Random Variables

<u>outcomes are numerical values</u>

A hypothesis:

Birthdays are uniformly distributed over

the first six months (181 days)

and

the second six months (184 days)

of the year.

We poll n "randomly selected" people, and calculate

$\hat{P}_{\text{first half}} = \#(\text{occurrences of Jan-June})/n$

$\hat{P}_{\text{second half}} = \#(\text{occurrences of July-Dec})/n$

Should we "<u>accept</u>" or "<u>reject</u>" our hypothesis?

Discrete Random Variables

Introduce sample space

$\{x_1, \ldots, x_J\}$ real numbers.

Then

$\underset{\text{r.v.}}{X} = x_j$ with probability p_j, $1 \le j \le J$

where $\begin{cases} 0 \le p_j \le 1, \ 1 \le j \le J \\ \sum_{j=1}^{J} p_j = 1 \end{cases}$.

Probability mass function (pmf):

$$\underbrace{f_X}_{\text{pmf}}(\underbrace{x_j}_{\text{outcome}}) = \underbrace{p_j}_{\text{probability of outcome}}, \quad 1 \le j \le J.$$

(r.v.: X)

Note $\begin{cases} 0 \le f_X(x_j) \le 1, \quad 1 \le j \le J \\ \text{and} \\ \sum_{j=1}^{J} f_X(x_j) = 1 \end{cases}$.

Example: uniform distribution J

Let $x_j = j$, $1 \le j \le J$:

$J = 6$: die roll face ?

$J = 12$: birthmonth ;

define

$$f_X^{\text{unif},J}(x_j) = \underbrace{1/J}_{p_j}, \quad 1 \le j \le J.$$

(Note $0 \le p_j \le 1$, $1 \le j \le J$, and $\sum_{j=1}^{J} p_j = 1$.)

Example: Bernoulli parameter θ, $0 \le \theta \le 1$

Let $J = 2$, and

$x_1 = 0$ ("tail"), $x_2 = 1$ ("head").

then "fair" coin: $\theta = 1/2$

$$f_X^{\text{Bernoulli}}(x;\theta) = \begin{cases} 1-\theta & \text{if } x = x_1 = 0 \quad p_1 \\ \theta & \text{if } x = x_2 = 1 \quad p_2 \end{cases}.$$

Note $0 \le p_1, p_2 \le 0$ and $p_1 + p_2 = 1$ for any admissible value of θ.

Random Variate Generation (Simulation)

X: a random variable

a sample space and probability law

x : a random variate — a realization of X

a number

Physical generation:

flip a coin, roll a die, ...

OR

Pseudo-random variate generation:

in MATLAB,

randi(J)

draws a member from the uniform pmf $f_X^{\text{unif},J}$ "population"

- a virtual roll of the die, OR
- a virtual flip of a (fair) coin, OR ...

DEMO

Expectation

Given a r.v. X with pmf $f_X(x)$, and a univariate function g,

$$\underbrace{\mathbb{E}}_{\substack{\text{expectation} \\ \text{(not random)}}} \text{ of } (\underbrace{g(X)}_{\substack{\text{random} \\ \text{quantity}}}) \equiv \sum_{j=1}^{J} \underbrace{g(x_j)}_{\text{outcome}} \cdot \underbrace{p_j}_{\text{probability } f_X(x_j)}$$

Note

$$\mathbb{E}(\underbrace{g(X) = C}_{\text{constant}}) = \sum_{j=1}^{J} C\, p_j = C \sum_{j=1}^{J} p_j = C.$$

μ, σ^2, and σ

mean, μ: center of mass

$$\mu = \mathbb{E}(X) = \sum_{j=1}^{J} x_j p_j$$

note
$$\begin{aligned}\mathbb{E}(X-\mu) &= \sum_{j=1}^{J} (x_j - \mu)\, p_j \\ &= \sum_{j=1}^{J} x_j p_j - \sum_{j=1}^{J} \mu p_j \\ &= \mathbb{E}(X) - \mu \sum_{j=1}^{J} p_j \\ &= \mu - \mu = 0\end{aligned}$$

variance, σ^2: "spread^2"

$$\begin{aligned}\sigma^2 &\equiv \mathbb{E}\left((X-\mu)^2\right) \\ &= \sum_{j=1}^{J} (x_j - \mu)^2 p_j \qquad (= E(X^2) - \mu^2)\end{aligned}$$

standard deviation, σ spread
(std dev)

