

# Random Variables and Probability Distributions

IE231 - Lecture Notes 5

Mar 14, 2017

## Some Preliminary Information

### Scales on Measurement

- Nominal scale: These are categorical values that has no relationship of order or rank among them. (e.g. colors, species)
- Ordinal scale: These are categorical values that has relationship of order or rank among them (e.g. military ranks, competition results). Though the relative order has no defined magnitude (e.g. Champion can get 40 points, runner up 39 and third place 30).
- Interval scale: There is a numerical order but the difference can only be defined in intervals, *since there is no absolute minimum*. We cannot compare in relative values. For instance, we cannot say 10 degree celsius is twice as hot as 5 degree celsius; what about -5 vs +5?
- Ratio scale: Scale with an absolute minimum. (e.g. If I have 50TL and my friend has 100TL, I can say that she has twice the money that I have.) Height, weight, age are similar examples.

See more on [https://en.wikipedia.org/wiki/Level\\_of\\_measurement](https://en.wikipedia.org/wiki/Level_of_measurement).

### Infinity

The concept of infinity is very broad. Currently, you just need to keep the distinction of countable and uncountable infinities in mind.

- Countably infinite: 1, 2, 3, 4, ... (e.g. natural numbers, integers, **rational numbers**)
- Uncountably infinite: 1, 1.01, 1.001, 1.0001, 1.00001, ... (e.g. real numbers)

How many real numbers are there between 0 and 1?

### Descriptive Statistics

Here are brief descriptions of mean (expectation), median, mode, variance, standard deviation, quantile.

- Mean:  $\bar{X} = \sum_i^N X_i$
- Median: Let's say  $X_k$  are ordered from smallest to largest and there are  $n$  values in the sample. Median( $X$ ) =  $X_{(n+1)/2}$  if  $n$  is odd and (usually) Median( $X$ ) =  $\frac{X_{(n/2)} + X_{(n/2+1)}}{2}$ .
- Quantile: On an ordered list of values for quantile ( $\alpha$ ) provides the  $(\alpha * n)^{th}$  smallest value of the list. For instance, if  $\alpha = 70\% = 0.7$  quantile value is the 7th smallest value in a list of 10 values.  $\alpha = 1$  means the maximum. Quantile is an important parameter in especially statistics.
- Mode:  $X_k$  with the highest frequency in the sample. In a sample of (1, 2, 2, 3, 4, 5), 2 is the mode.
- Variance:  $V(X) = \frac{\sum_i^N (X_i - \bar{X})^2}{n - 1}$
- Standard Deviation:  $\sigma(X) = \sqrt{\frac{\sum_i^N (X_i - \bar{X})^2}{n - 1}}$

```

set.seed(231)
#Let's pick 10 values from the numbers between 1 and 50.
numbers <- sample(1:50,10,replace=TRUE)
#The sorted version of the numbers
sort(numbers)

## [1] 1 9 15 16 16 18 26 31 32 35
#The mean values of the numbers
sum(numbers)/10

## [1] 19.9
#or in R
mean(numbers)

## [1] 19.9
#Median of the numbers
median(numbers)

## [1] 17
#Quantile 7/9 of the numbers
quantile(numbers,7/9)

## 77.77778%
##      31
#Quantile 0 of the numbers (also the min)
quantile(numbers,0)

## 0%
##      1
#Quantile 1 of the numbers (also the max)
quantile(numbers,1)

## 100%
##      35
#No simple solution for mode in R
freq_table<-table(numbers)
freq_table

## numbers
## 1 9 15 16 18 26 31 32 35
## 1 1 1 2 1 1 1 1 1
names(freq_table[which.max(freq_table)])

## [1] "16"
#Sample variance of numbers
sum((numbers - mean(numbers))^2)/(10-1)

## [1] 118.7667
#For large values you can take n ~ n-1
#in R
var(numbers)

```

```
## [1] 118.7667
#Sample standard deviation of values
sqrt(sum((numbers - mean(numbers))^2)/(10-1))

## [1] 10.89801
#in R
sd(numbers)

## [1] 10.89801
```

## Random Variables

Random variables are the abstractions of uncertain events so that we can generalize events in formal functions instead of explicitly enumerating the outcomes. For instance, assume  $X$  is the number of tails in 2 coin tosses.

$$P(X = 0) = P(\{H, H\}) = 1/4 \tag{1}$$

$$P(X = 1) = P(\{H, T\}, \{T, H\}) = 2/4 \tag{2}$$

$$P(X = 2) = P(\{T, T\}) = 1/4 \tag{3}$$

$$\tag{4}$$

$X$  can take values 0, 1 and 2.  $X$  is a discrete random variable.

