Probability & Statistics

1. Descriptive Statistics

Given a **SAMPLE** data of size n.

1.1 Quantitative (discrete and continuous) variable/data:

a. sample mean: $\bar{x} = \frac{\sum x_i}{n}$ or $\bar{x} = \frac{\sum x_i f_i}{n}$

b. sample variance: $s^2 = \frac{\sum (x_i - \bar{x})^2}{n}$ or $s^2 = \frac{1}{n}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]$

$$s^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{n} \quad \text{or} \quad s^2 = \frac{1}{n}\left[\sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n}\right]$$

c. sample standard deviation: $s = \sqrt{s^2}$

1.2 Qualitative (categorical) variable/data:

a. sample proportion: $\hat{p} = \frac{f}{n}$

2. Probability Theory

Given a sample space S, and events A, B.

2.0 Interpretations of probability:

a. definition(classical): $P(A) = \frac{|A|}{|S|}$

b. definition(frequentist): $P(A) = \lim_{n\to\infty} \frac{f}{n}$

2.1 Counting:

a. mn rule: $n_1 n_2 ... n_k$

b. permutation: $P_r^n = \frac{n!}{(n-r)!}$

c. combination: $C_r^n = \frac{n!}{(n-r)!r!}$

2.2 Probability of **ANY** events A and B:

a. rules: $P(\phi) = 0$
$P(S) = 1$
$0 \le P(A) \le 1$

b. complement: $P(A^c) = 1 - P(A)$

c. intersection: $P(A \cap B)$

d. union: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

e. exhaustive: $P(A \cup B) = P(S) = 1$

2.3 Conditional probability of **ANY** events A and B:

a. definition: $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$

b. intersection: $P(A \cap B) = P(A \mid B)P(B)$

2.4 **MUTUALLY EXCLUSIVE** events A and B:

a. definition: $P(A \cap B) = \phi$

b. union: $P(A \cup B) = P(A) + P(B)$

2.5 **INDEPENDENT** events A and B:

a. definition: $P(A \mid B) = P(A)$

b. intersection: $P(A \cap B) = P(A)P(B)$

c. union: $P(A \cup B) = P(A) + P(B) - P(A)P(B)$

2.6 **MUTUALLY EXCLUSIVE** and **EXHAUSTIVE** events B and B^c:

a. total probability: $P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c)$

b. Bayes' theorem: $P(B \mid A) = \dfrac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)}$

2.7 **MUTUALLY EXCLUSIVE** and **EXHAUSTIVE** events B_1, B_2, ..., B_k:

a. total probability:
$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + ... + P(A \mid B_k)P(B_k)$

b. Bayes' theorem:
$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + ... + P(A \mid B_k)P(B_k)}$$

3. Probability Distribution

3.1 Discrete random variable X:

a. definition: $P(X = x) \geq 0$ and $\sum P(X = x) = 1$

b. mean (expectation): $\mu = E(X) = \sum xP(X = x)$

c. variance: $\sigma^2 = Var(X) = \sum (x - \mu)^2 P(X = x)$

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

where $E(X^2) = \sum x^2 P(X = x)$

d. standard deviation: $\sigma = \sqrt{\sigma^2}$

e. probability: $P(X = x)$

$$P(a \leq X \leq b) = \sum_{a}^{b} P(X = x)$$

3.2 Continuous random variable X:

a. definition: $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$

b. mean: $\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$

c. variance: $\sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$

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

where $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx$

d. standard deviation: $\sigma = \sqrt{\sigma^2}$

e. probability: $P(X = x) = 0$

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$

3.3 Combining **ANY** random variables X, Y and constants a, b:

a. $E(aX + b) = aE(X) + b$

b. $Var(aX + b) = a^2 Var(X)$

c. $E(aX + bY) = aE(X) + bE(Y)$

d. $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$ if X and Y are **INDEPENDENT**

