

# Introduction to Probability and Inference

HSSP Summer 2017, Instructor: Alexandra Ding

July 19, 2017

Please fill out the attendance sheet!

**Suggestions Box:** Feedback and suggestions are important to the success of this class and my experience as a teacher, so please send comments to alexawding@gmail.com!

## 2 Lecture 2 Recap: Random Variables and Distributions

- **Random Variable:** maps the outcomes in the sample space to the real line
- Random Variables can be **Continuous** (Height, Car mileage) or **Discrete** (coin toss, die roll, number of rain drops falling on my car)
- **Probability Mass Function (PMF):** for a **discrete** RV  $X$ , the PMF, denoted  $f_X(x)$  or  $f(x)$  tells us what the probability that our Random Variable equals some value. In other words:

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

- **Cumulative Distribution Function (CDF):** the distribution of a discrete RV can also be described via the CDF. The CDF answers the question of "what is the probability that my RV is less than some  $x$ ?" CDF for a RV  $X$  is denoted as  $F_X$ , and is basically a cumulative sum of PMFs.

$$F_X(x) = P(X \leq x)$$

- **Bernoulli Distribution:** a discrete distribution. It has one **parameter**,  $p$ . A Bernoulli distributed RV has only two possible outcomes (1 or 0) and is 1 with probability  $p$ , and 0 with probability  $1 - p$ .

$Y \sim \text{Bern}(p)$   $Y$  is distributed Bernoulli with parameter  $p$

$$f_Y(y = 1) = p \text{ Bernoulli PMF}$$

- **Binomial Distribution:** a discrete distribution with two parameters  $n$  and  $p$ . The Binomial reflects the number of successes in  $n$  independent trials, where the probability of success on each trial is  $p$ . The Binomial RV is the sum of independent Bernoulli RVs!

$X \sim \text{Bin}(n, p)$   $X$  is distributed Binomial with some  $n$  and  $p$

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

- **Probability Density Function (PDF):** for a continuous random variable, the PDF gives the distribution. This function gives the *likelihood or density* of observing some value of a RV (not the probability). We usually have to evaluate integrals to find the probability that a RV takes on a certain interval of values.
- **Normal Distribution:** a VERY special continuous distribution. If  $X$  is a RV and is distributed Normal, the notation is:

$$X \sim N(\mu, \sigma^2)$$

Where  $\mu$  is the mean and  $\sigma^2$  is the variance (a measure of spread). The density function is

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

## 2.1 Warmup Puzzles

### 1. Act Normal!:

Recall that the **Empirical Rule** (also known as the 68, 95, 99.7 Rule) reflects the probability that an observation of a RV falls within 1, 2 and 3 standard deviations from the mean. Suppose that blood glucose in a patient population is distributed Normally with mean 15 and variance 4. In other words:

$$X \sim N(\mu = 15, \sigma^2 = 4)$$

What is the probability that blood glucose is between 13 and 17? What is the probability that blood glucose is between 11 and 15?

### 2. Z Scores: A Z score tells you how many standard deviations $\sigma$ your observation is from the mean, $\mu$ .

$$\text{Z Score} = \frac{|\text{observation} - \mu|}{\sigma}$$

Suppose you, a sabermetrician (baseball statistician) have measured the number of Home Runs that every team in Major League Baseball has hit in 2017, and find that the Mean is 115, with a standard deviation of 18. The Boston Red Sox have hit 94 home runs.<sup>1</sup> Calculate the Z Score of this observation.

### 3. Simpson's Paradox: You're still a sabermetrician and are at home looking at batting averages of different players. A batting average is the number of safe hits divided by the number of total at-bats. Suppose you're looking at the batting averages of two players, Derek Jeter and David Justice, in the years 1995 and 1996, as well as in both years combined.<sup>2</sup> Here's what you observe:

	1995	1996	Combined
Jeter	0.250	0.314	0.310
Justice	0.253	0.321	0.270

Who has the higher average in 1995? Who has the higher average in 1996? Who has the higher combined average? Does this strike you as strange?

