INTRO. TO MACHINE LEARNING HW1

SNIR HORDAN 205689581, ALEXANDER SHENDER 328626114

1.
By Law of Total Probability,
$P\,(\,7\,out\,of\,10\,heads)$=P($7\,out\,of\,10\,heads|forged)$P(forged)+$P\,(\,7\,out\,of\,10\,heads\,|fair)$P(fair)
By Bayes Theorem,

$$P\left(forged|7\,out\,of\,10\,heads\right)=\frac{P\,(\,forged\ \cap\ 7\,out\,of\,10\,heads)}{P\,(\,7\,out\,of\,10\,heads)}$$
$$=\frac{\frac{1}{1000}\left(\begin{array}{c}10\\7\end{array}\right)(0.8)^7\,(0.2)^3}{\frac{1}{1000}\left(\begin{array}{c}10\\7\end{array}\right)(0.8)^7\,(0.2)^3+\frac{999}{1000}\left(\begin{array}{c}10\\7\end{array}\right)(0.5)^7\,(0.5)^3}$$
$$\approx 0.00171675$$

2.Denote X-number of boys, Y-number of girls
Each time a famliy has a new child is equivelant to performing a bernoulli experiment (such as flipping a coin) with Pr(boy)=Pr(girl)=0.5 because each birth is independent from another birth.

P$(X=1)=P\left(boy\,is\,born\right)=1$ because families keep giving birth until boy is born and then stop.
Therefore in all cases a family will have precisely one boy.
Then by definition of expected value:
$EX=1P\left(X=1\right)=1$

$P\left(Y=0\right)=$ P(boy)=P(boy)=0.5because if boy is born then stop giving birth
P$(Y=n)=P\left(n\,girls\,then\,boy\right)$=P$(girl)^n\,P\left(boy\right)=0.5^{n+1}$
Then by the definition of expected value:
EY=$\sum_{n=0}^{\infty}P\left(Y=n\right)n=\sum_{n=1}^{\infty}n\left(0.5\right)^{n+1}=1$

Expected value is the average value of a random variable over a large amount of experiments.
The population of a country is very large and is comprised of all the people born there according to the above underlying probability distribution.
Then EX=EY implies that there will be the same number of boys and girls in the population.

3.a. $X_i \sim N\left(\mu_i, \sigma_i^2\right)$ iff probability density function of X_i is $f_{X_i}\left(x_i\right) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i}}$

$$f_{X_1}\left(x_1\right) = \int_{-\infty}^{\infty} f_{X_1,X_2}\left(x_1,x_2\right) dx_2 =$$

$$= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\} dx_2$$

Let $y_2 = \frac{x_2-\mu_2}{\sigma_2}$. $dy_2 = dx_2 \frac{1}{\sigma_2}$

$...{=}\int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right) y_2 + y_2^2 + \rho^2\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - \rho^2\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right]\right\} \sigma_2 dy_2{=}...$

$...{=}\left(\frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right\}\right)\left(\int_{-\infty}^{\infty} \exp\left\{-\frac{\left(y_2+\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\right)^2}{2(1-\rho^2)}\right\} dy_2\right) \underset{(1)}{=} \sqrt{1-\rho^2}\sqrt{2\pi}\frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right\} = ...$

${=}\frac{1}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right\}$

$(1)\int_{-\infty}^{\infty} \exp\left\{-\alpha\left(x+u\right)^2\right\} dx = \sqrt{\frac{\pi}{\alpha}}$

Analogously for $f_{X_2}\left(x_2\right)$, switching x_1, x_2 and letting $y_2 = \frac{x_1-\mu_1}{\sigma_1}$ gives us same result for X_2 from symmetry of the joint probability function.

3.b.By Bayes Law, (variation of Bayes Law applied to density functions)

$$f_{X_{1|}|X_2=x_2}(x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)} =$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\} \ldots.$$

$$\ldots/\frac{1}{\sigma_2\sqrt{2\pi}}\exp\left\{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right\} = \frac{\sqrt{2\pi}\sigma_2}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - (1-\rho^2)\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\} = \ldots$$

$$\ldots = \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(p\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right)^2\right]\right\} = \ldots$$

$$= \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right) - p\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]^2\right\} = \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}\exp\left\{-\frac{\left(x_1-\left(\mu_1+\frac{\rho\sigma_1(x_2-\mu_2)}{\sigma_2}\right)\right)^2}{2\sigma_1^2(1-\rho^2)}\right\}$$

Therefore, $X_1|X_2 = x_2 \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2), \sigma_1^2(1-\rho^2)\right)$

This shows that if X_1, X_2 are Gaussian and $Cov(X_1, X_2) = 0$ then $X_1, X_2$are independent.