$$\sigma \equiv \sqrt{\sigma^2} \qquad \text{definition}$$

Example: uniform distribution J

$$x_j = j,\ 1 \le j \le J \qquad p_j = 1/J,\ 1 \le j \le J$$

$$\mu = \mathbb{E}(X) = \sum_{j=1}^{J} x_j p_j = \frac{1}{J} \sum_{j=1}^{J} j = \frac{1}{J}\left(\frac{J(J+1)}{2}\right) = \frac{J+1}{2}$$

$$\sigma^2 = \mathbb{E}\left((X-\mu)\right)^2 = \frac{J^2 - 1}{12}$$

$$\sigma = \sqrt{\frac{J^2-1}{12}}$$

Example: Bernoulli, θ $J = 2$

$$x_1 = 0,\ x_2 = 1 \qquad p_1 = 1-\theta,\ p_2 = \theta$$

$$\mu = \mathbb{E}(X) = \sum_{j=1}^{2} x_j p_j = 0 \cdot (1-\theta) + 1 \cdot \theta = \theta$$

$$\begin{aligned}\sigma^2 = \mathbb{E}\left((X-\mu)^2\right) &= \sum_{j=1}^{2} (x_j - \overset{\theta}{\mu})^2 p_j \\ &= \theta^2 \cdot (1-\theta) + (1-\theta)^2 \theta = \theta \cdot (1-\theta)\end{aligned}$$

$$\sigma = \sqrt{\theta(1-\theta)}$$

Note for $\theta \to 0$ or $\theta \to 1$, $\sigma \to 0$: "sure thing".

Functions of Random Variables

Let

$$\underset{\text{new r.v.}}{Y} = \overset{\text{given function}}{g}(X) \quad \text{for} \quad X \sim f_X$$

X distributed according to...

Then for Y,

$$\text{Sample space} \equiv \{g(x_1), \ldots, g(x_{J_X})\} \text{ pruned}$$

$$\{y_1, y_2, \ldots y_{J_Y}\}$$

$$f_Y(y_i) = P(X = \text{any } x_j \text{ s.t. } g(x_j) = y_i) \quad \cup$$

$$= \sum_{g(x_j) = y_i} f_X(x_j), \quad 1 \le i \le J_Y .$$

Note

$$E_Y(Y) = \sum_{i=1}^{J_Y} y_i f_Y(y_i)$$

$$= \sum_{i=1}^{J_Y} y_i \sum_{g(x_j)=y_i} f_X(x_j)$$

$$= \sum_{i=1}^{J_Y} \sum_{g(x_j)=y_i} y_i f_X(x_j)$$

$$= \sum_{i=1}^{J_Y} \sum_{g(x_j)=y_i} g(x_j) f_X(x_j) = \sum_{j=1}^{J_X} g(x_j) f_X(x_j)$$

each x_j appears once and only once

$$= E_X(g(Y)) .$$

Example: uniform to Bernoulli

$$X \sim f_X^{\text{unif}, J=3} \qquad x_j = j, \; p_j = \tfrac{1}{3}, \quad 1 \le j \le J \equiv 3$$

$$g(x) = \begin{cases} 0 & \text{if } x = 1 \text{ or } x = 2 \\ 1 & \text{if } x = 3 \end{cases}$$

$\Rightarrow J_Y = 2, \; y_1 = 0, \; y_2 = 1$, and

$$\begin{cases} f_Y(y_1) = P(Y = 0) = P(X = 1 \text{ OR } X = 2) \\ \qquad\qquad = f_X(1) + f_X(2) = \tfrac{2}{3} \\ f_Y(y_2) = P(Y = 1) = P(X = 3) = \tfrac{1}{3} \end{cases}$$

Bernoulli with parameter $\theta = \frac{1}{3}$

Random Vectors

Joint pmf:

(X,Y) r. vector sample space $\{(x,y)_1, \ldots, (x,y)_J\}$

$\rightarrow \{(x_i,y_j),\ 1 \le i \le J_X,\ 1 \le j \le J_Y\}$

$$f_{X,Y}(x_i,y_j) = P(X = x_i,\ Y = y_j) \quad \text{AND}$$
$$= p_{i,j}^{X,Y}, \quad 1 \le i \le J_X,\ 1 \le j \le J_Y$$

where

$$\begin{cases} 0 \le p_{i,j}^{X,Y} \le 1, & 1 \le i \le J_X,\ 1 \le j \le J_Y \\ \displaystyle\sum_{i,j}^{J_X,J_Y} p_{i,j}^{X,Y} = 1 \end{cases}$$