<sup>1</sup>SOURCE: [http://www.espn.com/mlb/stats/team/\\_/stat/batting](http://www.espn.com/mlb/stats/team/_/stat/batting)

<sup>2</sup>[https://en.wikipedia.org/wiki/Simpson%27s\\_paradox](https://en.wikipedia.org/wiki/Simpson%27s_paradox)

# 3 Lecture 3: Introduction to Inference

## 3.1 Inference: Into the Wild

**Motivation:** *When we collect data, we are usually interested in estimating real-world quantities and answering relevant questions.*

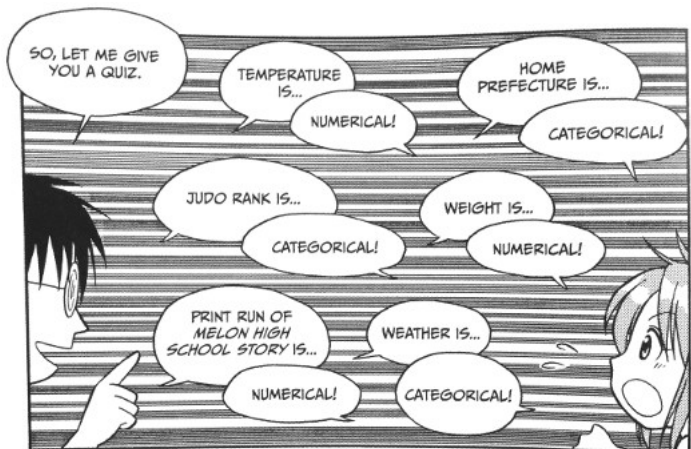
- In **Statistical Inference**, we use observable data to make a statement/decision about a statistical model.
- Email Marketing Campaign- What is the estimated Bounce Rate of email type A vs. type B?
- Disease Incidence- Has the incidence of cholera in Pakistan changed over the past 5 years?
- Transportation- What is the average wait time of passengers at Downtown Crossing?
- Disease Treatment- What is the 5-year survival rate of patients on drug vs. control?
- **Survey Sampling:** obtaining information about a larger population by examining a small fraction of observations. For a **population** of size  $N$ , measuring some characteristics of a subset of  $n < N$ .
- **Simple Random Sample:** Out of a population of  $N$ , each sample of size  $n$  is equally likely to be selected. *How many unique samples are there?*



- Different sampling techniques exist. More on this in Expt. Design

## 3.2 Random Variables and Measurement

- Consider each measurement in our sample to be a **Random Variable** (Capitalized). Let  $X_i$  be a RV denoting the height of the  $i$ th person in our sample. The **Observed** value of height is  $x_i$  (lowercase).
- *Notation: Random Variables are CAPITALIZED. Observations are LOWERCASE.*
- Suppose we are interested in measuring the height in inches of  $n$  randomly selected people, denoted as  $X_1, X_2, \dots, X_n$ . Then we might observe that  $x_1 = 64, x_2 = 75$  etc.
- **Categorical vs. Quantitative/Numerical Data (Variables):** *Examples?*



3

<sup>3</sup>from The Manga Guide to Statistics

### 3.3 Statistics summarize data

- **Statistic:** Numerical summary of observed variables (data). Formally, a real-valued function of Random Variables. Statistics are also RVs.
- Statistics can be calculated on the whole **population** or on a **sample** of the population.
- Since the goal of inference is to validate statistical models, statistics are used to approximate parameters of statistical models
- **Parameter:** Generally notated as  $\theta$ . Summary of a statistical model or of a population. Determines the shape of a distribution for a set of RVs of interest. Recall the parameters in

$$X \sim \text{Bin}(n, p) \text{ and } Y \sim N(\mu, \sigma^2)$$

- **Estimator:** A type of statistic that aims to estimate a model or population parameter.
- Examples of Population and Sample statistics (most common)

Statistic	Population	Sample
Mean	$\mu$	$\bar{X}$
Proportion	$p$	$\hat{p}$
Variance	$\sigma^2$	$S^2$
Standard Deviation	$\sigma$	$S$

- **Independence and Distribution Assumption:** We assume that members of the same population are IID (independent and identically distributed). And if our sample is random, the members of our sample are also assumed to be IID. Note that IID does not always hold! (see example below)
- **Random Variables X and Y are independent if**

$$P(X = x, Y = y) = P(X = x)P(Y = y) \text{ for Discrete RV X and Y}$$

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \text{ for both Discrete and Continuous RV X and Y}$$

- **Independence Example** <sup>4</sup>: Suppose you are an investment banker at Goldman Sachs and are evaluating a Collateralized Debt Obligation (CDO). A (simplified explanation of a) CDO is a "bet" taken on a "pool" of assets, for example, mortgages. In this case, imagine that you have a pool of 5 mortgages, each with an estimated 5% chance of defaulting (failing). You have a few options for the types of "bets" you can take:

- Bet Alpha: pays out K cash unless all five of the mortgages default
- Bet Epsilon: pays out K cash unless any one of the mortgages default

Which type of bet has greater risk? What is the probability of losing?

