Fundamental sampling distribution and data description

Measures of location: Mean, Median

Mean: The mean is simply a numerical average. Suppose that the observation in a sample with size n, are $x_1, x_2, \dots, x_n$, the sample mean denoted by $\bar{X}$. If mean belongs to the population, we denoted it by μ.

$$Sample\ mean = \bar{X} = \frac{x_1+x_2+\cdots+x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n},\ Population\ mean = \mu = \frac{x_1+x_2+\cdots+x_N}{N} = \sum_{i=1}^{N} \frac{x_i}{N}$$

Median (measure of central tendency): The middle number (in a sorted list of numbers).
To find the Median, place the numbers you are given in value order and find the middle number.

$$\tilde{X} = \begin{cases} x_{\frac{n+1}{2}} & if\ n\ is\ odd \\ \frac{1}{2}\left[x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right] & if\ n\ is\ even \end{cases}$$

Example: Find $\bar{X}$ and $\tilde{X}$ of these values, 5, 7, 3, 2, 8→ $2, 3, 5, 7, 8$

$$\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{2+3+5+7+8}{5} = 5,\ \tilde{X} = x_{\frac{n+1}{2}} = 5$$

Example: Find $\bar{X}$ and $\tilde{X}$ of these values, 5, 7, 3, 2, 8, 11→ $2, 3, 5, 7, 8, 11$

$$\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{2+3+5+7+8+11}{6} = 7.2,\ \tilde{X} = \frac{1}{2}\left[x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right] = \frac{5+7}{2} = 6$$

Measures of variability: Range, Variance, Standard Deviation

Sample variability plays an important role in data analysis. Process and product variability is a fact in engineering and scientific systems.

$$Range = R = x_{max} - x_{min},$$

$$Sample\ Variance: S^2 = \sum_{i=1}^{n} \frac{(x_i-\bar{X})^2}{n-1} \qquad Population\ Variance: \sigma^2 = \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N}$$

Example: Find R, S^2, and S of these values, 23, 41, 18, 37, 54, 73, 38, 29

$$\rightarrow 18, 23, 29, 37, 38, 41, 54, 73$$

$$Range = R = x_{max} - x_{min} = 73 - 18 = 55,$$

$$\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{18+\ 23+\ 29+\ 37+\ 38+\ 41+\ 54+\ 73}{8} = 42.5$$

$$S^2 = \sum_{i=1}^{n} \frac{(x_i-\bar{X})^2}{n-1}$$
$$= \frac{(18-42.5)^2 + (23-42.5)^2 + (29-42.5)^2 + (37-42.5)^2 + (38-42.5)^2 + (41-42.5)^2 + (54-42.5)^2 + (73-42.5)^2}{7}$$
$$= 325.4286$$

$$S = \sqrt{S^2} = \sqrt{325.4286} = 18.04$$

Mode:

The sample mode is the most frequently occurring value in a sample

Example: Find mode of these values: 1, 1, 2, 2, 3, 4, 6, 9, 11, 11, 11, 11, 12, 15, 20, 21→ $Mode = 11$

Note: If instead of using the actual values of the data, we want to use a frequency table to calculate the mean, median, and variance, we can use the following formulas to calculate their estimates.

$$\bar{X} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}, \quad S^2 = \sum_{i=1}^{k} \frac{(x_i - \bar{X})^2 f_i}{k-1}$$

İn these formulas, x_i is the midpoint for interval of i

Example: Use data of 50 passengers from chapter 1, calculate exact and estimated values of mean, median, and variance and discuss about symmetric type of histogram.

8, 4, 6, 3, 7, 3, 7, 5, 4, 8, 3, 7, 15, 16, 15, 8, 4, 4, 3, 3, 9, 5, 12, 8, 7, 5, 9, 3, 8, 9, 22, 10, 3, 37, 7, 6, 8, 3, 5, 16, 4, 15, 3, 12, 6, 8, 12, 12, 3, 5

3	3	3	3	3	3	3	3	3	3	4	4	4
	4	4	5	5	5	5	5	6	6	6	7	7
	7	7	7	8	8	8	8	8	8	8	9	9
	9	10	12	12	12	12	15	15	15	16	16	22
	37											

