend gives you a dollar, and if only they score, you give them a dollar. You know that you're a better shooter than your friend, you make 80% of your shots while they only make 70% of their shots. Assume all shots are i.i.d.

Part 1

(5 points) How many rounds do you need to play before you expect to make 100 dollars?

Solution

First calculate the probability of each outcome:

Both miss: $0.2 * 0.3 = 0.06$

You make it, they miss: $0.8 * 0.3 = 0.24$

You miss, they make it: $0.2 * 0.7 = 0.14$

You both make it: $0.8 * 0.7 = 0.56$

Now you can descrive the outcome as a random variable **X** (note that both make it and both miss have values of 0):

0	-1	1
0.62	0.14	0.24

Now you can compute the expected value:

$$E[X] = (0)(0.62) + (-1)(0.14) + (1)(0.24) = 0.1$$

In order to expect to make 100 dollars, you need to set $n * E[X] = 100$. Solving for n tells us it would take **1000 rounds.**

Part 2

(5 points) After 100 rounds, there is a 68% chance that your earnigns will be in what range?

Solution

First you need to compute the Variance **X**. This could either be calculated directly:

$$Var(X) = (0.62)(0 - 0.1)^2 + (0.14)(-1 - 0.1)^2 + (0.24)(1 - 0.1)^2 = 0.37$$

Or by the variance formula:

$$Var(X) = E[X^2] - E[X]^2 = ((0)^2(0.62) + (-1)^2(0.14) + (1)^2(0.24)) - (.1)^2 = 0.37$$

The 68% confidence interval for 100 rounds is one standard deviation around the mean. We can use linearity of expectation, and because shots are i.i.d, we can also sum the variances:

$$n * E[x] \pm \sqrt{n * Var(x)}$$

$$100 * 0.1 \pm \sqrt{100 * 0.37}$$

$$10 \pm 6.1$$

or in other notation

$$[3.9, 16.1]$$

Problem E (10 points)

You sample heights of college freshman and college sophomores at random. You manage to collect 45 freshman heights and 61 sophomore heights. You immediately take your measuremnts and calculate their means and standard deviations. You find that the mean of the first sample is 66 inches with a standard deviation of 6 and the mean of the second sample is 70 inches with a standard deviation of 5. Note: assume the distribution of heights in the overall population is normal.

Part 1

(4 points) What are the 95% confidence intervals for the average heights of the two populations?

Solution

The 95% confidence interval is caclulated from the sample mean and standard deviation as:

$$\mu \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

So for freshmen we get a 95% CI of:

$$66 \pm 1.96 * \frac{6}{\sqrt{45}} = 66 \pm 1.75 = [64.25, 67.65]$$

And for sophomores:

$$70 \pm 1.96 * \frac{5}{\sqrt{61}} = 70 \pm 1.25 = [68.75, 71.25]$$

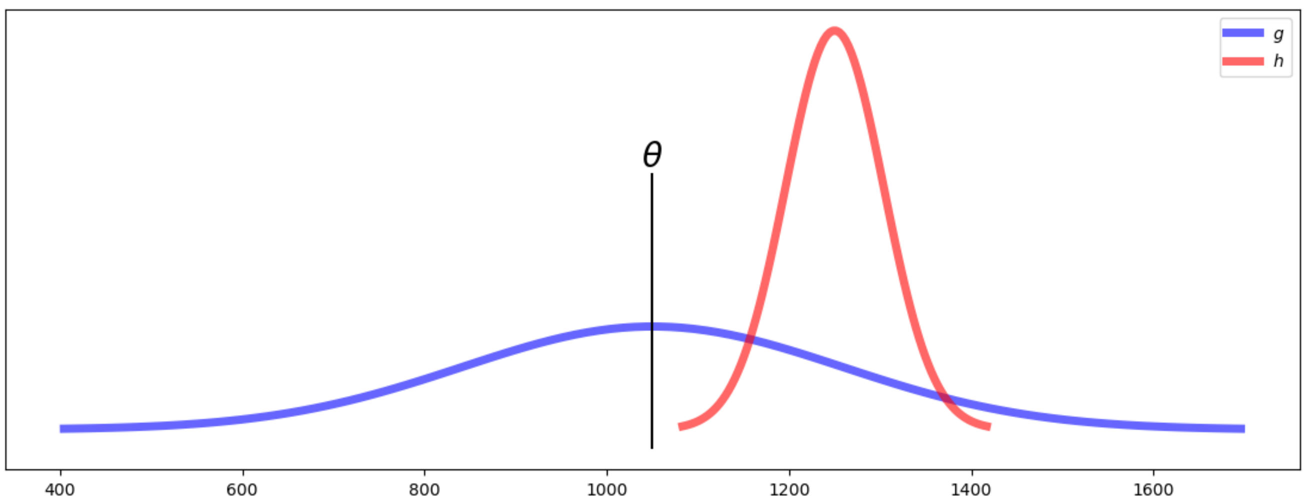
(1 point) In words, what you can conclude about the mean heights of the populations relative to each other based on your confidence intervals?

__Solution__

The confidence interval do not overlap. This can be interpreted to mean that because the true means of the populations do not lie in each others 95% confidence intervals, we can say with 95% confidence that the difference between the population means is not 0. Could also observe this as a one tailed test, and claim with 97.5% confidence that that the mean heigh of sophomores is greater than the mean height of freshman.

Problem F (4 points)

Consider two estimators, $g()$ and $h()$, that both use data sampled from a parametric distribution to estimate that distribution's parameter θ. The plot below shows the distribution of values one obtains for both $g()$ and $h()$ over different data samples, where all samples are the same size (eg, each sample contains the same number n of data items).



Part 1

(2 points) Contrast the two estimators: which has higher bias, and which has higher variance?

Solution:

g has higher variance, and h has higher bias.

Part 2

(2 points) State which estimator(s) if any appear to be unbiased and why.

Solution:

g appears to be unbiased because its mean looks to be close to θ.

Problem G (10 points)

Assuming that we have a standard Binomial distribution of n independent Bernoulli trials each with a probability of success p, and let x_i be the outcome of the i-th trial, where $x_i = 1$ if the trial is successful and $x_i = 0$ otherwise, find the maximum likelihood estimator for the probability of success for this Binomial distribution.

Hint: start with probability of a single Bernoulli trial: $P(x, p) = p^x(1 - p)^{1-x}$ and consider it's likelihood over n

Solution:

Assuming that we have observed n independent Bernoulli trials with a probability of success p, and let x_i be the outcome of the i-th trial, where $x_i = 1$ if the trial is successful and $x_i = 0$ otherwise, then the likelihood function is given by:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1 - p)^{1-x_i}$$

Taking the logarithm of the likelihood function, we have:

$$\log L(p) = \sum_{i=1}^{n}[x_i \log(p) + (1 - x_i) \log(1 - p)]$$

To find the MLE for p, we differentiate the log-likelihood with respect to p and set it equal to zero:

$$\frac{\partial}{\partial p}\log L(p) = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\sum_{i=1}^{n}(1-x_i)}{1-p} = 0$$

Simplifying the above equation, we get:

$$\hat{p}\mathrm{MLE} = \frac{\sum i=1^n x_i}{n}$$

Therefore, the MLE for the probability of success in a binomial distribution is the sample mean of the observed outcomes.