CS 760: Machine Learning © Fall 2020

Homework 1: Review

Instructor: Daniel L. Pimentel-Alarcón Due 09/24/2020

GO GREEN. AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

In this homework you will review some basic linear algebra, probability, and optimization concepts. Each sub problem is worth 10 points.

Problem 1.1. Show that $\mathbb{R}^{\mathrm{D}}$ is a subspace.

Problem 1.2. Subspaces spaces are, by definition, *closed* under linear combinations. For example, when you add or multiply elements of $\mathbb{R}^{\mathrm{D}}$, you end up with an element of $\mathbb{R}^{\mathrm{D}}$ (as shown in Problem 1.1). In other words, you cannot *fall* out of $\mathbb{R}^{\mathrm{D}}$ by adding or multiplying. Subspaces are not necessarily closed under *all* mathematical operations.

(a) Show that $\mathbb{R}^{\mathrm{D}}$ is *not* closed under element-wise square roots.

(b) Give an example of a subspace that is closed under element-wise square roots (besides being closed under linear combinations, as *all* subspaces must be).

Problem 1.3. Let $\mathbf{u}_1, \ldots, \mathbf{u}_{\mathrm{R}} \in \mathbb{R}^{\mathrm{D}}$. Show that $\mathbb{U} = \mathrm{span}[\mathbf{u}_1, \ldots, \mathbf{u}_{\mathrm{R}}]$ is a subspace.

Problem 1.4 (Diabetes testing)**.** With 9.3% of the U.S. population having diabetes, there is an increasing interest in studying this disease. Geneticists have determined that 95% of the people that develop diabetes have the following genes inactive:

- TCF7L2. Affects insulin secretion and glucose production.
- ABCC8. Helps regulate insulin.
- GLUT2. Helps move glucose into the pancreas.

(a) If you sequence your genome and find out that these genes are inactive, what is the probability that you develop diabetes?

(b) What other information would you need to know?

(c) Based on this information, when should you be concerned?

Problem 1.5 (Snapchat's delays)**.** Suppose that you are sending pics to your girlfriend/boyfriend overseas. Each time you send a picture through the Internet it takes a certain amount of time to reach your gf/bf. Assume that you can measure the time delay. The delay won't be constant, since it depends on the traffic of the Internet (in particular at the routers that handle your messages). You and your gf/bf measure the delays of several packet transmissions. It appears that there is a minimal time delay, say t_0 (msec). Based on your observations, it seems that larger delays are rarer than shorter ones. Propose a probabilistic model for the delays with a single free parameter θ. The value of θ should govern the expected delay characteristics. Let x denote a random variable that represents the delay. The observations you have made are assumed to be independent realizations of this random variable. Let $\mathbb{P}(\mathrm{x}|\theta)$ denote the probability density of x. Give an explicit form for $\mathbb{P}(\mathrm{x}|\theta)$ and explain the rationale of your model.

Problem 1.6 (Simulating random variables)**.** In this problem you will simulate random variables and study their distributions.

(a) Generate $\mathrm{N} = 1000$ i.i.d. Uniform$(0, 1)$ random variables $x_1, \ldots, x_\mathrm{N}$, and plot their histogram. Does it look fairly uniform?

(b) Let

$$y_\mathrm{i} = \begin{cases} 1 & \text{if } x_\mathrm{i} \leq \mathrm{p} \\ 0 & \text{otherwise.} \end{cases}$$

What is the distribution of y_i?

(c) Plot the histogram of the y_i's with $\mathrm{p} = 1/4, 1/2, 3/4$. Do these histograms match the distribution of your answer from (b)?

(d) Let z_k be the sum of the k^{th} batch of n y_i's. What is the distribution of z_k?

(e) Plot the histogram of the z_k's with $\mathrm{n} = 10$ and $\mathrm{p} = 1/4, 1/2, 3/4$. Do these histograms match the distribution of your answer from (d)?