There are also the continuous random variables. Continuous random variables are usually defined in intervals instead of individual values. For instance, define  $Y$  as any real number between 0 and 1 and all values within the interval are equally probable (i.e. uniform distribution).

$$P(Y \leq 0.25) = 1/4 \tag{5}$$

$$P(X \leq 0.5) = 2/4 \tag{6}$$

$$P(X \leq 0.75) = 3/4 \tag{7}$$

$$\tag{8}$$

## Fundamental Concepts

There are several fundamental concepts to keep in mind.

**Probability Mass Function (pmf):** pmf is the point probability for discrete distributions (i.e.  $f(x) = P(X = x)$ ). For instance  $P(X = H) = 1/2$ ,  $P(X = T) = 1/2$ .

$$\sum_i^n f(x_i) = 1$$

- **Probability Density Function (pdf):** pdf is the interval probability for continuous distributions (i.e.  $f(x) = P(a < X < b) = \int_a^b f(x)dx$ ). Since almost all point probabilities in continuous distributions are 0 (due to infinity), intervals.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

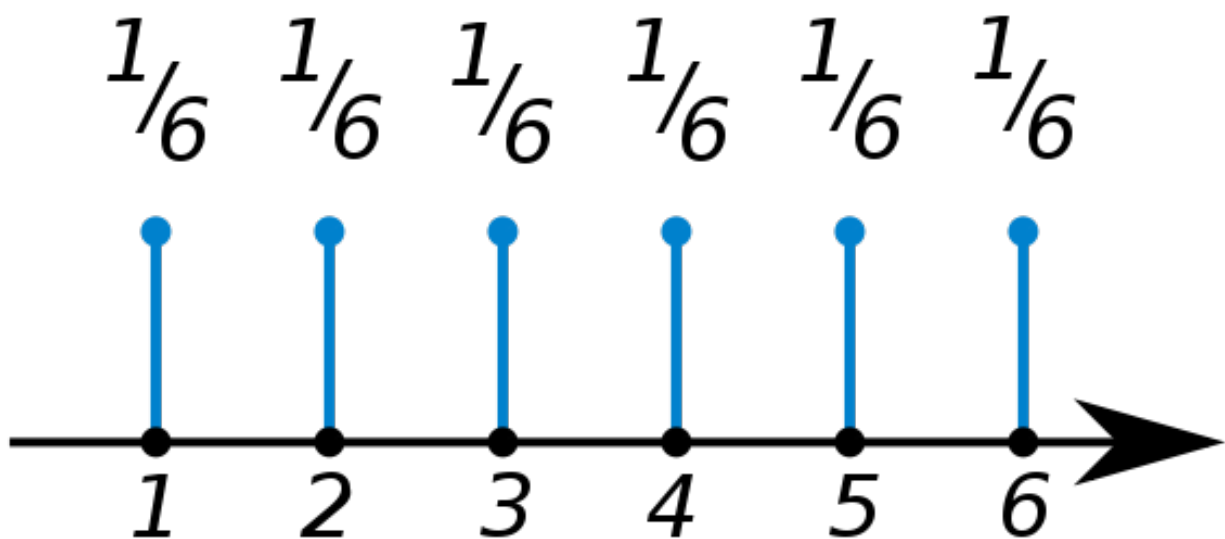


Figure 1:

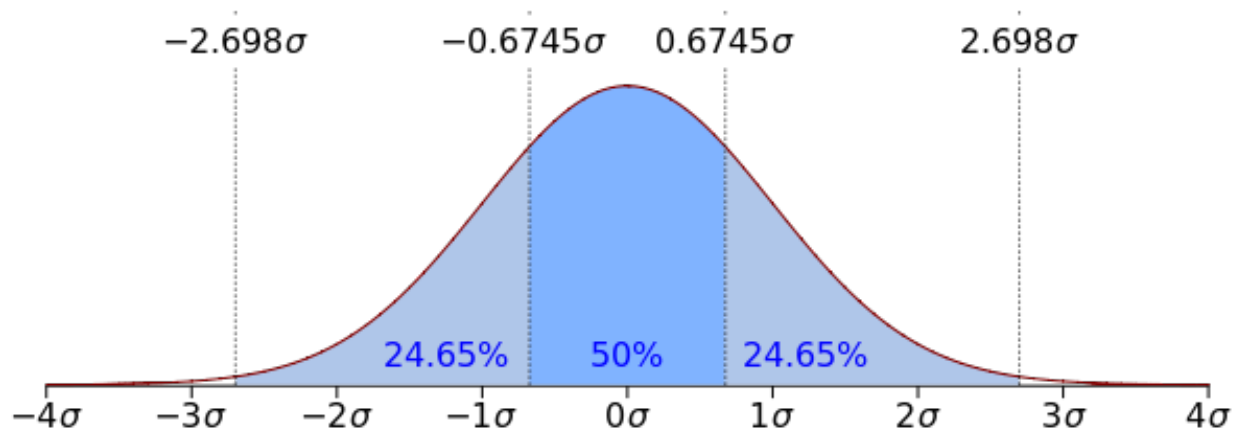


Figure 2:

- **Cumulative Distribution Function (cdf):** cdf is the cumulative probability for all values smaller than  $x$  (i.e.  $F(x) = P(X \leq x)$ ). For the coin toss an example cdf would be two or less tails ( $P(X \leq 2)$ ).

Main relationship between cdf and pdf is  $(F(X \leq a) = \int_{-\infty}^a f(x)dx)$ .

- **Expected Value ( $E[X]$ ):** Expected value of a probability distribution is calculated as follows.

$$\mu = E[X] = \sum_i^n x_i f(x_i)$$

for discrete distributions.

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

for continuous distributions.

Example: Calculate the expected value of number of tails in two coin tosses.

$$E[X] = \sum_i^n x_i f(x_i) = 0 * P(X = 0) + 1 * P(X = 1) + 2 * P(X = 2) \tag{9}$$

$$= 0 * 1/4 + 1 * 1/2 + 2 * 1/4 \tag{10}$$

$$= 1 \tag{11}$$

- **Variance ( $V(X)$ ):** Variance is calculated as follows

$$V(X) = E[(X - \mu)^2] = \sum_i^n (x_i - \mu)^2 f(x_i)$$

for discrete distributions.

$$V(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

for discrete distributions.

Variance can also be calculated as  $V(X) = E[X^2] - (E[X])^2$ .

## Some Discrete Distributions

### Bernoulli Distribution

It can also be called “single coin toss distribution”. For a single event with probability of success  $p$  and failure  $q = 1 - p$ , the distribution is called Bernoulli.

- pmf:  $f(x = 0; p) = q, f(x = 1) = p$
- $E[X] = 0 * (1 - p) + 1 * p = p$
- $V[X] = pq$

Example: Coin Toss

- $p = 0.5, q = 1 - p = 0.5$
- pmf:  $f(x = 0) = 0.5, f(x = 1) = 0.5$

- $E[X] = 0 * (1 - 0.5) + 1 * 0.5 = 0.5$
- $V(X) = 0.5 * 0.5 = 0.25$

## Binomial Distribution

Think of multiple Bernoulli trials (e.g. several coin tosses).

- pmf:  $f(x; p, n) = \binom{n}{x} p^x q^{(n-x)}$
- $E[X] = np$
- $V(X) = npq$
- cdf:  $F(X \leq x) = \sum_{i=0}^x f(i)$

Example: Multiple Coin Tosses (x5 coins, p = 0.5)

- pmf:  $f(x = 3; n = 5) = \binom{5}{3} (0.5)^3 (1 - 0.5)^{(5-3)} = 0.3125$

```
#R way
#(d)ensity(binomial)
dbinom(x=3, size=5, prob=0.5)
```

```
## [1] 0.3125
```

- $E[X] = 5 * 0.5 = 2.5$
- $V(X) = 5 * 0.5 * 0.5 = 1.25$
- cdf:  $F(X \leq 3; n = 5) = \sum_{i=0}^3 f(i) = 0.8125$

```
#R way
pbinom(q=3, size=5, prob=0.5)
```

```
## [1] 0.8125
```

## Multinomial Distribution

Now suppose there is not one probability ( $p$ ) but there are many probabilities ( $p_1, p_2, \dots, p_k$ ).