Marginal pmf's

$$f_X(x_i) = P(X = x_i)$$
$$= P(X = x_i, Y = y_1 \text{ OR } X = x_i, Y = y_2 \text{ OR} \cdots)$$
$$= \sum_{j=1}^{J_Y} P(X = x_i, Y = y_j)$$
$$= \sum_{j=1}^{J_Y} f_{X,Y}(x_i,y_j), \quad 1 \le i \le J_X$$

$$f_Y(y_j) = \sum_{i=1}^{J_X} f_{X,Y}(x_i,y_j), \quad 1 \le j \le J_Y$$

Conditional pmf's

$$f_{X|Y}(x_i|y_j) = \frac{f_{X,Y}(x_i,y_j)}{f_Y(y_j)} \qquad 1 \le i \le J_X$$

$$1 \le j \le J_Y$$

$$f_{Y|X}(y_j|x_i) = \frac{f_{X,Y}(x_i,y_j)}{f_X(x_i)}$$

... Bayes' Theorem.

Independence

X and Y are independent if

$$f_{X,Y}(x_i, y_j) = f_X(x_i)\, f_Y(y_j) \sim p_{i,j}^{X,Y} = p_i^X \cdot p_j^Y$$

$1 \le i \le J_X$

$1 \le j \le J_Y$

or

$$f_{X|Y}(x_i | y_j) = f_X(x_i)$$

$$f_{Y|X}(y_j | x_i) = f_Y(y_j) \;.$$

Expectation of sums

$$X \sim f_X\,,\; Y \sim f_Y$$

$$\mathbb{E}_{X,Y}(g(X) + h(Y)) = \sum_{i,j} p_{i,j}^{X,Y} (g(x_i) + h(y_j))$$

$$= \sum_{i,j} p_{i,j}^{X,Y} g(x_i) + \sum_{i,j} p_{i,j}^{X,Y} h(y_j)$$

$$= \mathbb{E}_{X,Y}(g(X)) + \mathbb{E}_{X,Y}(h(Y))$$

$$\left(= \mathbb{E}_X(g(X)) + \mathbb{E}_Y(h(Y)) \quad \text{if } X, Y \text{ independent} \right)$$

Expectation of products

$X \sim f_X$, $Y \sim f_Y$ independent r.v.'s

$$\mathbb{E}(g(X) \cdot h(Y)) = \sum_{i,j} p_{i,j}^{X,Y} g(x_i)\, h(y_j)$$

$$= \sum_{i,j} p_i^X p_j^Y g(x_i)\, h(y_j)$$

$$= \sum_i p_i^X g(x_i) \sum_j p_j^Y h(y_j)$$

$$= \mathbb{E}_X(g(X))\, \mathbb{E}_Y(g(Y))$$

The Binomial Distribution

i.i.d. Bernoulli trials:

Let

$$X_1 \sim f_X^{\text{Bernoulli}}, \; X_2 \sim f_X^{\text{Bernoulli}}, \; \ldots, \; X_n \sim f_X^{\text{Bernoulli}}$$

sample from Bernoulli population for given θ

be n independent identically distributed (i.i.d.) r.v.'s.

Define new random variables

$$Z_n = \sum_{i=1}^{n} X_n \;\; (\text{\# of 1's}), \quad \underbrace{\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_n}_{\text{sample mean}} \;\; (\text{fraction of 1's}).$$

Note each experiment draws

n Bernoulli r.v.'s $\to Z_n, \bar{X}_n$.

(Pseudo) random variate generation: $\bar{X}_n$ $\theta = 1/2$

```
n = ?   % size of Bernoulli sample (r. vector)
num_exp = ?   % # of realizations of X̄n
xbar_n_vec = zeros(1, num_exp)
for i_exp = 1: num_exp
    bern_r_vector = randi([0,1], 1, n)
    xbar_n_vec(i_exp) = sum(bern_r_vector)/n
end
```

DEMO

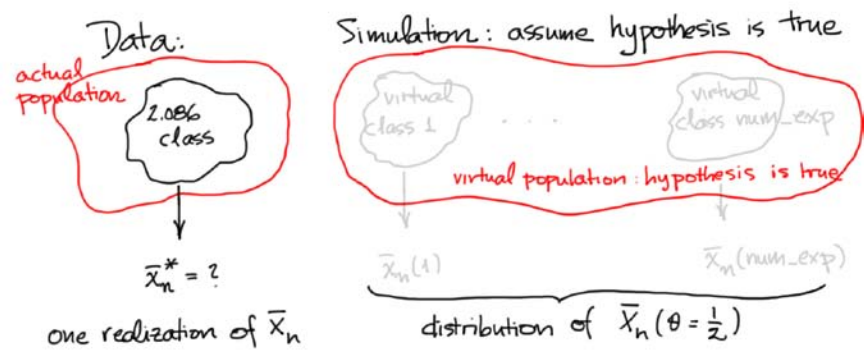
Birthmonth Revisited:

Hypothesis:

$$X = \begin{cases} 0 & \text{if birthmonth is [Jan – June]} \\ 1 & \text{if birthmonth is [July – Dec]} \end{cases}$$

is Bernoulli with parameter $\theta = 1/2$,

$$X \sim f_X^{\text{Bernoulli}}(x; \theta = 1/2).$$



If $\bar{x}_n^*$ is extremely unlikely with respect to distribution (pmf) of $\bar{X}_n(\theta = \frac{1}{2})$

REJECT hypothesis; otherwise, ACCEPT.

properties of binomial distribution: parameter θ

pmf: or $Z_n = k$

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$

$$P\left(\bar{X}_n = \frac{k}{n}\right) = \underbrace{\binom{n}{k}\theta^k(1-\theta)^{n-k}}_{\text{binomial}}, \quad k = 0,1,2,\dots n$$

note

$$P(\bar{X}_n = 0) = 1\cdot\theta^0(1-\theta)^n = (\text{for } \theta = \tfrac{1}{2})\ \left(\tfrac{1}{2}\right)^n \quad n \to \infty$$

$$\llcorner\ P(X_1 = 0 \text{ AND } X_2 = 0 \text{ AND } \dots X_n = 0) \underset{\theta=\frac{1}{2}}{=} \tfrac{1}{2}\cdot\tfrac{1}{2}\cdots\tfrac{1}{2}$$

only one "way" to get $\bar{X}_n = 0$: 0, 0, 0, …, 0

but many ways to get $\bar{X}_n \approx \frac{1}{2}$:

0, 1, 0, 1, … OR 1, 0, 1, 0, … OR 1, 0, 0, 1, 1, 0, 0, 1 … OR

mean:

$$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^n X_n\right) \overset{\text{i.i.d.}}{=} \frac{1}{n}\sum_{i=1}^n \overbrace{E_{X_n}(X_n)}^{\theta} = \theta \qquad \mu_{\bar{X}_n}$$

hence $\bar{X}_n$ is an estimator for θ

$\bar{x}_n$ is an estimate for θ

variance, std dev: Appendix A $\qquad \sigma^2_{\bar{X}_n}, \sigma_{\bar{X}_n}$

$$E((\bar{X}_n - \theta)^2) = \frac{1}{n}E((X_\cdot - \theta)^2) = \frac{\theta(1-\theta)}{n}$$

$$\Rightarrow \sigma^2_{\bar{X}_n} = \frac{\theta(1-\theta)}{n}, \quad \sigma_{\bar{X}_n} = \sqrt{\frac{\theta(1-\theta)}{n}}$$

hence $\bar{X}_n$ is a good estimator for θ for large n,

since large deviations $|\bar{X}_n - \theta|$ are unlikely

Appendix A

$$\sigma^2_{\bar{X}_n} = E\left((\bar{X}_n - \theta)^2\right) = E\left(\left(\frac{1}{n}\sum_{i=1}^n X_i - \theta\right)^2\right)$$

$$= E\left(\left(\frac{1}{n}\sum_{i=1}^n (X_i - \theta)\right)\left(\frac{1}{n}\sum_{k=1}^n (X_k - \theta)\right)\right)$$

$$= \frac{1}{n^2}E\left(\sum_{i=1}^n\sum_{k=1}^n (X_i - \theta)(X_k - \theta)\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^n\sum_{k=1}^n E\left((X_i - \theta)(X_k - \theta)\right)$$

but if $i \neq k$,

$$E((X_i - \theta)(X_k - \theta)) \overset{\text{independence}}{=} E_{X_i}(X_i - \theta)\cdot \overset{\text{mean}}{E_{X_k}(X_k - \theta)} = 0$$

and hence

$$\sigma^2_{\bar{X}_n} = \frac{1}{n^2}\sum_{i=1}^n \overset{\text{variance of Bernoulli r.v.}}{E((X_i - \theta)^2)} = \frac{1}{n^2}\cdot n\cdot\theta(1-\theta)$$

$$= \frac{\theta(1-\theta)}{n}$$

$$\sigma_{\bar{X}_n} = \sqrt{\frac{\theta(1-\theta)}{n}}$$

quite famous $\sqrt{n}$