Problem 1.7 (Logistic Gradient)**.** The following expression describes the log-likelihood of the logistic regression model:

$$\ell(\boldsymbol{\theta}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \left[\mathrm{y_i} \log \left(\frac{1}{1 + e^{-\boldsymbol{\theta}^\mathsf{T} \mathbf{x}_\mathrm{i}}} \right) + (1 - \mathrm{y_i}) \log \left(\frac{1}{1 + e^{\boldsymbol{\theta}^\mathsf{T} \mathbf{x}_\mathrm{i}}} \right) \right].$$

The goal in logistic regression is to maximize this quantity.

(a) Derive an expression for its gradient (w.r.t. $\boldsymbol{\theta}$).

(b) Derive an expression for its Hessian.

(c) Is $\ell(\boldsymbol{\theta})$ a scalar, a vector, or a matrix? What about its gradient? What about its Hessian?

Problem 1.1. Show that $\mathbb{R}^D$ is a subspace.

R^D is a vector space over field $K = R$

Consider $\forall \vec{w}, \vec{v} \in R^D$

$\vec{w} = (w_1 \cdots w_D)^T$ $\quad$ $\vec{v} = (v_1 \cdots v_D)^T$

$\forall \alpha, \beta \in K = R$

$\alpha\vec{w} + \beta\vec{v} = (\alpha w_1 + \beta v_1, \cdots, \alpha w_D + \beta v_D) \in R^D$

R^D is a subspace ∎

Problem 1.2. Subspaces spaces are, by definition, *closed* under linear combinations. For example, when you add or multiply elements of $\mathbb{R}^D$, you end up with an element of $\mathbb{R}^D$ (as shown in Problem 1.1). In other words, you cannot *fall* out of $\mathbb{R}^D$ by adding or multiplying. Subspaces are not necessarily closed under *all* mathematical operations.

(a) Show that $\mathbb{R}^D$ is *not* closed under element-wise square roots.

(b) Give an example of a subspace that is closed under element-wise square roots (besides being closed under linear combinations, as *all* subspaces must be).

(a) Consider $W \subseteq R^D$ $\quad \forall \vec{w} \in W$.

$\vec{w} = (w_1 \cdots w_D)^T$ $\quad w_i \in Q$ $\quad i = 1 \cdots D$ i.e. w_i is rational number

Then W is closed under linear combinations

Consider $\vec{w} = (2, 0, \cdots 0)^T$ $\quad$ V is element-wise square roots of W

$v = (\sqrt{2}, 0 \cdots 0)^T \notin W$ ∎

(b) Consider $W = \{ a \in R : a \geq 0 \}$. is a vector space over K

$K = \{ b \in R : b \geq 0 \}$. $\forall \alpha, \beta \in K$ $\quad v_1 = a \in W$ $\quad v_2 = b \in W$

$\alpha v_1 + \beta v_2 = \alpha a + \beta b \in W$

$\sqrt{v_1} \geq 0 \in W$ ∎

Problem 1.3. Let $\mathbf{u}_1, \ldots, \mathbf{u}_R \in \mathbb{R}^D$. Show that $\mathbb{U} = \text{span}[\mathbf{u}_1, \ldots, \mathbf{u}_R]$ is a subspace.

Since $U = \text{span}(u_1 \cdots u_R)$

we have $\forall V_1, V_2 \in U \quad \exists \; a = (a_1 \cdots a_R)^T, \quad b = (b_1 \cdots b_R)^T$

$$V_1 = \sum_{i=1}^{R} a_i u_i$$

$$V_2 = \sum_{i=1}^{R} b_i u_i$$

$\forall \alpha, \beta \in R$

$$\alpha V_1 + \beta V_2 = \sum_{i=1}^{R} (\alpha a_i + \beta b_i) u_i \in U$$

Then U is a subspace ∎

Problem 1.4 (Diabetes testing)**.** With 9.3% of the U.S. population having diabetes, there is an increasing interest in studying this disease. Geneticists have determined that 95% of the people that develop diabetes have the following genes inactive:

- TCF7L2. Affects insulin secretion and glucose production.
- ABCC8. Helps regulate insulin.
- GLUT2. Helps move glucose into the pancreas.

