Stat Definitions

Sreejith Sreekumar

December 30, 2021

Contents

1 Confidence Interval

- A confidence interval gives the PROBABILITY that our true value lies within the range of values. Bigger interval = higher probability

2 Probability

- Area under an interval of a distribution curve.

3 Likelihood vs Probability

- Answer from Cross Validated Likelihood: What is the best values of the parameters so that the data that we observed follows a <some> distribution? Probability: Assuming that the data comes from a certain distribution what is the chance. Probability is the area under the PDF curve

4 Percentage of normal distribution lies within 1 std of mean? 2, 3 std?

- 68%, 95%, 99.7%

5 SGD Update Rule

$$\theta = \theta - \alpha \Delta J(\theta)$$

$$(Current\ \theta\ vector) - learning\ rate * (Gradient\ of\ Slope)$$

6 Probability and Statistics

The problems considered by probability and statistics are inverse to each other. In probability theory we consider some underlying process which has some randomness or uncertainty modeled by random variables, and we figure out what happens. In statistics we observe something that has happened, and try to figure out what underlying process would explain those observations.

7 Law of Large Numbers

If you repeat an experiment independently a large number of times and average the result, what you obtain should be close to the expected value

8 Central Limit Theorem

9 Type I and Type II Error

Type 1: False Positive, Type 2: False Negative

10 Inverse Document Frequency

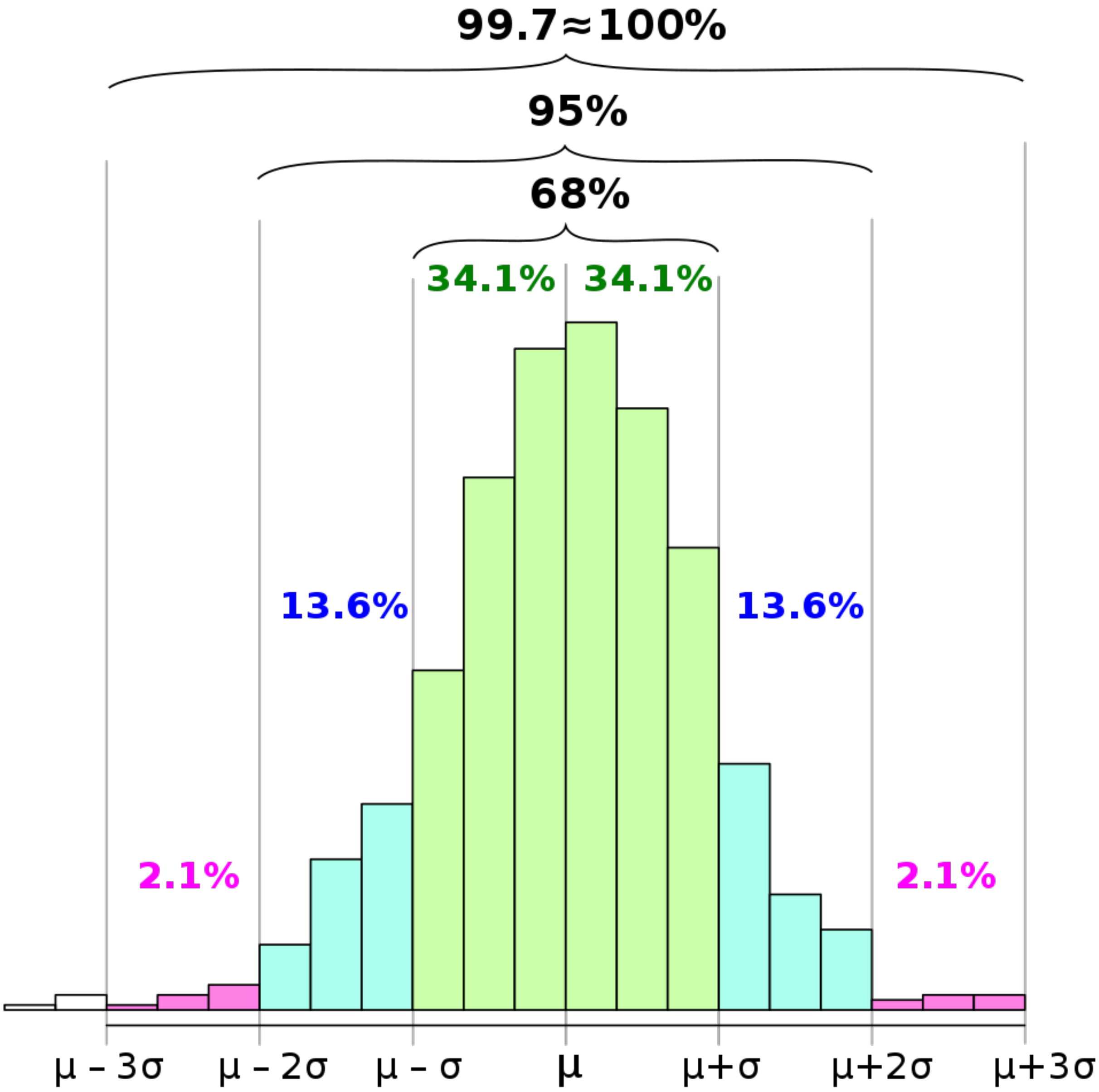
$$idf = log\frac{|D|}{d : ti \in d}$$

where | D | is the number of documents in our corpus, and | {d : ti ∈ d} | is the number of documents in which the term appears.

11 Kolmogorov - Smirnov Test

Tests whether sample fits a distribution well.

12 Normal Distribution Standard Deviations



13 Confidence Interval

There are 100 products and 25 of them are bad. What is the confidence interval?

p = 25/100 = 0.25

CI = 0.25 +/- 1.96 sqrt((0.25(1-0.25)) * 100)

CI = p +/- Z * sqrt(variance of binom dist)

CI = (16.5,33.5)

95% confidence = plus or minus 1.96 STDEV

13.1 Margin of Error

Margin of Error = $t * \frac{s}{\sqrt{n}}$

14 Isolation Forest

Creates splits like a random forest, but find out how difficult is it to isolate the path to split an instance. Answer: Quora

15 T-Test (Two Sample)

- Both has to be normal distributions
- Both needs to have similar variance
- σ is unknown
- Smaller sample sizes (Ideally in the range of 20-30) or else we use **z-test**
- $\frac{\bar{x1}-\bar{x2}}{\sqrt{\frac{s1^2}{n1}-\frac{s2^2}{n2}}}$

16 T-Test and Z-Test for one sample

- T-test is used when the sample size is less than 30
- t-Test = $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$
- From CLT, as n increased sample sd will be similar to population sd
  z-test = $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$
- when sample size is $>$ 30 and σ is available, use **Z-test**, else use **T-Test**

17 Chi-Squred Test

If there is a statistically significant difference in the observed vs expected counts.

$\chi 2 = \Sigma \frac{(O_i - E_i)^2}{E_i}$

Example:

Coin tossed 50 times.

Expected: 25H 25T
Observed: 28H 22T

$\frac{(28-25)^2}{25} + \frac{(22-25)^2}{25}$

$= \frac{9}{25} + \frac{9}{25}$

$= \frac{18}{25}$

$= 0.72$

H0: There is no statistically significant difference between observed values and expected values.
For a critical value of 0.05, and degree of freedom (n-1 = 1) $\chi^2 = 3.84$ Since this value is greater than 0.72, we accept Null Hypothesis

18 Probability Distributions

Mathematical Function that gives the probabilities of occurrence of different possible outcomes for an experiment.

18.1 Binomial:

Coin toss event repeated n times, with probability p of success.

- $nCr.P^r.(1-P)^{n-r}$
- Discrete with parameters (n, p)
- n independent experiments
- Success p
- Failure (1-p)
- Mean: np,
- Median: $\lfloor np \rfloor$, $\lceil np \rceil$
- Expected Value: np

18.2 Exponential

- Exponential distribution is often concerned with the amount of time until some specific event occurs.
- $f(x) = me^{-mx}$
- Decay parameter, $m = \frac{1}{\mu}$
- $\mu = \sigma$
- Expected value = p

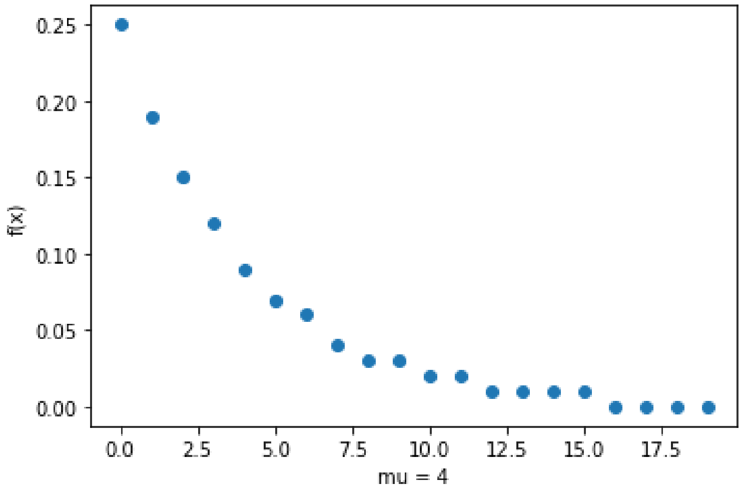
```
import matplotlib.pyplot as plt
import numpy as np

def get_ex(m, x):
    return round(m * np.power(np.e, -m*x), 2)
```

```
points = [(x, get_ex(0.25, x)) for x in list(range(20))]

xs = [x[0] for x in points]
ys = [y[1] for y in points]

f, ax = plt.subplots()
ax.scatter(xs, ys)
ax.set_ylabel("f(x)")
ax.set_xlabel("mu = 4")
plt.show()
```



19 Maximum Likelihood Estimation

Location which maximizes the likelihood of the data we measured.

Likelihood given all data points
L(μ, σ| x1, x2,. . . ,xn) = L(μ, σ| x1) * L(μ, σ| x2). . .

- Take derivative wrt μ , and treat σ like it is constant; and equate it to 0
- Take derivate wrt σ, and treat μ like it is a constant
- Take ln
- MLE calculation of Normal Distribution:

  Derivation

20 EM Algorithms

Goal: $\theta_{MLE} \in argmax_\theta P_\theta(x)$ Problem: $P_\theta(x) = \sum_x P_\theta(x)$ is difficult to maximize

- Start with a random estimate (and random parameters)
- Observe the data
- Adjust the parameters to some data so that the estimates of the parameter are better
- Observe more data
- Adjust . . .

21 Multicollinearity

- Measured by VIF
- Regress one variable with other variables: $\frac{1}{1-R^2}$
- VIFs are calculated by taking a predictor, and regressing it against every other predictor in the model.
- Useful Link: `https://etav.github.io/python/vif_factor_python.html`
- Any variable with VIF $>$ 5 must be removed

22 Normalization and Standardization

- Both are scaling techniques
- Normalization = Min-Max scaling
- Standardization = (x' - μ)/σ

23 Distances

- ${}_{i=1}{}^{n}(x_i - y_i)$