Max Tjen
CSDS 313
Homework 1

1) (5pts) Data is usually represented by a matrix, where rows are samples and columns are variables. However, it is not always the case. Give two different examples that data are not represented by a matrix.
    a) Genetic data
        i) There can be very few patients (rows) that have been documented but for each patient, there are very many biomarkers
    b) State data
        i) For each state, there can be hundreds of variables with each signifying either a continuous or discrete variable (ex. Median salary of residents, whether or not there was a natural disaster in the past year)

2) (15pts) Prove the following equations and use English to explain the meaning of each equation.
    a) $E(X-\mu)^2 = EX^2 - \mu^2$
        i) $E(X-\mu)^2$
        ii) $E((X-\mu)(X-\mu))$
        iii) $E(X^2 - 2X\mu + \mu^2)$
        iv) $EX^2 - 2EX\mu + \mu^2$
        v) $EX^2 - 2\mu\mu + \mu^2$
        vi) $EX^2 - 2\mu^2 + \mu^2$
        vii) $EX^2 - \mu^2$
        viii) This equation signifies a distribution's variance, as it uses data values and the mean value. The variance describes the spread of data points in a set relative to its mean
    b) $E(\hat{\mu}) = \mu$
        i) $E(\hat{\mu})$
        ii) $E((\mu_1 + \mu_2 + \ldots + \mu_n)/n)$
        iii) $(1/n)(E(\mu_1 + \mu_2 + \ldots + \mu_n))$
        iv) $(1/n)(n\mu)$
        v) μ
        vi) This equation signifies that the mean of sample means should be equal to the mean of the population. This is because of the central limit theorem,

where the distribution of the sample means should be approximately the same as the population mean.

c) $var(\hat{\mu}) = \sigma^2/n$

i) $var(\hat{\mu})$

ii) $var((\mu_1 + \mu_2 + \ldots + \mu_n)/n)$

iii) $(1/n)^2(var(\mu_1 + \mu_2 + \ldots + \mu_n))$

iv) $(1/n)^2(\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2)$

v) $(1/n)^2(n\sigma^2)$

vi) $(1/n^2)(n\sigma^2)$

vii) σ^2/n

viii) This equation signifies that the variance of sample means should be equal to the variance of the population. This is because of the central limit theorem, where the distribution of the sample standard deviations should be approximately the same as the population standard deviation. Variation is built off of standard deviation (s.d.^2), so the sample variations should be approximately the same as the population variation

3) (30) For each of the six distributions: Bernoulli with parameter p, Binomial(n, p), Uniform U(a,b), Normal N(μ,σ2), Exponential with parameter λ, and power law with parameters α and xmin, please provide their 1) probability mass function or probability density function; 2) cumulative distribution function (CDF); 3) mean; and 4) variance.

a) Bernoulli

i) Probability mass function: $f(x) = \{ p \; if \; x = 1$

$\{ 1 - p \; if \; x = 0$

ii) Cumulative distribution function: $F(x) = \{ 0 \; if \; x = 0$

$\{ 1 - p \; if \; 0 \leq x < 1$

$\{ 1 \; if \; x \geq 1$

iii) Mean: $E[X] = p$

iv) Variance: $Var(X) = p(1 - p)$

b) Binomial

i) Probability mass function: $f(x) = \binom{n}{x} p^x (1 - p)^{n-x}$

ii) Cumulative distribution function: $F(x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$

iii) Mean: $E[X] = np$

iv) Variance: $Var(X) = np(1-p)$

c) Uniform

i) Probability density function: $f(x) = \{ 1/(b-a) \; if \; a \leq x \leq b$

$\{ 0 \; if \; otherwise$

ii) Cumulative distribution function: $F(x) = \{ 0 \; if \; x < a$

$\{ (x-a)/(b-a) \; if \; a \leq x \leq b$

$\{ 0 \; if \; x > b$

iii) Mean: $E[X] = (a+b)/2$

iv) Variance: $Var(X) = (b-a)^2/12$

d) Normal

i) Probability density function: $f(x) = (1/\sigma\sqrt{2\pi}) e^{-(x-\mu)^2/2\sigma^2}$

ii) Cumulative distribution function: $F(x) = \int_{-\infty}^{x} (1/\sqrt{(2\pi)}) e^{-x^2/2}$

iii) Mean: $E[X] = \mu$

iv) Variance: $Var(X) = \sigma^2$

e) Exponential

i) Probability density function: $f(x) = \{ \lambda e^{-\lambda x} \; if \; x \geq 0$

$\{ 0 \; if \; x < 0$

ii) Cumulative distribution function: $F(x) = \{ 0 \; if \; x \leq 0$

$\{ 1 - e^{-\lambda x} \; if \; x > 0$

iii) Mean: $E[X] = 1/\lambda$

iv) Variance: $Var(X) = 1/\lambda^2$

f) Power law

i) Probability density function: $f(x) = ([\alpha - 1]/x_{min})(x/x_{min})^{-\alpha}$

ii) Cumulative distribution function: $F(x) = (x/x_{min})^{-\alpha+1}$

iii) Mean: $E[X] = ([\alpha - 1]/[\alpha - 2])x_{min}$

iv) Variance: $Var(X) = ([\alpha - 1]/[\alpha - 3])x_{min}^{2}$