3.4 (Discrete) Binomial random variable X:

a. notation: $X \sim B(n, p)$

b. formula: $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$

c. mean: $\mu = np$

d. variance: $\sigma^2 = np(1 - p)$

3.5 (Continuous) Normal random variable X:

a. notation: $X \sim N(\mu, \sigma^2)$

b. formula: $f(x) = ...$ (too complex, unable to compute directly)

c. mean: μ

d. variance: σ^2

e. transform to N(0,1): $Z = \frac{X - \mu}{\sigma}$

3.6 (Continuous) Standard normal random variable Z:

a. notation: $Z \sim N(0,1)$

b. formula: $f(x) = ...$ (too complex also, use Z-table)

c. table: Z-table

d. mean: 0

e. variance: 1

3.7 Approximate Binomial using Normal:

a. condition: $np \geq 5$ **AND** $nq \geq 5$

b. binomial: $X \sim B(n, p)$

c. normal: $Y \sim N(np, npq)$, where $q = 1 - p$

d. approximation: $P(X = x) \approx P(x - 0.5 \leq Y \leq x + 0.5)$

4. Sampling Distribution

4.1 Sample mean (σ known):

a. n≥30, any population: $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

b. n<30, population $\sim N(\mu, \sigma^2)$: $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

4.2 Sample mean (σ unknown):

a. n≥30, population $\sim N(\mu, s^2)$: $\overline{X} \sim N\left(\mu, \frac{s^2}{n}\right)$

b. n<30, population $\sim N(\mu, s^2)$: $\overline{X} \sim t\left(\mu, \frac{s^2}{n}\right)$, $\nu = n - 1$

4.3 Sample proportion:

a. $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$, population $\sim B(n, p)$:

$$\hat{P} \sim N\left(p, \frac{pq}{n}\right)$$

5. Inferential Statistics for Means

Given a **SAMPLE** data of size n.

5.1 Unbiased point estimates:

a. population mean: $\bar{x} = \frac{\sum x}{n}$

b. population variance: $s^2 = \frac{1}{n-1}\sum (x-\bar{x})^2$

or $s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]$

5.2 Interval estimates (σ unknown):

a. confidence level: 1-α

b. n≥30, population normal: $\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$

c. n<30, population normal: $\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$

5.3 Testing population mean μ (σ unknown):

$H_0 : \mu = \mu_0$

$H_A : \mu \neq \mu_0$ or $\mu > \mu_0$ or $\mu < \mu_0$

estimate σ^2 by: $s^2 = \frac{1}{n-1}\sum (x-\bar{x})^2$

a. n≥30, population normal:

test statistic: $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

b. n<30, population normal:

test statistic: $t_\nu = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

with degree of freedom: $\nu = n-1$

5.4 Testing difference between two population means μ_X and μ_Y (σ unknown):

$H_0 : \mu_X - \mu_Y = d_0$

$H_A : \mu_X - \mu_Y \neq d_0$ or $\mu_X - \mu_Y > d_0$ or $\mu_X - \mu_Y < d_0$

a. samples independent, $n_X \geq 30$ and $n_Y \geq 30$, population normal

estimate σ_X^2 by: $s_X^2 = \frac{1}{n_X - 1}\sum(x - \bar{x})^2$

estimate σ_Y^2 by: $s_Y^2 = \frac{1}{n_Y - 1}\sum(y - \bar{y})^2$

test statistic: $z = \frac{(\bar{x} - \bar{y}) - d_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$

b. samples independent, $n_X < 30$ or $n_Y < 30$, population normal

assume $\sigma_X^2 \approx \sigma_Y^2 = \sigma^2$

estimate σ^2 by: $s_p^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}$

where $s_X^2 = \frac{1}{n_X - 1}\sum(x - \bar{x})^2$ and $s_Y^2 = \frac{1}{n_Y - 1}\sum(y - \bar{y})^2$

test statistic: $t_\nu = \frac{(\bar{x} - \bar{y}) - d_0}{s_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$