Interval	1	2	3	4	5	6	7	8
CI	3-7.8	7.8-12.6	12.6-17.4	17.4-22.2	22.2-27	27-31.8	31.8-36.6	36.6-41.4
x_i	5.4	10.2	15	19.8	24.6	29.4	34.2	39
f_i	28	15	5	1	0	0	0	1
$x_i f_i$	151.2	153	75	19.8	0	0	0	39
$x_i - \bar{x}$	-2.7	2.1	6.9	11.7	16.5	21.3	26.1	30.9
$(x_i - \bar{X})^2 f_i$	204.12	66.15	238.05	136.89	0	0	0	954.81

$$Exact\ \bar{X} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{3+3+\cdots+37}{50} = \frac{405}{50} = 8.1,\ Exact\ \tilde{X} = \frac{1}{2}\left[x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right] = \frac{7+7}{2} = 7$$

$$Exact\ S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{n-1} = \frac{(3-8.1)^2 + (3-8.1)^2 + (3-8.1)^2 + \cdots + (37-8.1)^2}{49} = \frac{1774.5}{49} = 36.21$$

$$Estimated\ \bar{X} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{438}{50} = 8.76,\ Estimated\ \tilde{X} = \frac{1}{2}[x_{mid} + x_{mid+1}] = \frac{19.8+24.6}{2} = 22.2$$

$$Estimated\ S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2 f_i}{n-1} = \frac{204.12 + 66.15 + 238.05 + 136.89 + 954.81}{49} = 32.65$$

Exact value of $\bar{x}$ is greater than exact value of $\hat{x}$, but estimated value of $\bar{x}$ is smaller than estimated value of $\hat{x}$. our decision should be according to comparing exact values. so our histogram is skewed to the right or positively skewed

Chapter 6

Discrete and continuous distributions

Binomial distribution

A binomial random variable with parameters n and p represents the number of success with probability p independently. If X is such a random variable, then

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0,1,\dots,n$$

Expectation value and variance of binomial distribution:

If X is a discrete binomial random variable with parameters n and p, then

$$E(X) = np \qquad var(X) = np(1-p) = npq$$

Example: suppose you will be attending 6 hockey games. If each game independently will go to overtime with probability 0.1, find the probability that:

- At least one of the games will go to the overtime
- At most one of the games will go into overtime

$$p(x \geq 1) = 1 - p(x = 0) = 1 - \binom{6}{0} 0.1^0 (1-0.1)^{6-0} = 0.4686$$

$$p(x \leq 1) = p(x = 0) + p(x = 1) = \binom{6}{0} 0.1^0 (1-0.1)^{6-0} + \binom{6}{1} 0.1^1 (1-0.1)^{6-1}$$
$$= 0.5314 + 0.3543 = 0.8857$$

Example: A fair dice is to be rolled 20 times, find the expected value and variance of the number of times:

- 6 appears
- 5 or 6 appears
- An odd number appears

$$A = 6\ appears, B = 5\ or\ 6\ appears, C = An\ odd\ number\ appears$$

$$E(A) = np_A = 20 * {}^1/_6 = {}^{20}/_6, \quad V(A) = np_A q_A = 20 * {}^1/_6 * {}^5/_6 = {}^{100}/_{36}$$

$$E(B) = np_B = 20 * {}^2/_6 = {}^{40}/_6, \quad V(B) = np_B q_B = 20 * {}^2/_6 * {}^4/_6 = {}^{160}/_{36}$$

$$E(C) = np_C = 20 * {}^3/_6 = {}^{60}/_6, \quad V(C) = np_C q_C = 20 * {}^3/_6 * {}^3/_6 = {}^{180}/_{36}$$

Example: At a certain airport, 70% of the flights arrive on time. A sample of 10 flights is studied. Find

- Probability that exactly 8 flights were on time
- Probability that less than or equal to 8 flights were on time

$$p = 0.7, n = 10$$

$$p(x = 8) = \binom{10}{8} 0.7^8 (1-0.7)^{10-8} = 0.233$$

$$p(x \leq 8) = 1 - p(x > 8) = 1 - [p(x = 9) + p(x = 10)]$$
$$= 1 - \left[\binom{10}{9} 0.7^9 (1-0.7)^{10-9} + \binom{10}{10} 0.7^{10} (1-0.7)^{10-10}\right] = 1 - 0.149$$
$$= 0.851$$

Hypergeometric distribution

(Binomial distribution in which the trials are not ındependent)

The probability distribution of the hyper geometric random variable is the number of success in a random sample of size n selected from N items of which k are labeled success and N-k labeled failure.

If X is such a random variable, then

$$H(X,N,n,k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \qquad x = 0,1,\dots,k$$

Expectation value and variance of Hypergeometric distribution:

If X is a discrete hyper geometric random variable with parameters n and p, then

$$E(X) = \frac{nk}{N} \qquad var(X) = \left(\frac{N-n}{N-1}\right)\frac{nk}{N}\left(1-\frac{k}{N}\right) = (\frac{N-n}{N-1})npq$$

Example: Draw 6 cards from a deck without replacement. What is the probability of getting two hearts?

$X = \text{getting two hearts}, \qquad \mathrm{n = 6, k = 13, N = 52}, \;\; p(x=2) = \frac{\binom{13}{2}\binom{52-13}{6-2}}{\binom{52}{6}} = 0.31513$

Example: A crate contains 50 light bulbs of which 5 are defective and 45 are not. A Quality Control Inspector randomly samples 4 bulbs without replacement. Let *X* = the number of defective bulbs selected. Find the probability mass function, *f*(*x*), of the discrete random variable *X*.

$x = \text{the number of defective bulbs}, \qquad \mathrm{n = 4, k = 5, N = 50, x = 0, 1, 2, 3, 4}$

$p(x=0) = \frac{\binom{5}{0}\binom{50-5}{4-0}}{\binom{50}{4}} = 0.6469, \;\; p(x=1) = \frac{\binom{5}{1}\binom{50-5}{4-1}}{\binom{50}{4}} = 0.3081$

$p(x=2) = \frac{\binom{5}{2}\binom{50-5}{4-2}}{\binom{50}{4}} = 0.04298, \;\; p(x=3) = \frac{\binom{5}{3}\binom{50-5}{4-3}}{\binom{50}{4}} = 0.001954$

$p(x=4) = \frac{\binom{5}{4}\binom{50-5}{4-4}}{\binom{50}{4}} = 0.0000217, \qquad p(X=x) = \begin{cases} \frac{\binom{5}{x}\binom{45}{4-x}}{\binom{50}{4}} & x = 0, 1, 2, 3, 4 \\ 0 & otherwise \end{cases}$

Poisson distribution

A random variable X has Poisson distribution with parameterλ, if probability mass function of that random variable is:

$$p(X=x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x = 0,1,2,\dots$$

Expectation value and variance of Poisson distribution:

If X is a discrete Poisson random variable with parameter λ, then

$$E(X) = var(X) = \lambda$$

Example: If electricity power failures occur according to a Poisson distribution with an average of 3 failures every twenty weeks.

- calculate the probability that there will not be more than one failure during a particular week.
- Find expectation and variance of electricity power failures during a particular week

$\lambda = \frac{3}{20} = 0.15 \text{ failures every week}$

$p(X \le 1) = p(X = 0) + p(X = 1) = \frac{e^{-0.15}0.15^0}{0!} + \frac{e^{-0.15}0.15^1}{1!} = 0.98951$

$E(X) = var(X) = 0.15$

Note: Poisson random variable arise approximations to Binomial random variable, if the number of trials is large and probability of success is small ($\lambda = np$)

Example: the probability of producing a defective item is 0.1.

- What is the probability that a sample of 10 items will contain at most 1 defective item?
- What is the Poisson approximation for this probability?

$p = 0.1, n = 10$

$$p(x \le 1) = p(x = 0) + p(x = 1) = \left[\binom{10}{0} 0.1^0(1 - 0.1)^{10-0} + \binom{10}{1} 0.1^1(1 - 0.1)^{10-1}\right]$$
$$= 0.7361$$

$\lambda = np = 10 * 0.1 = 1$

$$p(x \le 1) = p(x = 0) + p(x = 1) = \left[\frac{e^{-1}1^0}{0!} + \frac{e^{-1}1^1}{1!}\right] = 0.7358$$

Geometric distribution

Suppose that repeated independent Bernoulli trials each one having probability of success P are to be performed. Let X be the number of trials needed until the first success occurs. We say that X follows the geometric probability distribution with parameter p .

Probability mass function of X

$p(X = x) = (1 - p)^{x-1}p \qquad x = 1,2, \ldots$

What are the Differences between the Geometric and the Binomial Distributions?