4. If Y=0 then $0 = \text{Cov(X,0)} \leq 0\sigma_X^2 = 0$and $\rho = 0$

If X=0 then $0 = \text{Cov(Y,0)} \leq 0\sigma_Y^2 = 0$and $\rho = 0$

Otherwise,

Let $Z = X - \frac{Cov(X,Y)}{\sigma_Y^2}Y$

$0 \leq_{(1)} \sigma_Z^2 \leq Cov\left(X - \frac{Cov(X,Y)}{\sigma_Y^2}Y, X - \frac{Cov(X,Y)}{\sigma_Y^2}Y\right) =_{(2)} \sigma_X^2 - 2Cov(\frac{Cov(X,Y)}{\sigma_Y^2}Y, X) + \frac{Cov(X,Y)}{\sigma_Y^4}\sigma_Y^2$

$= \sigma_X^2 - 2\frac{Cov(X,Y)^2}{\sigma_Y^2} + \frac{Cov(X,Y)^2}{\sigma_Y^2}$

(1)variance is non-negative

(2)lineaerity and homogeneity of covariance

Then $Cov(X,Y)^2 \leq \sigma_X^2\sigma_Y^2 \implies |Cov(X,Y)| \leq \sigma_X\sigma_Y \implies -1 \leq \frac{Cov(X,Y)}{\sigma_X\sigma_Y} \leq 1$

By definition, $-1 \leq \rho \leq 1$.

(Upper and lower bounds are recieved when Y=cX for c≠0)

QED

5. a. $\sigma_i\tilde{}Bin(p,10)$ from definition of binomial distribution, which is a sum of independent bernoulli experiments with probability p.

5.b. Eσ_i=np=10p .

Let $\{X_j\}_{j=1}^{10}$ be independent bernoulli experiments (independent coin flips are bernolli experiments).

Then by definition $\sigma_i = X_1 + X_2 + ... + X_{10}$.

Then $E\sigma_i = EX_1 + X_2 + ... + X_{10}$=$EX_1 + ... + EX_{10} = p + ... + p = 10p$. (because $\{X_j\}$ i.i.d.)

5.c.

$\{\sigma_i\}_{i=1}^{n=1000}$ are independent random variables because the coin flips are independent, and distributed with binomial distribution $Bin(p,10)$ as we showed in 5.a. Eσ_i=10p as seen in 5.b. $0 \leq \sigma_i \leq 10$ because σ_i is increased by 0 when coin i flips to "tails" and 1 when coin i is flipped to "heads", then maximum amount of head flips for each σ_i is 10 and minimum is 0. Therefore P$(0 \leq \sigma_i \leq 10) = 1$, (a=0,b=10)

By Hoeffding's Inequality, for any $\epsilon > 0$, P$\left(\left|\hat{\theta} - \mu\right| \geq \epsilon\right) \leq 2\exp\left\{-\frac{2n\epsilon^2}{(b-a)^2}\right\}$, i.e.

$$2\exp\left\{-\frac{2\,(1000)\,\epsilon^2}{(10-0)^2}\right\} = 2\exp\left\{-20\epsilon^2\right\} \leq 0.05$$

$$\implies 20\epsilon^2 \geq -\ln(0.025) \implies \epsilon \geq \sqrt{\frac{-\ln(0.025)}{20}}$$

Then $\epsilon_{min} = \sqrt{\frac{-\ln(0.025)}{20}} \approx 0.4295$ is the smallest error margin posible with confidence 0.95.

6.**Proof**

Let $\{A_i\}_{i=1}^{n}$ be events in a probability space and P a probability function.

Notice that $\bigcup\limits_{i=1}^{n} A_i = \bigcup\limits_{i=1}^{n}\left[A_i \backslash \bigcup\limits_{j=1}^{i-1} A_j\right]$ from basic set theory and $\left\{A_i \backslash \bigcup\limits_{j=1}^{i-1} A_j\right\}_{i=1}^{n}$ are pairwise-disjoint events.

Claim If $\{B_i\}_{i=1}^{n}$ pairwise-disjoint events then $P\left(\bigcup\limits_{i=1}^{n} B_i\right) = \sum\limits_{i=1}^{n} P(B_i)$

Proof Base case: n=1 trivial P$(B_1) = P(B_1)$

n=2 :

By Inclusion-Exclusion Formula $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for events A, B.

Then if B_1, B_2 are disjoint $P(B_1 \cup B_2) = P(B_1) + P(B_2)$.

Assume induction hypothesis for n and induce for n+1.

$\bigcup\limits_{i=1}^{n-1} B_i$ and B_n are disjoint because $B_n \cap B_i = \emptyset$ for all $1 \leq i \leq n-1$, then

$P\left(\bigcup\limits_{i=1}^{n} B_i\right) = P\left(\bigcup\limits_{i=1}^{n-1} B_i \cup B_n\right) = P\left(\bigcup\limits_{i=1}^{n-1} B_i\right) + P(B_n) \underset{induction\,hypothesis}{=} \sum\limits_{i=1}^{n-1} P(B_i) + P(B_n) = \sum\limits_{i=1}^{n} P(B_i)$

QED

By claim, $P\left(\bigcup\limits_{i=1}^{n} A_i\right) = P\left(\bigcup\limits_{i=1}^{n}\left[A_i \backslash \bigcup\limits_{j=1}^{i-1} A_j\right]\right) = \sum\limits_{i=1}^{n} P\left(\left[A_i \backslash \bigcup\limits_{j=1}^{i-1} A_j\right]\right) \leq \sum\limits_{i=1}^{n} P(A_i)$

because if $A \subset B \implies P(A) \leq P(B)$ because P is a probability function

then for every $1 \leq i \leq n$, P$\left(A_i \backslash \bigcup\limits_{j=1}^{i-1} A_j\right) \leq P(A_i) \implies \sum\limits_{i=1}^{n} P\left(\left[A_i \backslash \bigcup\limits_{j=1}^{i-1} A_j\right]\right) \leq \sum\limits_{i=1}^{n} P(A_i)$.

QED