Introduction to Statistical Modeling
MSDS 598

Probability and Statistics
(Quick Overview)

Michael Ruddy

From Last Time

- Probability: "The likelihood that an event will occur"
- An **experiment** is an activity or procedure that produces distinct, well-defined possibilities called **outcomes.**
- The set of all possible outcomes is called the **sample space.**
- **Empirical Probability**: Probability estimate obtained by running the experiment many times.
  - Flip a coin many times
  - Historical weather patterns

$$\boxed{P = \frac{\text{\# of times outcome occurred}}{\text{\# of times experiment performed}}}$$

- A **random variable** is a *function* that assigns a numerical value to each outcome in the sample space

Uniform Random Variable

Pick a random number (discrete)

- Experiment: Pick a random number between one and ten
- Random Variable X: Value of that number
    - (This is a *discrete* random variable)

- Each value of the random variable is equally likely, meaning that X is **ormly distributed random variable**.

- What does the probability density function look like for X here?

Continuous Random Variable

Pick a random number (continuous)

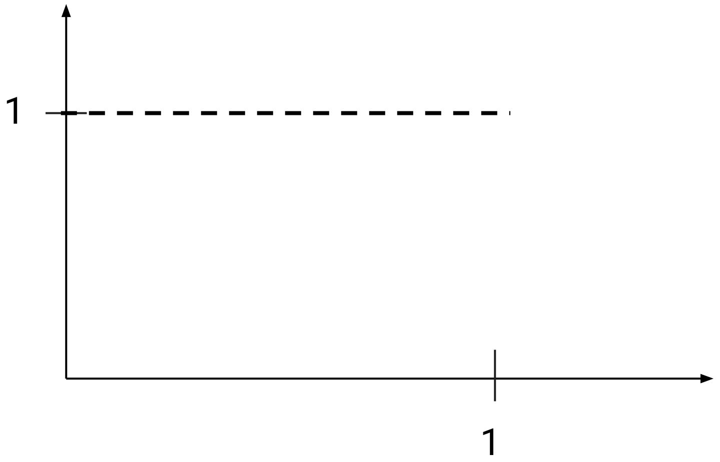
- Experiment: Pick a random real number between 0 and 1
- Random Variable X: Value of that number
    - (This is a *continuous* random variable)

- What is the probability that $X = 0.5$? $P(X = 0.5)$
- What is the probability that X is between 0 and 1? $P(0 < X < 1)$

Continuous Random Variable

Pick a random number (continuous)

- Experiment: Pick a random real number between 0 and 1
- Random Variable X: Value of that number
  - (This is a *continuous* random variable)



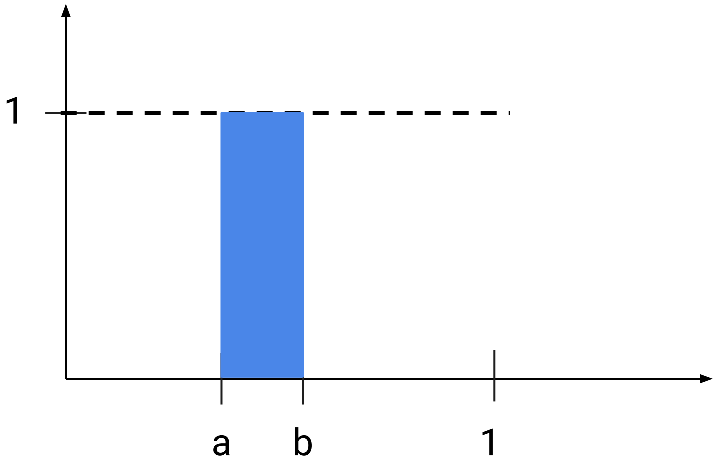
Probability density function

$$f_X(x) = 1$$

Continuous Random Variable

Pick a random number (continuous)

- Experiment: Pick a random real number between 0 and 1
- Random Variable X: Value of that number
    - (This is a *continuous* random variable)



Probability density function

$$f_X(x) = 1$$

$$\int_a^b f_X(x)dx = P(a < X < b)$$

"Area under the curve from *a* to *b*"

The **Mean** or **Expected Value**

Discrete

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

Continuous

$$\mu = \int_x x f_X(x) dx$$

- "Weighted sum of the values of the random variable"

- Recall the experiment **Pick a random number (discrete)**
  - What is the mean of the random variable X here?
  - How this relate to the "colloquial" understanding of the average?
  - What happens to the mean if we change the distribution so that numbers higher than 5 are more likely to be chosen?

The **Variance**

Discrete

$$\sigma^2 = \mathrm{var}X = \sum_x (x-\mu)^2 P(X=x)$$

Continuous

$$\sigma^2 = \mathrm{var}X = \int_x (x-\mu)^2 f(x)dx$$

- "How far values tend to stray from the mean"
- The **Standard Deviation** is the square root of the variance
    - Same units as the mean

- Is the variance of temperature in San Francisco or in Sacramento higher?

Estimating Mean for Continuous Random Variables

- We usually don’t have an explicit Probability Density Function for a continuous random variable.
  - We estimated it earlier!

N random sample of random variable X

Sample Mean $\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$

$$\{x_1, x_2, \ldots, x_N\}$$

Estimating Mean for Continuous Random Variables

- We usually don’t have an explicit Probability Density Function for a continuous random variable.
    - We estimated it earlier!

