Applied Statistical Methods

Continuous Distributions of One Variable

Xuemao Zhang
East Stroudsburg University

February 17, 2023

Outline

- Background: Probability Distributions
- Characterizing a Distribution
- Normal Distributions
  - Normal Distributions
  - Central Limit Theorem
- Continuous Distributions Derived from the Normal Distribution
  - Chi-Square Distribution
  - t-Distribution
  - F-Distribution
- Other Continuous Distributions
  - Lognormal Distribution
  - Weibull Distribution
  - Exponential Distribution
  - Uniform Distribution
- scipy.stats

Background: Probability Distributions

- **Distribution** is the description of data values and frequencies of a data set
- A random variable is a variable resulted from a random procedure. It takes numerical values only.
  - Flip a coin
  - Toss a die
  - Measure heights of a population
- **Probability distribution** is the distribution for a population. The population could be a population of individuals or a population of samples
- The probability distribution of a given statistic based on random samples of a same size is called sampling distributions.
  - For example, the sampling distribution of the sample mean is the distribution of all possible sample means, with all samples having the same sample size n taken from the same population.

Discrete Probability Distributions

- Discrete Random Variable: if the random variable can assume only a finite or countable number of values
- Discrete probability distribution lists all possible values of a discrete random variable X and the probability associated with each value x.
- The probability mass function (PMF)

$$P(x) = P(X = x)$$

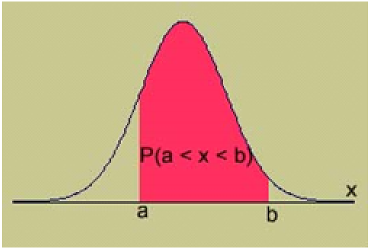
  must satisfy two conditions:

  - $0 \leq P(x) \leq 1$ for any x
  - $\sum_x P(x) = 1$

Continuous Probability Distributions

- Continuous Random Variable: if the random variable can assume infinitely many values corresponding to the points on a real line interval.
- Probability density function (PDF) $f(x) \geq 0$ describes the probability distribution of a continuous random variable.
- The PDF $f(x)$ must satisfy the following properties
  - $f(x) \geq 0$ for any $x \in R$
  - $\int\limits_{-\infty}^{\infty} f(x)dx = 1$

Continuous Probability Distributions



- $P(a \leq x \leq b) =$ area under the curve between a and b.
- There is no probability attached to any single value of x. That is, $P(x = a) = 0$.

Expected Value and Variance

- Let X be a discrete random variable with pmf $P(x)$. Then the expected value of X , denoted by $E(X)$ or μ, is defined to be

$$\mu = E(X) = \sum_x xP(x).$$

- Let X be a continuous random variable with pdf $f(x)$. Then the expected value of X , denoted by $E(X)$ or μ, is defined to be

$$\mu = E(X) = \int_{-\infty}^{\infty} tf(t)dt.$$

Expected Value and Variance

- Let X be a random variable with pmf or pdf $f(x)$. If g is a function, then

$$E[g(X)] = \sum_x g(x)f(x) \text{ or } E[g(X)] = \int_{-\infty}^{\infty} g(t)f(t)dt.$$

- Let X be a random variable. The variability is characterized by its variance.

$$\sigma^2 = Var(X) = E[(X - EX)^2] = E(X^2) - (EX)^2.$$

  - σ is called the standard deviation of X.

Characterizing a Distribution

- We deal with samples instead of populations in practical data analysis
- Let's use library `numpy` to characterize a data set -NumPy Reference https://numpy.org/doc/stable/reference/

```
import numpy as np
import pandas as pd
```

Characterizing a Distribution - Center

- **Mean**
  - $\bar{x} = \sum_{i=1}^{n} x_i / n$
  - Location parameter: for example $\mu = E(X)$

```
mtcars=pd.read_csv("../data/mtcars.csv")
np.mean(mtcars.mpg)
```

```
## 20.090625000000003
```

- **Median**

```
np.median(mtcars.mpg)
```

```
## 19.2
```

- **Mode** is the most frequently occuring value in a data set.
- The easiest way to find the mode of a data set is to use `scipy.stats`.
  - https://docs.scipy.org/doc/scipy/tutorial/stats.html

```
from scipy import stats
stats.mode(mtcars.mpg)
```

Characterizing a Distribution - Center

- In some situations the **geometric mean** can be useful to describe the location of a distribution.
- It is formula is

$$mean_{geometric} = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = exp\left(\frac{\sum_{i=1}^{n} ln(x_i)}{n}\right)$$

```
stats.gmean(mtcars.mpg)
```

```
## 19.25006404155361
```

Characterizing a Distribution - Spread

- **Range**: max-min
  - ptp stands for ‘peak-to-peak’

```
range=np.ptp(mtcars.mpg)
print(range)
#use Python built-in functions
```

```
## 23.5
```

```
max(mtcars.mpg)-min(mtcars.mpg)
```

```
## 23.5
```

- The **cumulative distribution function** or cdf of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = P(X \leq x) \text{ for all } x.$$

If X is continuous,

$$F_X(x) = \int_{-\infty}^{x} f(t)dt.$$

Characterizing a Distribution - Spread

- **Percentiles** are just the inverse of the CDF, and give the value below which a given percentage of the data values occur.
  - The 50th percentile is the median.
  - https://numpy.org/doc/stable/reference/generated/numpy.quantile.html

```
np.quantile(mtcars.mpg, q=[0.32, 0.50, 0.97])
```

```
## array([16.352, 19.2  , 32.505])
```

Characterizing a Distribution - Spread

- Sample **variance**
  - $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$
  - https://numpy.org/doc/stable/reference/generated/numpy.var.html
- Sample **standard deviation**
  - $s = \sqrt{s^2}$
  - https://numpy.org/doc/stable/reference/generated/numpy.std.html

```
np.var(mtcars.mpg, ddof=1)
```

```
## 36.32410282258064
```

```
np.std(mtcars.mpg, ddof=1)
```

```
## 6.026948052089104
```

- The **standard error** is the estimate of the standard deviation of a statistic when the statistics is considered as a random variable.
- For normally distributed data, the standard error (SE) of the sample mean $\bar{x}$ is $SE(\bar{x}) = \frac{s}{n}$.

Characterizing a Distribution - Spread

- In statistical analysis of a data set it is common to find the **confidence interval** of an unknown parameter
- For example, the $100(1-\alpha)\%$ of the mean parameter is

$$estimate \pm quantile_{1-\alpha/2} \cdot SE(estimate).$$

- Let's calculate a confidence interval using this formula
  - https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.t.html#scipy.stats.t

```
from scipy import stats
data=mtcars.mpg
stats.t.interval(confidence=0.95, df=len(data)-1,
loc=np.mean(data), scale=stats.sem(data))
#stats.sem compute standard error of the mean
```

```
## (17.91767850874625, 22.263571491253757)
```

Normal Distribution

A random variable Y is said to have a normal probability distribution if and only if, for $\sigma > 0$ and $-\infty < \mu < \infty$, the pdf of Y is

$$\boxed{f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, -\infty < y < \infty.}$$

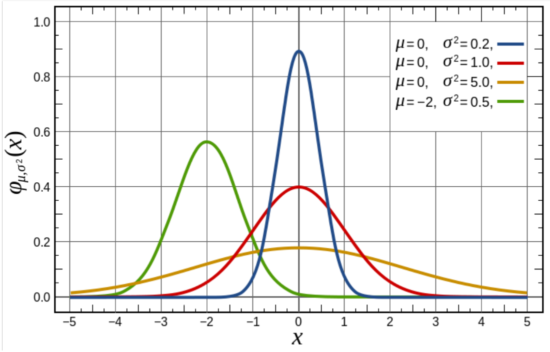
- If Y is a normally distributed random variable with parameters μ and σ, then

$$\boxed{E(Y) = \mu \text{ and } Var(Y) = \sigma^2.}$$

- Let $Y \sim N(\mu, \sigma^2)$. Then

$$\boxed{Z = \frac{Y - \mu}{\sigma} \sim N(0, 1).}$$

Normal Distribution



1. Mean $= \mu$; Standard deviation $= \sigma$.
2. Symmetric about $x = \mu$.
3. Total area under the curve is 1.

Normal Distribution - CLT

- The Central Limit Theorem: Let $X_1, \ldots, X_n$ be a sequence of iid random variables. Let $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Define $\overline{X}_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$. Let $G_n(x)$ denote the cdf of $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$. Then, for any $x, -\infty < x < \infty$,

$$\lim_{n\to\infty} P(G_n(x) \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

That is, $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ has a limiting standard normal distribution.