<sup>4</sup>From Nate Silver's "The Signal and the Noise: Why so many predictions fail, but some don't", 2012

### 3.4 Population Statistics

- Suppose we have a population of size  $N$ . Population Statistics are computed on the measurements of **all members of the population**.
- **Population Mean:** reflects the center of the dataset.

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

- **Population Variance:** reflects how spread out the datapoints are around the population mean,  $\mu$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

- **Population Proportion:** For observations of categorical  $Y_i$ , reflects the proportion of positive observations.

$$p = \frac{\sum_{i=1}^N y_i}{N}$$

- Suppose you, a guidance counselor, are interested in the average GPA of all 300 students in your class year. You have access to all of these measurements. What population parameter would you calculate?

- Suppose that in the population of sea turtles on this beach, 900 eggs have been laid, and 88 hatch. What is the population proportion of eggs that have hatched?

### 3.5 Sample Statistics and Standard Error

- **Population statistics are often unknown or difficult to measure.** Thus, taking a sample of the population and calculating statistics on this is often useful.
- For population of size  $N$ , take sample of size  $n$ . A good rule of thumb is that  $n > 30$ .
- What makes a useful estimator of a population parameter? **Unbiasedness and Consistency.**
- **Sample Mean:** reflects the center of the measurements  $x_i$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

Sample mean is a good estimator of  $\mu$  because  $E(\bar{X}) = \mu$ .

- *Intuition:* in the long run, we expect sample mean to be centered around pop mean = **Unbiasedness!** As we take more samples (as  $n$  gets larger), the **standard error** of our estimate will decrease (on average sample mean gets closer to pop mean) = **Consistency.**

- **Sample Proportion:** reflects the proportion of positive measurements in our sample (for a categorical variable)

$$p = \frac{\sum_{i=1}^n y_i}{n}$$

- **Sample Variance:** reflects how spread out the datapoints are around the sample mean,  $\bar{X}$ .

$$S^2 = \frac{\sum (x_i - \bar{X})^2}{n - 1}$$

*Compare with population proportion, and take a closer look at that denominator!*

- **Standard Error:** Standard Error is similar to a "measurement error". Each sample of  $n$  from the population of  $N$  will be slightly different, and thus you'll have different values of your sample statistic. In other words, Standard Error is the Error in our estimate of the parameter, NOT the spread of the parameter itself!
- Standard Error decreases with sample size, and depends on the estimator you're using.

### 3.6 Practice with Sample Statistics

1. **Jersey Shore Infrastructure:** Suppose you, a transportation official, are interested in estimating the average number of cars traversing bridges in New Jersey at 2pm on a Tuesday. Of NJ's 6500 bridges, you get your employees to take a sample of 10 and count the number of cars that pass within 30 minutes.

60, 40, 30, 35, 50, 10, 30, 15, 20, 65

- (a) Calculate the sample mean, median and mode
- (b) Calculate the sample variance

2. **AB Testing:** You are working for a marketing analytics company, and your client, who hosts a website and is interested in adding new features, is concerned about the website's bounce rate. The bounce rate is the proportion of visitors to a website who navigate away from the site after viewing only one page. Her web developers create two new versions of the website, A and B, and try these new versions on a sample of the website visitors. Out of the 1000 visitors who saw website A, 850 of them left after viewing one page. Out of the 1000 who saw B, 900 of them left.

Calculate the sample proportion (bounce rate) for websites A and B. Is one better than the other? Do you *think* they are significantly different from each other?