- pmf:  $f(x_1, \dots, x_k; p_1, \dots, p_k; n) = \binom{n}{x_1, \dots, x_k} p_1^{x_1} * \dots * p_k^{x_k}$
- where  $\binom{n}{x_1, \dots, x_k} = \frac{n!}{x_1! \dots x_k!}$ ,  $\sum_i x_i = n$  and  $\sum_i p_i = 1$ .

Example: Customers of a coffee shop prefer Turkish coffee with probability 0.4, espresso 0.25 and filter coffee 0.35. What is the probability that out of the first 10 customers, 3 will prefer Turkish coffee, 5 will prefer espresso and 2 will prefer filter coffee?

$$f(3, 5, 2; 0.4, 0.25, 0.35; 10) = \binom{10}{3, 5, 2} * 0.4^3 * 0.25^5 * 0.35^2 = 4.3 * 10^{-6} = 0.0193$$

```
#Explicit form
factorial(10)/(factorial(3)*factorial(5)*factorial(2))*0.4^3 * 0.25^5 * 0.35^2
```

```
## [1] 0.01929375
```

```
#Density multinomial
dmultinom(x=c(3, 5, 2), prob=c(0.4, 0.25, 0.35))
```

```
## [1] 0.01929375
```

Binomial distribution is a special case of multinomial distribution.

## Hypergeometric Distribution

Hypergeometric distribution can be used in case the sample is divided in two such as defective/nondefective, white/black, Ankara/Istanbul. Suppose there are a total of  $N$  items,  $k$  of them are from group 1 and  $N - k$  of them are from group 2. We want to know the probability of getting  $x$  items from group 1 and  $n - k$  items from group 2.

- pmf:  $f(x, n; k, N) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}$
- $E[X] = \frac{nk}{N}$
- $V[X] = \frac{N-n}{N-1} * n * \frac{k}{N} * (1 - \frac{k}{N})$

Example: Suppose we have a group of 20 people, 12 from Istanbul and 8 from Ankara. If we randomly select 5 people from it what is the probability that 1 of them is from Ankara and 4 of them from Istanbul.

$$f(1, 5; 8, 20) = \frac{\binom{8}{1} \binom{20-8}{5-1}}{\binom{20}{5}} = 0.256$$

```
#Explicit form
```

```
x=1
```

```
n=5
```

```
k=8
```

```
N=20
```

```
(choose(k,x)*choose(N-k,n-x))/choose(N,n)
```

```
## [1] 0.255418
```

```
#Density hypergeometric, see ?dhyper for explanations
```

```
dhyper(x=1,m=8,n=12,k=5)
```

```
## [1] 0.255418
```

## Negative Binomial Distribution

Negative Binomial distribution answers the question “What is the probability that  $k$ -th success occurs in  $n$  trials?”. Differently from the binomial case, we fix the last attempt as success.

- pmf:  $f(x; p, n) = \binom{n-1}{x-1} p^x q^{(n-x)}$

Example: Suppose I'm repeatedly tossing coins. What is the probability that 3rd Heads come in the 5th toss?

$$f(3; 0.5, 5) = \binom{5-1}{3-1} 0.5^3 0.5^{(5-3)} = 0.1875$$

```
#Explicit form
```

```
choose(5-1,3-1)*0.5^3*0.5^(5-3)
```

```
## [1] 0.1875
```

```
#Binomial way
```

```
dbinom(3-1,5-1,0.5)*0.5
```

```
## [1] 0.1875
```

```
#Negative binomial way
```

```
dnbinom(x=5-3,size=3,prob=0.5)
```

```
## [1] 0.1875
```

## Geometric Distribution

Geometric distribution answers “What is the probability that first success comes in the n-th trial?”

- pmf:  $f(x; p, n) = q^{(n-1)}p$
- $E[X] = 1/p$
- $V[X] = \frac{1-p}{p^2}$