(a) If you sequence your genome and find out that these genes are inactive, what is the probability that you develop diabetes?

(b) What other information would you need to know?

(c) Based on this information, when should you be concerned?

Use A denote the event people have diabetes

B denote people have genes inactive

$P(A) = 9.3\%$

$P(B|A) = 95\%$

$P(A \cap B) = P(A) \cdot P(B|A)$

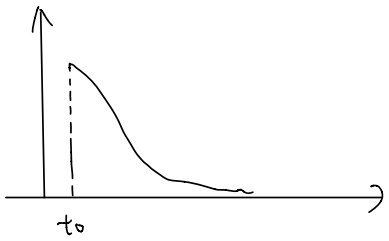
(a) $$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(A^c) \cdot P(B|A^c)}$$

I need the information of $P(B|A^c)$ to compute $P(A|B)$

(b) $P(B|A^c)$: What percent of people that do not have diabetes have the above genes inactive?

(c) We need to concern about the authenticity and veracity of the information. We need to verify whether the 95% in the description is reasonable to represent $P(B|A)$. To verify this, we need to make sure we have random fair sample and there is no other factors disturb the sample results.

Problem 1.5 (Snapchat's delays)**.** Suppose that you are sending pics to your girlfriend/boyfriend overseas. Each time you send a picture through the Internet it takes a certain amount of time to reach your gf/bf. Assume that you can measure the time delay. The delay won't be constant, since it depends on the traffic of the Internet (in particular at the routers that handle your messages). You and your gf/bf measure the delays of several packet transmissions. It appears that there is a minimal time delay, say t_0 (msec). Based on your observations, it seems that larger delays are rarer than shorter ones. Propose a probabilistic model for the delays with a single free parameter θ. The value of θ should govern the expected delay characteristics. Let x denote a random variable that represents the delay. The observations you have made are assumed to be independent realizations of this random variable. Let $\mathbb{P}(\mathrm{x}|\theta)$ denote the probability density of x. Give an explicit form for $\mathbb{P}(\mathrm{x}|\theta)$ and explain the rationale of your model.



The data follows normal distribution with mean t_0, $\sigma^2 = \theta$. and data only has right side of data.

The rationale behind this model is the fact larger delays are rarer than shorter ones, and there is a minimal time delay t_0. θ represents how higher the delay is.

Problem 1.7 (Logistic Gradient). The following expression describes the log-likelihood of the logistic regression model:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left[y_i \log\left(\frac{1}{1+e^{-\boldsymbol{\theta}^T \mathbf{x}_i}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{\boldsymbol{\theta}^T \mathbf{x}_i}}\right)\right].$$

The goal in logistic regression is to maximize this quantity.

(a) Derive an expression for its gradient (w.r.t. $\boldsymbol{\theta}$).

(b) Derive an expression for its Hessian.

(c) Is $\ell(\boldsymbol{\theta})$ a scalar, a vector, or a matrix? What about its gradient? What about its Hessian?

(a) $l(\theta) = \sum_{i=1}^{n} \left[y_i \theta^T x_i - \log\left(1+e^{\theta^T x_i}\right)\right]$

$$\frac{\partial l}{\partial \theta_j} = \sum_{i=1}^{n} \left[y_i \cdot x_{ij} - \frac{e^{\theta^T x_i}}{1+e^{\theta^T x_i}} \cdot x_{ij} \right]$$

$$= \sum_{i=1}^{n} \left(y_i - \frac{e^{\theta^T x_i}}{1+e^{\theta^T x_i}} \right) x_{ij}$$

$$\frac{\partial l}{\partial \theta} = \sum_{i=1}^{n} \left(y_i - \frac{e^{\theta^T x_i}}{1+e^{\theta^T x_i}} \right) x_i$$