with degree of freedom: $\nu = n_X + n_Y - 2$

c. samples dependent, $n_X = n_Y = n < 30$, population normal

let $d_i = x_i - y_i$

estimate σ^2 by: $s_d^2 = \frac{1}{n-1}\sum(d - \bar{d})^2$

or $s_d^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]$

test statistic: $t_\nu = \frac{\bar{d} - d_0}{s_d / \sqrt{n}}$

with degree of freedom: $\nu = n - 1$

6. Inferential Statistics for Proportions

6.1 Unbiased point estimate of binomial population proportion p :

a. population proportion: $\hat{p} = \frac{x}{n}$

6.2 Interval estimate of binomial population proportion p :

a. $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$

confidence level: 1-α

interval: $\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

6.3 Testing binomial population proportion p :

$H_0 : p = p_0$

$H_A : p \neq p_0$ or $p > p_0$ or $p < p_0$

a. $np_0 \geq 5$ and $nq_0 \geq 5$

test statistic: $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$

where $\hat{p} = \frac{x}{n}$, $q_0 = 1 - p_0$

6.4 Testing difference between two binomial population proportions p_1 and p_2 :

$H_0 : p_1 - p_2 = d_0$

$H_A : p_1 - p_2 \neq d_0$ or $p_1 - p_2 > d_0$ or $p_1 - p_2 < d_0$

a. $d_0 = 0$, $n_1\hat{p}_1 \geq 5$, $n_1\hat{q}_1 \geq 5$, $n_2\hat{p}_2 \geq 5$ and $n_2\hat{q}_2 \geq 5$

test statistic: $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$,

where $\hat{p}_1 = \frac{x_1}{n}$, $\hat{p}_2 = \frac{x_2}{n}$, $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$, $\hat{q} = 1 - \hat{p}$

b. $d_0 \neq 0$, $n_1\hat{p}_1 \geq 5$, $n_1\hat{q}_1 \geq 5$, $n_2\hat{p}_2 \geq 5$ and $n_2\hat{q}_2 \geq 5$

test statistic: $z = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$

where $\hat{p}_1 = \frac{x_1}{n_1}$, $\hat{q}_1 = 1 - \hat{p}_1$, $\hat{p}_2 = \frac{x_2}{n_2}$, $\hat{q}_2 = 1 - \hat{p}_2$

6. 5 Testing single multinomial population proportions $p_1, p_2, ..., p_k$

a. $e \geq 5$

$H_0 : p_1 = p_2 = ... = p_k = \frac{1}{k}$ (all proportions are the same)

or $H_0 : p_1 = c_1, p_2 = c_2, ..., p_k = c_k$ (proportions are not the same)

H_A : at least one proportion is not equal

test statistic: $\chi_v^2 = \sum \frac{(o_i - e_i)^2}{e_i}$

degree of freedom: $\nu = k - 1$

computation: $e_i = np_i$

6.6 Testing difference between *k* multinomial population proportions

a. $e \geq 5$

$H_0 : p_1 = p_2 = ... = p_k$ (no difference in proportions between them)

$H_A : p_1, p_2, ..., p_c$ are not all equal

test statistic: $\chi_v^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$

degree of freedom: $\nu = (r-1)(c-1)$

computation: $e_{ij} = \frac{r_i c_j}{n}$

where $r_i = \sum_{j=1}^{c} o_{ij}$, $c_j = \sum_{i=1}^{r} o_{ij}$

6.7 Testing independence between two multinomial variables A and B

a. $e \geq 5$

H_0 : Variables A and B are independent

H_A : Variables A and B are not independent

test statistic: $\chi_v^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$

degree of freedom: $\nu = (r-1)(c-1)$

computation: $e_{ij} = \frac{r_i c_j}{n}$

where $r_i = \sum_{j=1}^{c} o_{ij}$, $c_j = \sum_{i=1}^{r} o_{ij}$