Normal Distribution - CLT

- A simulation study: Consider repeating this process: Roll a balanced-die 5 times. Find the sample mean. What do we know about the behavior of all sample means that are generated as this process continues indefinitely?
- Let X be the random variable of the results of rolling the die. How is the probability distribution of X.

Normal Distribution - CLT

- Python code
  - `numpy.random`: https://numpy.org/doc/1.16/reference/routines.random.html

```
import matplotlib.pyplot as plt
plt.style.use('classic')
import numpy as np
from numpy import random
iter = 1000
sample_means=[] #create a list to store sample means
n=5      #sample size
for i in range(iter):
  x=random.randint(1,7, size=n)
  sample_means.append(np.mean(x))

print(len(sample_means))

np.mean(sample_means)
plt.hist(sample_means)
plt.show()
```

Chi-square Distribution

- Let $Y_1, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then $Z_i = (Y_i - \mu)/\sigma$ are independent, standard normal random variables, $i = 1, 2, \ldots, n$, and

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

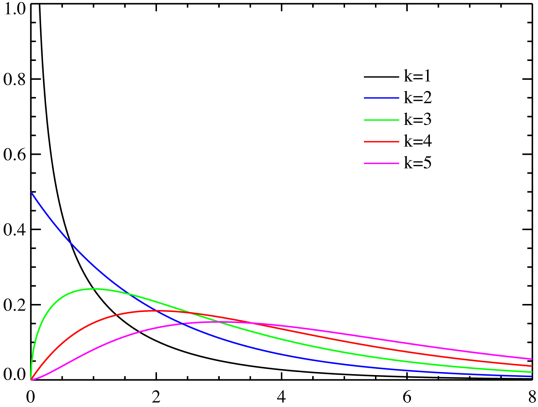
  has a χ^2 distribution with n degrees of freedom (df).

- Let $Y_1, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Let $S^2 = \dfrac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}$ be the sample variance. Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2}$$

  has a χ^2 distribution with $n-1$ degrees of freedom (df). Also, $\overline{Y}$ and S^2 are independent random variables.

Chi-square Distribution



1. The values of chi-square can be zero or positive, but it cannot be negative.
2. The chi-square distribution is not symmetric, unlike the Normal distributions. As the number of degrees of freedom increases, the distribution approaches a Normal distribution and thus becomes more symmetric.

Student's t-distribution

- t-distribution is proposed by W.S. Gosset in 1908. Due to Gosset's pseudonym "Student", it is known as "Student's t-distribution".
- Let Z be a standard normal random variable and let W be a χ^2-distributed variable with v df. If Z and W are independent, then

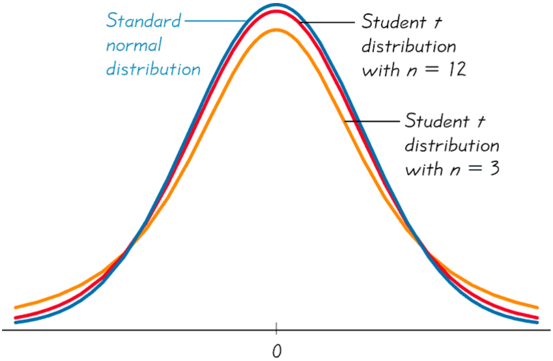
$$T = \frac{Z}{\sqrt{W/v}}$$

  is said to have a t-distribution with v df.
- Let $Y_1, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

  has Student's t-distribution with $n-1$ degrees of freedom.

Student's t-distribution



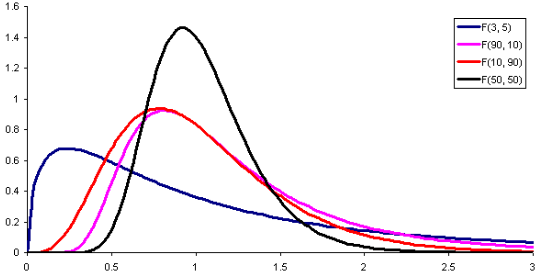
1. The density curves of the t-distribution look quite similar to the standard normal curve.
2. The spread of the t-distributions is a bit bigger than that of the standard normal curve.
3. As df gets bigger, the $t(df)$ density curve gets closer to the standard normal density curve.

F-Distribution

- Let W_1 and W_2 be independent χ^2-distributed random variables with v_1 and v_2 df, respectively. Then,
$$F = \frac{W_1/v_1}{W_2/v_2}$$
is said to have an F distribution with v_1 numerator degrees of freedom and v_2 denominator degrees of freedom.

- Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu_X, \sigma_X^2)$ population, and let $Y_1, \ldots, Y_m$ be a random sample from an independent $N(\mu_Y, \sigma_Y^2)$ population. Then
$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$
has an F-distribution with $n-1$ numerator degrees of freedom and $m-1$ denominator degrees of freedom.

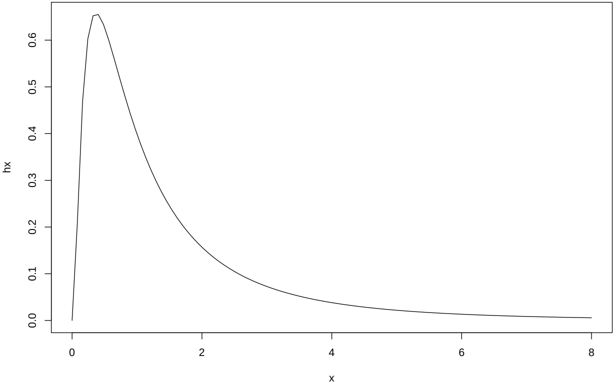
F-Distribution



1. The F distribution is not symmetric.
2. Values of the F distribution cannot be negative.
3. The exact shape of the F distribution depends on the two different dfs: Numerator df and Denominator df.

Lognormal Distribution

- In some circumstances a data set with a positive skewed distribution can be transformed into a symmetric normal distribution by taking logarithms.
- A lognormal (or log-normal) distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. That is, if X is lognormal, then $Y = \log X$ (log here is the natural log) is normal.

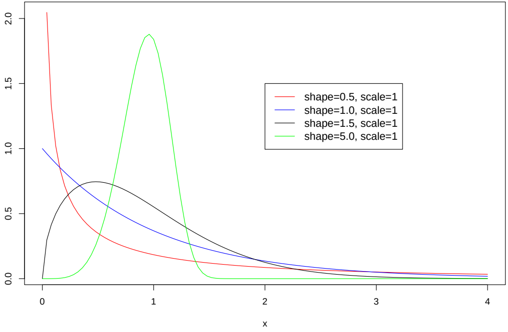


Weibull Distribution

- It has two parameters which allows it to handle increasing, decreasing, or constant failure-rates.

$$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

- $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter of the distribution.



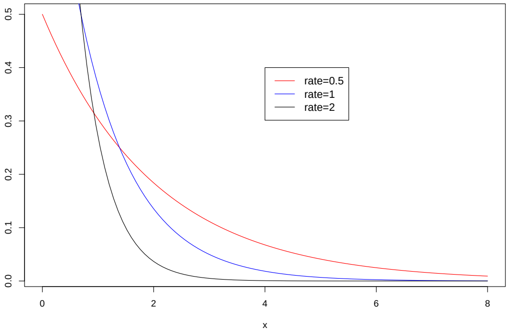
Weibull Distribution

- The Weibull distribution is the most commonly used distribution for modeling reliability data or "survival" data.
- If x is a "time-to-failure", the Weibull Distribution gives a distribution for which the failure rate is proportional to the power of time $k-1$.
- A value of $k<1$ indicates that the failure rate decreases over time.
- A value of $k=1$ indicates that the failure rate is constant over time.
- A value of $k>1$ indicates that the failure rate increases over time.

Exponential Distribution

- The pdf of an exponential distribution is given by

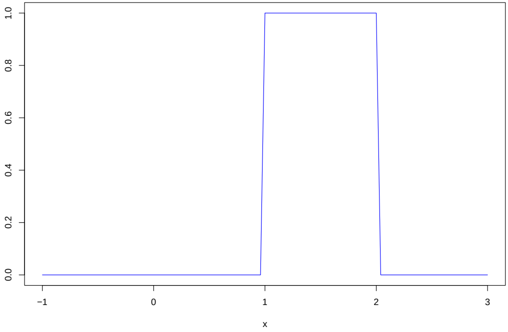
$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$



Uniform Distribution

- Uniform distribution: an even probability for all data values. It is not common for real data.

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & otherwise \end{cases}$$



scipy.stats

- `scipy.stats` can be used for random number generation, density, probability and quantile calculations.
- Read the tutorial
  https://docs.scipy.org/doc/scipy/tutorial/stats.html

License