Consider $\tilde{y}$ such that $\tilde{y}_i = \frac{e^{\theta^T x_i}}{1+e^{\theta^T x_i}}$ $\qquad X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$

$$\frac{\partial l}{\partial \theta} = X^T (y - \tilde{y})$$

(b) $$\frac{\partial}{\partial \theta_j}\left(\frac{\partial l}{\partial \theta}\right) = \sum_{i=1}^{n} \left(- \frac{e^{\theta^T x_i}}{(1+e^{\theta^T x_i})^2} \cdot x_{ij} \cdot x_i \right)$$

$$= - \sum_{i=1}^{n} \frac{e^{\theta^T x_i}}{(1+e^{\theta^T x_i})^2} x_i \cdot x_{ij}$$

$$\frac{\partial^2 l}{\partial \theta \partial \theta^T} = - \sum_{i=1}^{n} \frac{e^{\theta^T x_i}}{(1+e^{\theta^T x_i})^2} x_i \cdot x_i^T \quad \text{is Hessian}$$

(c) $l(\theta)$ is a scalar

$\frac{\partial l}{\partial \theta}$ is a vector

Hessian $\frac{\partial^2 l}{\partial \theta \partial \theta^T}$ is negative semi-definite matrix

Homework 1.6

September 21, 2020

1 hw1.6

[1]:
```
import numpy as np
import pandas as pd
```

2 (a)

[2]:
```
data = np.random.uniform(size = 1000)
data = pd.Series(data)
data.plot.hist()
None
```

It shows fairly uniform by the histogram.

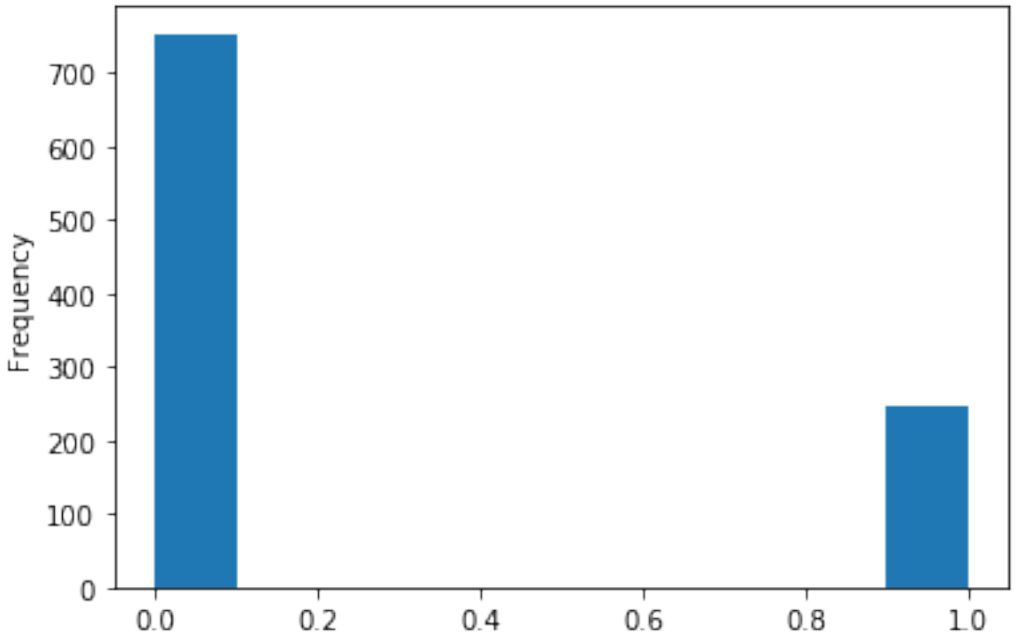
2.1 (b)