### 3.7 Linking Inference, Probability, and Distributions

- **REFER BACK TO WARMUP PUZZLE 1**

- **Statistical Model:** Collection of RVs that describe observable data, their distributions and distribution parameters
- **Parameter  $\theta$ :** Summary of a statistical model or population. Characteristic(s) that determine the joint distribution for RVs of interest. **Parameter Space** describes the set of all possible values of a parameter (ex:  $\lambda \in [0, \infty)$  )
- In statistics, we assume that we are observing a random sample of data from a distribution with parameter  $\theta$ , for example:

$$X_i \sim N(\mu, \sigma^2) \text{ for } i = 1 \dots N$$

And we observe:  $X_1 \dots X_n$ , and calculate some statistics (ex: sample mean)

- **Every Sample Statistic has a Sampling Distribution.** This Sampling Distribution is the distribution that a statistic will take given a specific population probability model with some set of parameters. For example, the distribution of the Sample Mean is (via the Central Limit Theorem)

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

*Sampling Distribution is different from the population distribution*

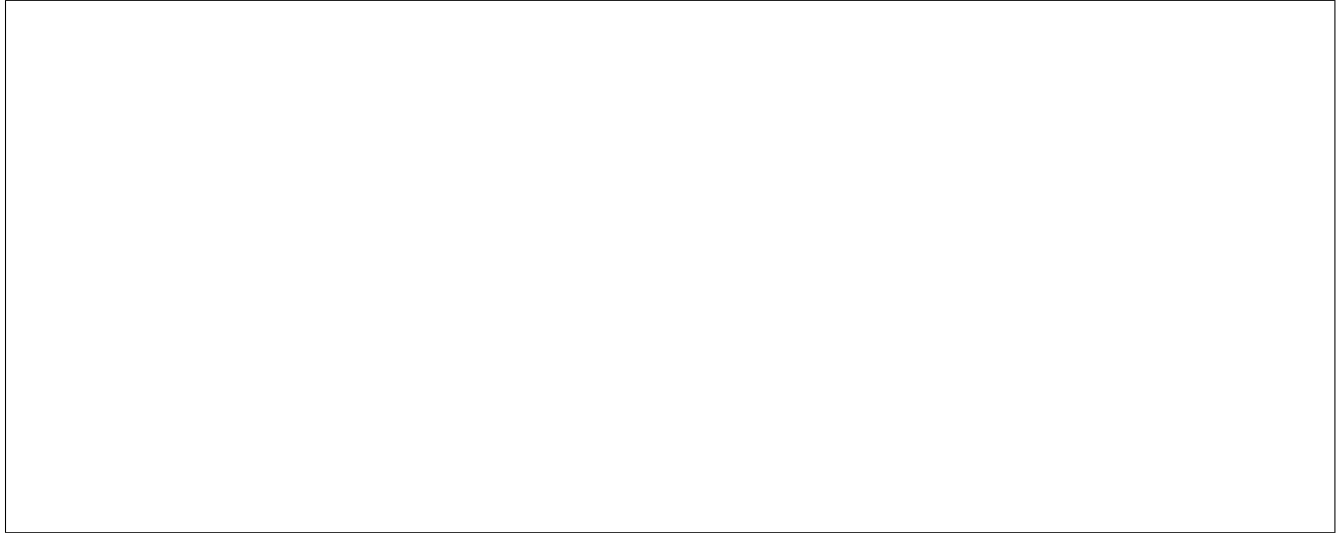
**CRITICAL NOTE:** This is NOT the SD of individuals- rather, it is the SD of our **estimate** of the mean, if we resample many times. The **true mean is just a constant** but what makes our RV random is the sampling procedure!

### 3.8 Normal Distribution is special

- The Normal Distribution is a good model of many natural phenomena. due to the Central Limit Theorem.
- **Central Limit Theorem (Less Casual Definition):** the sum of IID random variables (with finite variance) tends toward a normal distribution, even if the RVs themselves are not normally distributed.
- In other words, the **Normal Distribution can be used to approximate the distribution of the sample mean**, even if the original observations are not normally distributed!

- **Example:** Amount of Sleep for Harvard Students

Define a RV: Suppose that out of  $N = 6700$  undergrads, we sample  $n = 100$  students. Let  $X_i$  denote the amount of sleep the  $i^{th}$  student in our sample gets. Why is  $X_i$  a random variable? Draw a reasonable distribution for  $X_i$ ? Draw a reasonable distribution for  $\bar{X}$ , the sample mean?



- **Binomial Approximation:** Normal distribution can also be used to approximate the binomial distribution under certain conditions.

