Binomial, normal distribution and sampling distribution

Iván Andrés Trujilllo Abella

ivantrujillo1229@gmail.com

Random variable, expected value and variance

Die game

Lab

Bernoulli Distribution

In a single trial with only two possible outcomes; $\textit{Success}(1)$ or $\textit{Failure}(0)$.

Roll a dice

Assume that you win if lands $\{3, 4, 5, 6\}$ or lost if lands $\{1, 2\}$. The probability of win is $\frac{4}{6}$ and lost is $\frac{2}{6}$.

Probability Mass Function (PMF)

X	$\mathbf{0}$	$\mathbf{1}$
$f(X = x)$	$(1 - P)$	P

$$P(X = x) = P^x (1 - P)^{1-x} \tag{1}$$

Mean and variance

Mean

$$\mu = E(X) = \sum f(x)x = P \tag{2}$$

Variance

$$\begin{aligned} \sigma^2 = Var(X) &= E(X^2) - (E[X])^2 \\ &= P - P^2 \\ &= P(1 - P) \end{aligned} \tag{3}$$

Binomial

Flipping a coin

- Random variable x will take the value 1 (success) when the coin land Tail.
- Each trial is independent
- The probability is constant

Problem

Determine the probability of get three successes ($k = 3$) in five trials ($n = 5$).

5 trials and 3 success...

5 trials and 3 success...

Success and failures	Probability
$E_1 E_2 E_3 F_4 F_5$	$p^3(1-p)^2$
$E_1 E_2 E_4 F_3 F_5$	$p^3(1-p)^2$
$E_1 E_2 E_5 F_3 F_4$	$p^3(1-p)^2$
$E_1 E_3 E_4 F_2 F_5$	$p^3(1-p)^2$
$E_1 E_3 E_5 F_2 F_4$	$p^3(1-p)^2$
$E_1 E_4 E_5 F_2 F_3$	$p^3(1-p)^2$
$E_2 E_3 E_4 F_1 F_5$	$p^3(1-p)^2$
$E_2 E_3 E_5 F_1 F_4$	$p^3(1-p)^2$
$E_2 E_4 E_5 F_1 F_3$	$p^3(1-p)^2$
$E_3 E_4 E_5 F_1 F_2$	$p^3(1-p)^2$

Binomial distribution

We must said that $X \sim B(k, n, p)$

Probability Mass Function (*PMF*)

$$P(X = x) = \binom{n}{x} P^x (1 - P)^{n-x} \tag{4}$$

Cumulative Distribution Function (*CDF*)

$$P(X \leq x) = \sum_{i=0}^{x} \binom{n}{x} P^x (1 - P)^{n-x} \tag{5}$$

Python

```python
from scipy.stats import binom
binom.pmf(successes, trials, P)
binom.cdf(succeses, trials, P)
```

Test extreme cases:

- $CDF(n, n, 0.5)$
- $PMF(1, 1, 0.5)$

Mean and variance sum of independent Bernoulli trials

Let $B_1, ..., B_n$ iid random variables with $\mu = P$ and $\sigma^2 = P(1-P)$ if think binomial as the sum of this variables:

Mean of binomial distribution

$$E(X) = \sum_{i=1}^{n} E(B_i) = nP. \tag{6}$$

Variance of bionomial distribution

$$Var(X) = \sum_{i=1}^{n} var(B_i) = nP(1-P) \tag{7}$$

Normal distribution

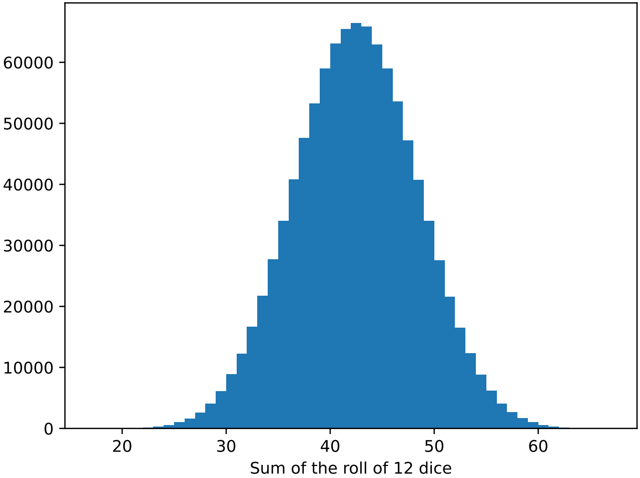
Is presented in some natural phenomenas or variables:

- Weight
- Height
- Math score

Properties:

- mean = median = mode
- Describe by its mean and standard deviation

Normal distribution



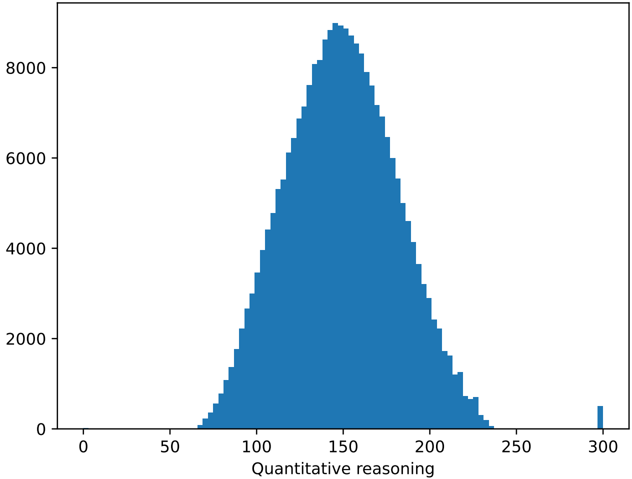
Normal distribution

...

$$f(x) = \frac{1}{2\sqrt{\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{8}$$

```
from scipy.stats import norm
norm.pdf(x, loc=mean, scale=std)
norm.cdf(x, loc=mean, scale=std)
```

Official data of universitaries students 2020

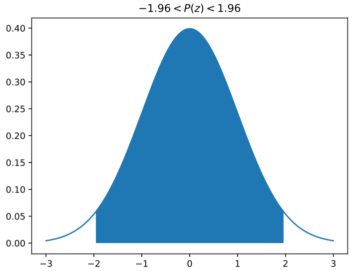


We can uses to compute more complex queries

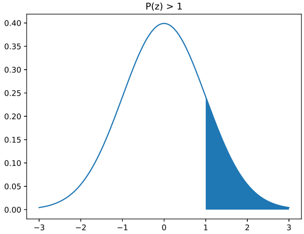
Standard deviations	1	2	3
Expected	0.6826	0.9544	0.9973
Observed	0.6662	0.9648	0.9979

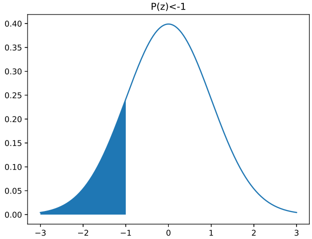
...

See lab



1. $CDF(x) - CDF(X)$
2. $1 - CDF(x)$
3. $CDF(X)$





Practical

...

- What is the probability
- What is the probability if $\bar{x} = 10$ and $\sigma^2 = 1$

Practical

...

- What is the probability
- What is the probability if $\bar{x} = 10$ and $\sigma^2 = 1$

Solutions

Quantile Function

Also known as Percent-Point function or inverse cumulative distribution function.
Here we pass the probability and pff give us the value...

...

$$Q(p) = F_x^{-1}(p) \quad p \in [0, 1] \tag{9}$$

```
from scipy.stats import norm
norm.ppf(quantile, loc=mean, scale=std)
```

Independent and identically distributed (iid)

Let $x_1, ..., x_n$ be a collection of random variables and $F_{X_i} = P(X_i \leq x_i)$ and $F_{X_1,...,X_n}(x_1, ...x_n) = P(X_1 \leq x1 \cap ... \cap X_n \leq x_n)$ accumulated join distribution then the variables

...

- Are identically distributed if

$$F_{X_1}(x) = F_{X_2}(x) = ... = F_{X_n}(x) \quad \forall x \tag{10}$$

- Are independent if

$$F_{X_1,...,X_n}(x_1, ...x_n) = F_{X_1}(x_1)...F_{X_n}(x_n) \tag{11}$$

Mean and variance of sampling mean

A set of random variables $x_1, ..., x_N$ drawn from certain distribution with mean μ and variance σ^2 finite, with sampling mean $\frac{\sum x_1}{n}$.