$$P(y_i = 1) = P(x_i \leq p) = p$$
$$P(y_i = 0) = P(x_i > p) = 1 - p$$

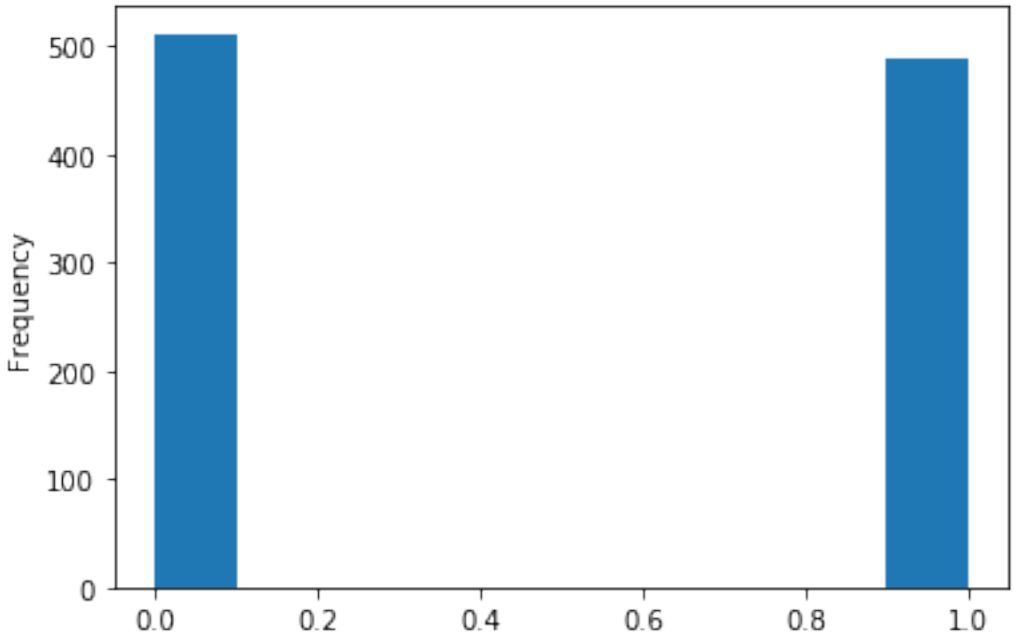
y_i is Bernoulli distribution.

2.2 (c)

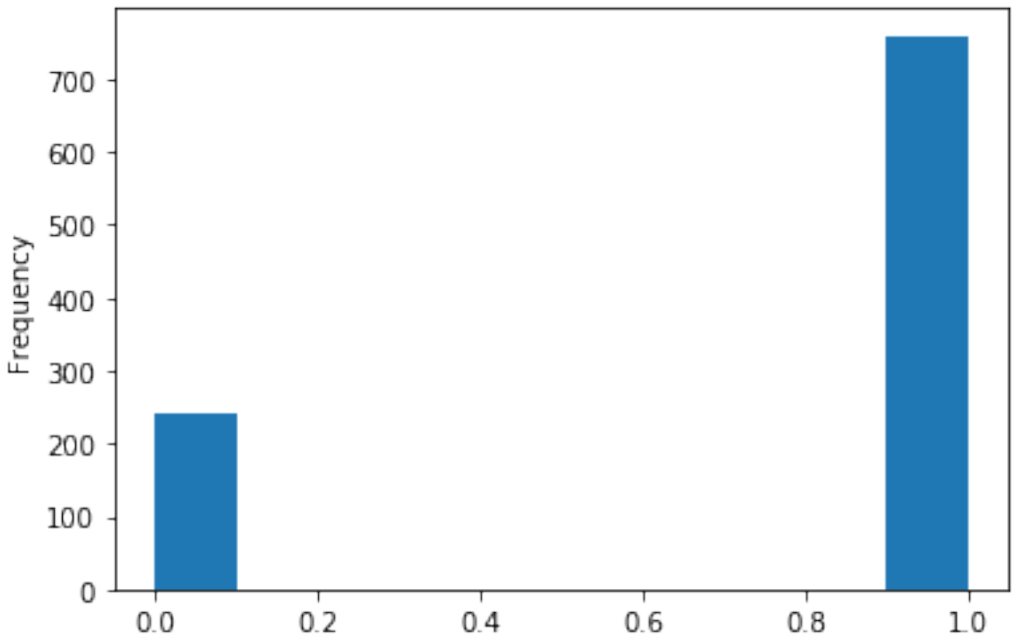
[3]:
```
y1 = data.apply(lambda x: 1 if x <= 1/4 else 0)
y1.plot.hist()
None
```