- The most obvious difference is that the Geometric Distribution does not have a set number of observations, n.
- The second most obvious is the question being asked:
  Binomial: asks for the probability of a certain number of success
  Geometric: asks for the probability of the first success

Expectation value and variance of Gmetric distribution:

If X is a discrete geometric random variable with parameters x and p, then

$$E(X) = \frac{1}{p} \qquad var(X) = \frac{1 - p}{p^2} = \frac{q}{p^2}$$

Example: If a production line has a 20% defective rate.

- What is the probability that first defective comes in third selection?
- What is the average number of inspections to obtain the first defective?

$p = 0.2$

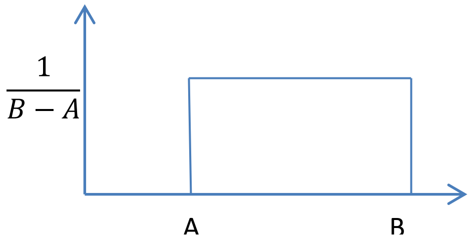
$p(X = x) = (1 - p)^{x-1}p \rightarrow p(X = 3) = (1 - 0.2)^{3-1}0.2 = 0.128$

$$E(X) = \frac{1}{p} = \frac{1}{0.2} = 5$$

Uniform distribution

The density function of continuous uniform random variable X on the interval $[A, B]$ is

$$f(X, A, B) = \begin{cases} \frac{1}{B - A} & A \le X \le B \\ 0 & otherwise \end{cases}$$



Expectation value and variance of Uniform distribution:

If X is a continuous uniform random variable in interval of A and B, then

$$E(X) = \frac{A + B}{2} \qquad var(X) = \frac{(B - A)^2}{12}$$

Example: Suppose that X is a Uniform R.V over the interval (0, 2). Find

- $p\left(X > {}^1/_3\right) =$
- $p(0.3 \le X \le 0.9) =$
- E(X) & Variance(X)

$$p\left(X > {}^1/_3\right) = \int_{{}^1/_3}^{2} \frac{1}{2} dx = \frac{1}{2} x \Big|_{{}^1/_3}^{2} = \frac{1}{2}\left(2 - \frac{1}{3}\right) = \frac{5}{6}$$

$$p(0.3 < X < 0.9) = \int_{0.3}^{0.9} \frac{1}{2} dx = \frac{1}{2} x \Big|_{0.3}^{0.9} = \frac{1}{2}(0.9 - 0.3) = 0.3$$

$$E(X) = \frac{A + B}{2} = \frac{0 + 2}{2} = 1 \qquad var(X) = \frac{(B - A)^2}{12} = \frac{(2 - 0)^2}{12} = \frac{1}{3}$$

Exponential distribution

The Random Variable X has an exponential distribution with parameter λ, if it's density function is given by

$$f(X) = \begin{cases} \lambda e^{-\lambda x} & x > 0, \lambda > 0 \\ 0 & otherwise \end{cases}$$

Expectation value and variance of Exponential distribution:

If X is a continuous Exponential random variable with parameters x and λ, then

$$E(X) = \frac{1}{\lambda} \qquad var(X) = \frac{1}{\lambda^2}$$

Example: If jobs arrive every 15 seconds on average, $\lambda = 4$ per minute,

- what is the probability of waiting less than or equal to 30 seconds

$$p(x \le 30) = \int_{0}^{30} 15e^{-15x}dx = \int_{0}^{0.5} 4e^{-4x}dx = -e^{-4x}\Big|_{0}^{0.5} = 1 - e^{2} = 0.86$$

Example: Accidents occur with a Poisson distribution at an average of 4 per week. i.e.λ= 4

- Calculate the probability of more than 5 accidents in any one week
- What is the probability that at least two weeks will pass between accident?

$$\begin{aligned} p(x > 5) = 1 - p(x \le 5) \\ &= 1 \\ &- [p(x=5) + p(x=4) + p(x=3) + p(x=2) + p(x=1) + p(x=0)] \\ &= 1 - \left[\frac{e^{-4}4^{5}}{5!} + \frac{e^{-4}4^{4}}{4!} + \frac{e^{-4}4^{3}}{3!} + \frac{e^{-4}4^{2}}{2!} + \frac{e^{-4}4^{1}}{1!} + \frac{e^{-4}4^{0}}{0!}\right] = 1 - 0.7851 \\ &= 0.2149 \end{aligned}$$

$$p(x > 2) = \int_{2}^{\infty} 4e^{-4x}dx = -e^{-4x}\Big|_{2}^{\infty} = \frac{1}{e^{8}} = 0.00034$$