...

$$E(\bar{x}) = \mu \tag{12}$$

$$\begin{aligned} E(\bar{x}) &= E\left(\frac{\sum x_i}{n}\right) \\ &= \frac{1}{n}\left(\sum E(x_i)\right) \\ &= \frac{1}{n}\sum \mu = \mu. \end{aligned} \tag{13}$$

Sampling variance

...

$$sdv(\bar{x}) = \frac{\sigma}{\sqrt{n}} \tag{15}$$

(See lab)

$$\begin{aligned} var(\bar{x}) &= \frac{1}{n^2} \sum var(x_i) \\ &= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \\ sdv(\bar{x}) &= \frac{\sigma}{\sqrt{n}} \end{aligned}$$

Distribution of sample statistics

Each sample have different values, then statistics are random variables, but what distribution follow?.

Distribution of sample statistics

Each sample have different values, then statistics are random variables, but what distribution follow?.

Sampling of mean

- if $X \sim N(\mu, \sigma)$ then $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
- CLT states if n is large then $\bar{X}$ is approximately normal with $N(\mu, \frac{\sigma}{\sqrt{n}})$

Covergence in probabality

$$X \xrightarrow{P} X' \tag{16}$$

The Probability of X differ from X' tend to **zero** when $n \to \infty$

Covergence in Distribution

$$X \xrightarrow{d} X' \tag{17}$$

$$\lim_{n\to\infty} CDF_n(X) = CDF(X') \tag{18}$$

Law a large of numbers

Weak

Let $x_1, ..., x_n$ a succession $(\{x_n\})$ of random variables **iid** with mean $E(x_i) = \mu$ then:

...

$$\frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow[n\to\infty]{P} \mu \tag{19}$$

(See simulation)

Central limit theorem

$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ Suppose that we draw samples from a population and get the mean of each sample for instance:

$$\begin{aligned}
\bar{x}_1 &= \frac{1}{k}(x_1^1 + x_2^1 + ... + x_k^1) \\
\bar{x}_2 &= \frac{1}{k}(x_1^2 + x_2^2 + ... + x_k^2) \\
&\vdots \\
\bar{x}_j &= \frac{1}{k}(x_1^j + x_2^j + ... + x_k^j)
\end{aligned} \tag{20}$$

Thus $\bar{x}_j$ is the $j-th$ sample mean composed of k terms.

CLT

Let $x_1, ..., x_n$ a succession $(\{x_n\})$ of random variables **iid** with mean $E(x_i) = \mu$ and $var(x_i) = \sigma^2 < \infty$ (Finite) then:

...

$$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to\infty]{d} N(0,1) \tag{21}$$

CLT simulation

(See simulation)

How big is?

n?

30 is a practical value

$$n \geq 30 \quad (22)$$

...

- Population - Parameters
- Sample - Statistics

for instance the mean μ and sample mean $\bar{x}$. in some books σ^2 and S^2 for population and sample variance respectively.

Poisson Distribution

According to the former binomial distribution $X \sim b(p, n)$ the two parameter are the shape a form of the distribution. the poisson distribution is the case when the variable follow a binomial distribution with a $n \rightarrow \infty$

...

In the limit case, the occurrence of a only event is only guaranteed in the measure that the space is very small, for instance if the ocurrence of the events is simultaneous, you should not consider a Poisson distribution. the FD we can dervied of a binomial distribution in the following way $E(x) = np = \lambda$, thus:

$$\frac{n!}{(n-k)!k!}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k}$$

$$\frac{(n-k+1)!}{n^k k!}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-k}$$

$e = \lim_{x\to\infty}(1+\frac{1}{n})^n$we must use $t = \frac{n}{k}$, and thus $\frac{n+k}{n} = 1 + \frac{k}{n}$

$$\lim_{n\to\infty} = \frac{e^{-k}\lambda^k}{k!}$$

thus a ramdon variable follow a poisson distrbution with a paramter λ
$X \sim p(\lambda)$ and its FD is rewritten as:

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