[4]:
```
y2 = data.apply(lambda x: 1 if x <= 1/2 else 0)
y2.plot.hist()
None
```



[5]:
```
y3 = data.apply(lambda x: 1 if x <= 3/4 else 0)
y3.plot.hist()
None
```



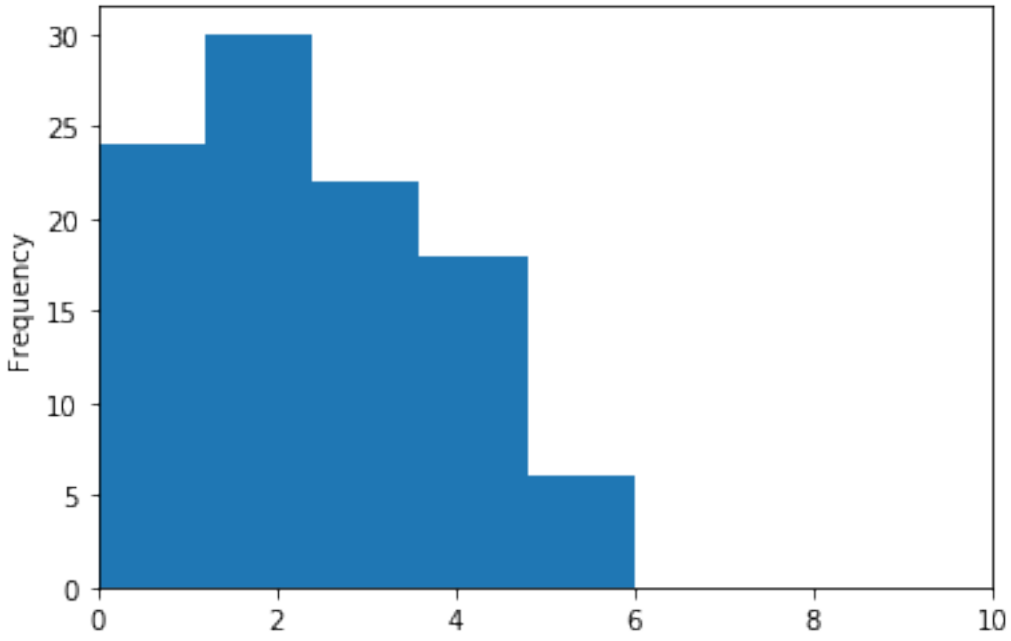
These histograms match with the comclusion from (b).

2.3 (d)