N random sample of random variable X

Sample Mean $\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$ $\leftarrow$ $\{x_1, x_2, \ldots, x_N\}$

Law of Large Numbers

$$\bar{X} \to \mu \text{ as } N \to \infty$$

Estimating Variance for Continuous Random Variables

Sample Variance: $$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{N-1}$$

Sample Mean

N random sample of random variable X

$$\{x_1, x_2, \ldots, x_N\}$$

Estimating Variance for Continuous Random Variables

Sample Variance: $$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{N-1}$$

Sample Mean

N random sample of random variable X

$$\{x_1, x_2, \ldots, x_N\}$$

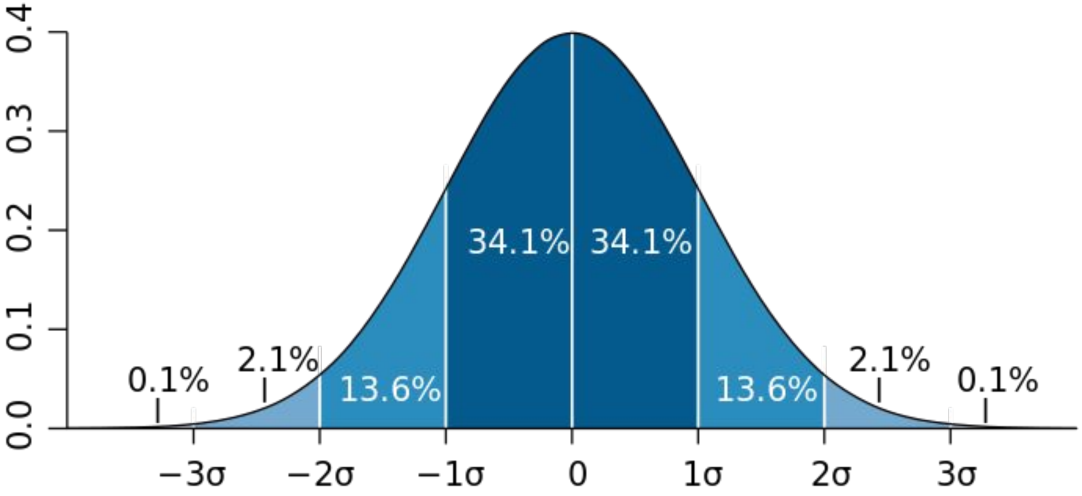
- The (*N*-1) comes from the fact that we using the sample mean rather than the population mean to estimate the variance.
- The *variance of the sample* uses *N* instead.

$$s^2 \to \sigma^2 \text{ as } N \to \infty$$

Normal Distribution

- Many random variables are distributed *normally* meaning they follow the **Normal Distribution** according to a mean and standard deviation

$$N(\mu, \sigma) \sim X$$



Central Limit Theorem

Sample Mean

Random Variable $\bar{X}_N$: Take the mean of a sample of size N

Central Limit Theorem

- The distribution of this random variable tends towards a normal distribution as N increases.
- **<u>Many say</u>** that it often looks very normal by N=30.

- Doesn’t matter what the distribution of the original population is!

Hypothesis Testing

- Null Hypothesis H_0
  - Assumed Fact/Default Position
- Alternative Hypothesis H_1/H_a
  - The opposite, or negation of the null hypothesis

- When we do Hypothesis testing we either
  - Reject the null hypothesis
  - Fail to reject the null hypothesis

- We can never conclude the Null Hypothesis.
- We can only say, “there is not enough evidence to reject”!

Hypothesis Testing

- Null Hypothesis H_0
    - Assumed Fact/Default Position
- Alternative Hypothesis H_1/H_a
    - The opposite, or negation of the null hypothesis

- When we do Hypothesis testing we either
    - Reject the null hypothesis
    - Fail to reject the null hypothesis

In theoretical American criminal justice a defendant is “innocent until proven guilty”

- Reject the null hypothesis (defendant is guilty)
- Fault to reject the null hypothesis (defendant is not guilty)

Hypothesis Testing

- Null Hypothesis H_0
  - Assumed Fact/Default Position
- Alternative Hypothesis H_1 / H_a
  - The opposite, or negation of the null hypothesis

Steps

1. Set a *significance level*, α
2. Run the experiment
3. Determine probability that experiment outcome occurred *given* H_0
4. If this is less than the significance level reject, otherwise fail to reject

Confidence Intervals

- Consider the random variable where μ and n are fixed.

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

- This calculates the Z-score for a hypothesis test (when the population variance is not known) and follows a distribution known as the "Student's t-distribution."

- What is the "Fail to Reject" interval around the sample mean?
    - Suppose we choose a level of significance α

$$\left[\bar{X} - c\frac{s}{\sqrt{n}}, \bar{X} + c\frac{s}{\sqrt{n}}\right]$$

$$P(-c \leq T \leq c) = 1 - \alpha$$

Confidence Intervals

- In this case μ is in the confidence interval if and only if the *p*-value is less than the level of significance (this is not always the case with Confidence Intervals!)

- If we were to take samples and calculate a confidence interval for each one, then 95% of the time we would contain the population mean.
- Everytime we construct a confidence interval there is a 95% chance it will contain the population mean.
  - **NOT** for a *given* interval there is 95% chance it contains the mean.

$$\left[\bar{X} - c\frac{s}{\sqrt{n}}, \bar{X} + c\frac{s}{\sqrt{n}}\right] \qquad P(-c \leq T \leq c) = 1 - \alpha$$