### 3.9 Tools of Inference: Confidence Intervals and Hypothesis Testing

- In inference, we want to measure population parameters using sample statistics.
- 2 major tools of statistical inference: Confidence Intervals and Hypothesis Testing.
- **Conceptual Example of CI:** We take a survey of a random sample of 1500 teenagers in the United States, and ask them how much they spent that year on movie tickets. Sample mean is \$55. Do you think the population mean (i.e. the average amount spent by teenagers in the US) is EXACTLY \$55? If we took another random sample of the population, do you expect to get a sample mean of \$55 again?

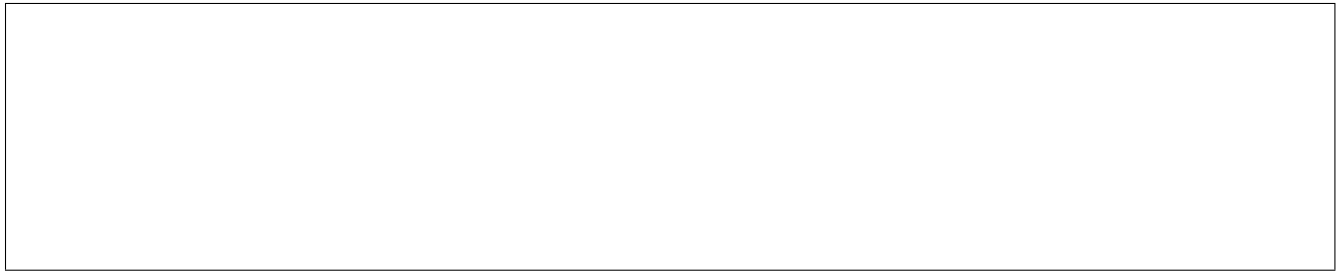


- **Confidence Interval:** describes the uncertainty associated with a sampling method/sampling statistic. For a 95% confidence interval, is interpreted as having a 95% Chance that the true parameter value is contained within this interval. Note that the **true parameter value is fixed**, but the bounds of the CI are random variables.

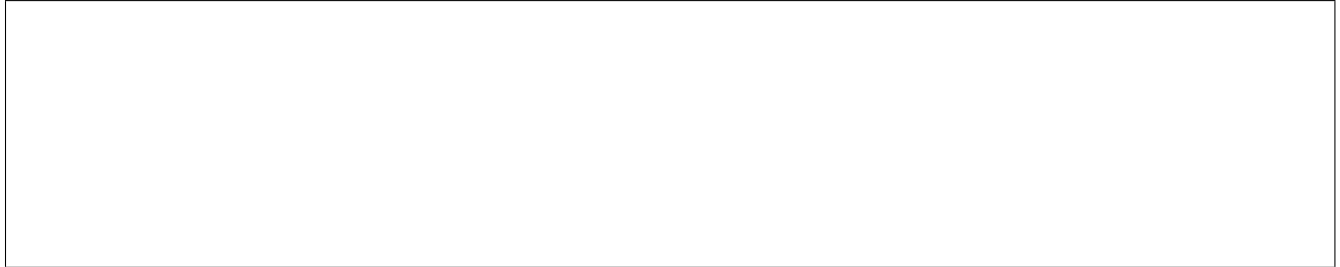
Example of CI:  $55 \pm 3.4$ . Also written as: (51.6, 58.4)

How would you interpret this confidence interval?





- Size of a CI depends on the **sample size** and **confidence level**.



### 3.10 Next Lecture:

#### Constructing a Confidence Interval (in brief)

1. Choose a Sample Statistic: calculate a point estimate using your sample.
2. Select a Confidence Level
3. Calculate the Standard Error
4. Calculate the Confidence Interval

$$\text{Estimate} \pm (\text{Z-Value}) \times (\text{SD of Estimate})$$

- **Hypothesis Testing:** A statistical **hypothesis** is an assumption about a population parameter. This assumption may or may not be true. Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses.<sup>5</sup>

### 3.11 Executive Summary

- The goal of inference is to use observable data to make a statement about a statistical model.
- Because population statistics are often unknown or difficult to measure, we can take a sample of the population and calculate statistics on it.
- Sample Statistics include the sample mean, proportion and variance
- Sample Statistics have Sampling Distributions, with an associated Standard Error
- The Normal Distribution is a good model of many types of data
- Confidence intervals reflect the range where we expect the true parameter value to be. The width of the CI depends on the sample size and confidence level.
- **Next Week:** Computing Confidence Intervals, Hypothesis Testing

---

<sup>5</sup>Definition from: <http://stattrek.com/hypothesis-test/hypothesis-testing.aspx>