$$z_k = \sum_{i=1}^{n} y_{ki} \sim Binomial(n, p)$$

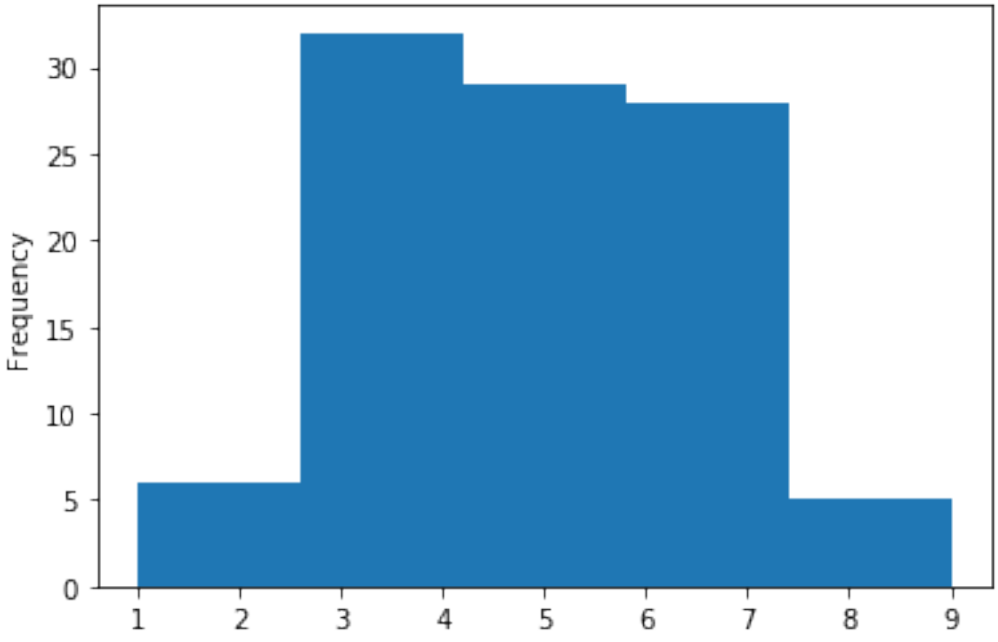
2.4 (e)

[11]:
```
z1 = np.array(y1).reshape(-1,10).sum(1)
ax = pd.Series(z1).plot.hist(bins = 5)
ax.set_xlim([0,10])
None
```



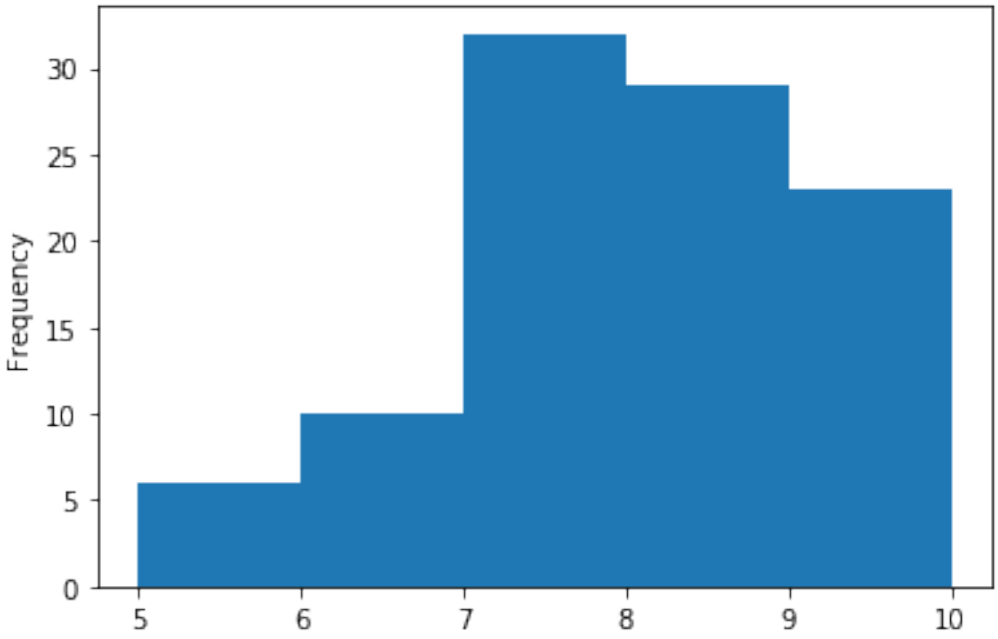
[12]:
```
z2 = np.array(y2).reshape(-1,10).sum(1)
pd.Series(z2).plot.hist(bins = 5)
ax.set_xlim([0,10])
```

[12]: (0, 10)



[13]:
```
z3 = np.array(y3).reshape(-1,10).sum(1)
pd.Series(z3).plot.hist(bins = 5)
ax.set_xlim([0,10])
```

[13]: (0, 10)



These histograms match the conclusion in (d). They are binomial distribution, with larger p, the data become more left skewed.

[ ]:

[ ]:

[ ]: