7. Hypothesis Testing

Discrete Distribution

If a random variable can have only discrete outcomes, we have a discrete probability distribution

e.g A coin flip
No of people claiming insurance in a month

Types of Discrete Distribution

→ Binomial or Bernoulli
→ Negative Binomial
→ Geometric
→ Poisson
→ Hypergeometric

* Binomial Distribution

e.g. A coin flip.
Gender of babies delivered in a Hospital

→ Can have only 2 possible outcomes

→ There are no external factors influencing probability of each outcome over time

→ The chances of each outcome are independent of previous results

Toss of coin Once: Bernoulli Trial

Multiple Bernoulli Trial → Binomial Distribution

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

where

x = outcome
n = no of trials
p = probability of each trial success on

In Excel : BINOM.DIST()

* Hypergeometric Distribution

A hypergeometric distribution is generated when you have Bernoulli trials but selections are not replaced

$$P\{X=k\} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where $\binom{A}{B} = \frac{A!}{(A-B)!\,B!}$

In Excel : hypgeomdist()

e.g.

In a company, there are 18 employees; out of which 9 are women. We have to promote 8 of them with 3 being women. Find probability.

HYPGEOM.DIST(Sample_s, number_sample, population_s, number_pop, cumulative)

HYPGEOM.DIST(3, 8, 9, 18, False)

* Negative Binomial

→ Used to find out no of trials ~~needed~~ needed to get x successes.

e.g. What is the probability that the 30th purchase in my store will happen with the 100th customer when the probability of purchase for any customer is 20%?

NEGBINOM.DIST(number_f, number_s, probability_s, cumulative)

NEGBINOM.DIST(70, 30, 0.2, False)

Geometric Distribution

→ Probability of the first success in the n^{th} trial

e.g. Supposing there ~~are~~ is a defect rate of 2%, with some mechanical component being produced. What is the probability that a QC inspector will need to review at most 20 items before finding a defect?

Geometric Distribution is a special case of negative binomial distribution (In excel)

NEGBINOM.DIST (19, 1, 0.02, TRUE)

Poisson Distribution

→ Used to model no of events occuring in a time frame

e.g. no of ~~int~~ insurance claims in month

no of calls in an hour

disease spread in a day

Conditions to apply Poisson Distribution

1) Events have to be counted as whole numbers

2) Events are independent

3) Average frequency of occurence for the given time period is known

4) Number of events that have already occurred can be counted

$$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where λ = avg. no of occurences in a given interval of time

No impact of sample size "n"

Eg. You are a Manager in a call centre with a staff of 55 people, who on average handle 330 calls in an hour. A holiday is coming up and 5 resources want leave. You estimate the 50 remaining resources can manage 20% greater calls, but want to plan for the chance of greater than 20% of increased call volume.

What are the chances that number of calls on that day will go up by more than 20%?

$\lambda = \frac{330}{55} = 6$ calls an hour

20% of greater calls with 5 less resources

$= \frac{330 \times 1.2}{50} = 7.2 \cong 7$ calls in a hour

We need probability of seeing 8 or more calls a hour when average is 6.

Poisson.DIST (x, mean, Cumulative)

= 1 - Poisson.DIST (7, 6, True)

Binomial OR Poisson

1) If a mean/average probability of an event happening per unit time is given and you are asked to calculate probability of n events happening in a given time then the Poisson distribution is used.

2) If an exact probability of an event happening is given or implied in the question and you are asked to calculate the probability of this event happening k times out of n, then the Binomial Distribution must be used.

Binomial Distribution

describes the distribution of binary data from a finite sample. Thus it gives the probability of getting "r" events out of "n" trials.

Poisson Distribution

describes the distribution of of binary data from an infinite sample. Thus it gives the probability of getting "r" events in a population.

Continuous Distributions

Continuous distributions are applicable when an event can take on any value within a given range.

e.g. Height of person

Avg waiting time

Per Capita income

Normal distributions are most common kind of a continuous probability distribution due to its useful applications in statistics.

Normal Probability distribution

Properties:

1) Symmetric about the (single) mean

2) Mean = Median = Mode

3) The two tails extend indefinitely and never touch the axis

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

where y = vertical height of pt on normal dist.

x = distance along horizontal axis

σ = Sd

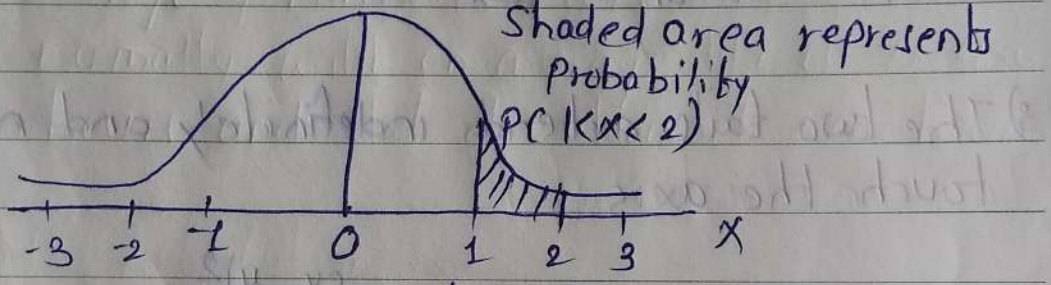
μ = mean

e = exponential constant = 2.71828...

π = 3.14159...

Area Under the Curve

The total area under a normal probability curve is always 1. This property allows us to think of the area as probability, and therefore we can compute probability two values on the curve

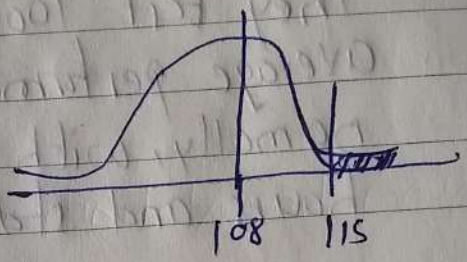


Normal Distribution

→ There is no as such a concept of point probability in continuous distribution

e.g. We test 100 students and find that IQ is normally distributed with an average of 108, with std deviation of 7.

Supposing you pick a random student from the 100, what are the chances he/she has an IQ > 115?



NORM.DIST (outcome, Mean, Stddev, cumulative)

= 1 - NORM.DIST (115, 108, 7, TRUE)

e.g. A manufacturer wants to state a guarantee for performance of their product, in hours, so that failure rates on the basis of the hours of performance are restricted to less than 5%.

They test 1000 samples, and find average performance hours to be distributed normally, with an average life of 71,450 hours and std deviation of 2700 hours

# Hours	P< hours.
71450	
70000	
69000	
68000	
67000	
66000	

Standard Normal Distribution

→ Special case with mean = 0 & std dev = 1

All normal distributions can be converted to std normal by following formula

$$Z = \frac{X - \mu}{\sigma}$$

where X = value of random variable

μ = mean

σ = std dev

$$Y_Z = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-Z^2}{2}}$$

where Y_Z = vertical height on the std normal dist

Computing Probabilities in R

Two Types of fns

1) density fn ⇒ cumulative = False

2) cumulative ⇒ cumulative = True

1) Binomial Distribution

Density fn

dbinom (no of success, no of trial, prob of success)

P(X=3) X ⇒ no of heads in a toss of coin 10 time

dbinom (3, 10, 0.5)

dbinom (1:10, 10, 0.5)

cumulative Probability fn
pbinom (no of success, no of trials, prob of success)

P(X=3) X ⇒ Upto 3 heads

pbinom (3, 10, 0.5)

pbinom (1:10, 10, 0.5)

2) Negative binomial Distribution

density fn

dnbinom (num_f, num_s, prob_s)

P(X=2) X ⇒ no of heads (2nd head occurs in 5th trial)

dnbinom (3, 2, 0.5)

Cumulative fn

#pnbinom (num_f, num_s, prob_s)

pnbinom (3, 2, 0.5)

3) Hypergeometric Distribution

Density fn success ↑ failure ↑

#dHyper (sample_s, pop_s, pop_f, sample_size)

#PC 2 red when 5 cards are drawn at random from a deck without replacement)

dhyper (2, 26, 26, 5)

cumulative fn

phyper (Sample_S, pop_S, pop_f, sample_size)

phyper (2, 26, 26, 5)

4) Poisson Distribution

Density fn

#dpois (x, mean)

P(x=4) mean=6), x=observed rate.

dpois (4, 6)

cumulative fn

ppois (x, mean)

ppois (4, 6)

5) Normal Distribution

#pnorm(x, mean, std dev)

#P(X <= 1.65)

pnorm (1.65, 0, 1)

---- ---- -XXXX- ----------

* Central Limit Theorem

As sample size grows sufficiently large, the sampling distribution of the means will tend towards a normal distribution (even if the underlying population is not normal)

Mathematically:

When we select simple random samples of size n, the distribution of these samples can be modeled means with a probability model that is

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow N(\text{Sample mean}, \text{Sample SD})$$

Implications

If sample size is sufficiently large (≥ 30), you can always use a normal distribuⁿ as your test distribution without worrying about true population distribution

Test statistics = t

$$= \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

where $\bar{X}$ → Sample mean
μ → Population mean
s → sd of sample
n → Sample size

Degrees of freedom = Sample size − 1

Hypothesis Testing in R

#t distribution

#P(t<=x)

Syntax: pt() → q = Probab } compulsary
df =
ncp =
lower.tail = } optional
log.p =

pt(1.65, 29)

q ⇒ t-stat value
df ⇒ degree of freedom

We take sample of 28 items, the sample mean is 30, sample sd = 5, the sample comes from a population whose mean is 35, what are the chances we will observe a sample mean of atmost 30?

$P(\bar{X} <= 30)$

$$t\text{-stat} = \frac{X - \mu}{\sigma/\sqrt{N}} = \frac{(30-35)}{\frac{5}{\sqrt{28}}} = -5.29$$

At $pt(-5.29, 27)$

t tests

Single sample t tests

create vector of data

set.seed(100)

X <- rnorm(16, 2, 1) #Data series of 16 points with mean = 2 & SD = 1

mean(x)

H_0 : Mean = 2 (Null)

H_1 : Mean > 2 (Alternate)

t.test(x, alternative = "greater", mu = 2)

H_0 : Mean = 2
H_1 : Mean not equal to 2

```
t.test(x, alternative = "two sided", mu=2)
```

Two Sample t-test

Independent Sample Test

Create Two Random Samples

```
set.seed(100)
x1 <- rnorm(20, 2, 1)
x2 <- rnorm(20, 3, 1.5)
mean(x1)
mean(x2)
```

H_0 : Mean 1 = Mean 2 (Mean diff is 0)

H_1 : Mean 1 ≠ Mean 2 are different

t.test (x_1, x_2, alternative = "two sided", mu=0)

Paired Sample

H_0 : Mean 1 = Mean 2 (Mean diff is 0)

H_1 : Mean 1 and Mean 2 are different

t.test (x_1, x_2, alternative = "two sided",
mu=0, paired = TRUE)