STAT 346/446 Lecture 1

Statistical Inference and Review

CB: Sections 5.1 - 5.4
DS: Section 7.1

1. Statistical Inference

2. Random samples – review
   - Random samples from a Normal distribution
   - Order Statistics

What is Statistical Inference?

- In Statistical inference we $\to$ properties of a random sample $X_1, \dots, X_n$
  - use a *sample* to learn about a *population* $\to$ described by a prob. distribution
  - aka. use *data* to learn about a *parameter* (of a population)
- The formal process of this *learning* uses probabilistic models
- We assume that a population of interest can be described by a probability distribution. For example:

models of a population

  - Life time of a Christmas light series follows the $\mathrm{Expo}(\theta)$
  - The diastolic Blood Pressure of all US adults can be modeled as $N(\mu, \sigma^2)$
  - Arrivals of customers can be modeled as $\mathrm{Poisson}$ process with unknown arrival rate λ

Statistical Inference

- We *model* data as realizations of random variables that have the population distribution, $f(x \mid \theta)$ depends on an unknown parameter θ.
- Given the data we have observed (and the model we chose for the population) what can we say about the unknown parameters θ?
  - I.e. we observe random variables

$$X_i \sim f(x \mid \theta) \qquad i = 1, \ldots, n$$

sample size

    and want to draw probabilistic conclusions about θ.
  - This is **parameter estimation**
- Once we have a good estimate of θ we may want to use $f(x \mid \theta)$ to predict new observations from the population
  - This is **prediction**

predict X_0 (or X_{n+1})

not observed

Statistical Inference

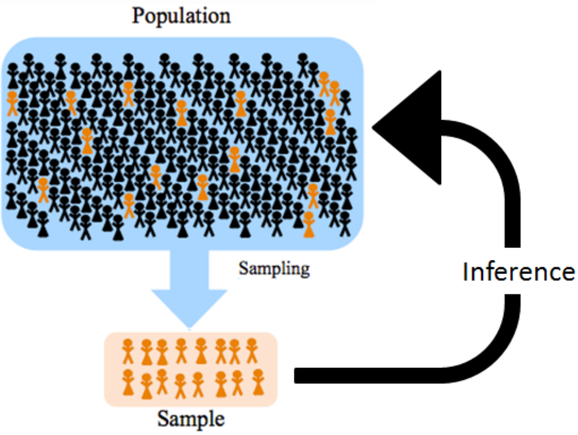
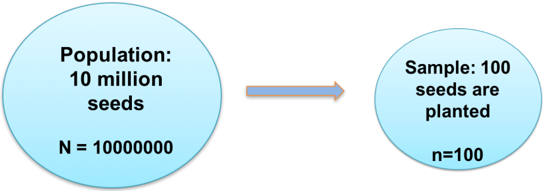


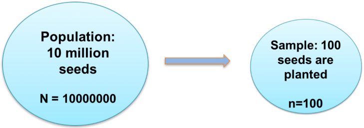
Figure source: `towardsdatascience.com`

Example: Flower seeds

- A storage bin contains 10 million flower seeds each of which either produce white or red flowers.
- How many (or what percentage) will produce white flowers?
- To answer that question we take a sample of 100 seeds, plant them and observe the color or the flowers they produce



Example: Flower seeds



- Population:
  - $p =$ proportion of seeds that give white flowers
  - $1 - p =$ proportion of seeds that give red flowers

The **parameter** p is unknown , a number (not a random variable)

θ

- Sample: n seeds are selected and planted

$$X_i = \begin{cases} 1 & \text{if seed } i \text{ produces a white flower} \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, \ldots, n$$

$X_i \sim \text{Bernoulli}(p)$

What is the distribution of X_i?

Example: Flower seeds

- Random variables

true if all seeds have equal prob. to be in the sample.

$$X_i = \begin{cases} 1 & \text{if seed } i \text{ produces a white flower} \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, \ldots, n$$

What is the distribution of the *statistic* $Y = \sum_{i=1}^{n} X_i$?

$\sim$ Binomial (n, p)

at least approx.

if: indep.
identically distributed Bernoulli (p)

- Data = realization of random variables

$$x_i = \begin{cases} 1 & \text{if seed } i \text{ produces a white flower} \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, \ldots, n$$

or $y =$ total number of seeds that produced a white flower

Example: Flower seeds

- Say we got 89 white flowers out of the $n = 100$ seeds we planted. What does that tell us about p?

- Common sense estimate of p:

$$\hat{p} = \frac{y}{n} = \frac{89}{n} = 0.89$$

= point estimate.

$\longrightarrow$ This is called *Point estimation*

- How sure am I about this number?
  - Uncertainty bounds on the estimate $\longrightarrow$ *Interval estimation*
  - How confident am I that $p < 0.9$? $\longrightarrow$ *Hypothesis testing*

- A customer gets a random sample of 300 new seeds. Based on my data what can I tell them about how many white flowers they can expect? $\longrightarrow$ *Prediction*

Review from STAT 445

Random samples – Review

Sections 5.1-5.4

Key topics

- Random sample
- Statistic, Order Statistic
- Sampling distribution
- Sampling distributions of the sample mean $\overline{X}$ and the sample variance S^2
  - Special case: Normal random sample
- The t_p, and $F_{p,q}$ distributions

Definition of a random sample

Random sample

Random variables $X_1, \ldots, X_n$ are called a

random sample of size n from the population $f(x \mid \theta)$

if $X_1, \ldots, X_n$ are

- mutually independent, and
- marginal pmf/pdf of each X_i is $f(x)$ ← $f(x|\theta)$

- Alternative name for a random sample:
  independent and identically distributed (iid) random variables with pdf or pmf $f(x)$

i.i.d. $f(x)$ = random sample from $f(x)$

Definition of a statistic

A statistic

Let

- $X_1, \ldots, X_n$ be a random sample of size n
- $T(x_1, \ldots, x_n)$ be a real-valued (or vector-valued) function with domain that includes the sample space of $(X_1, \ldots, X_n)$

then

- The random variable (or random vector) $Y = T(X_1, \ldots, X_n)$ is called a **statistic**.
- The probability distribution of Y is called the **sampling distribution of Y**

- In short: A statistic is a function of a random sample.
- Note: Cannot be a function of a parameter.

Commonly seen statistics

- **sample mean:**

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **sample variance:**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- **sample standard deviation:**

$$S = \sqrt{S^2}$$

The random variables $\overline{X}$, S^2 and S all have a (sampling) distribution

Moments of sums

Lemma

Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $g(x)$ be a function such that $\mathrm{E}\,(g(X_1))$ and $\mathrm{Var}\,(g(X_1))$ exist. Then

$$\mathrm{E}\left(\sum_{i=1}^{n} g(X_i)\right) = n\mathrm{E}\,(g(X_1))$$

$$\text{and } \mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = n\mathrm{Var}\,(g(X_1))$$

say $E(X_i) = \mu$ $V(X_i) = \sigma^2$

$E(\sum X_i) = n\mu$

$V(\sum X_i) = n\sigma^2$

Give info about distr. of $\sum_{i=1}^{n} X_i$

or $\sum_{i=1}^{n} g(X_i)$

Moments of some common statistics

Theorem

Let $X_1, \ldots, X_n$ be a random sample of size n from a population with mean μ and variance $\sigma^2 < \infty$. Then

1. $\mathrm{E}\left(\overline{X}\right) = \mu$
2. $\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$
3. $\mathrm{E}\left(S^2\right) = \sigma^2$

Useful fact: For any numbers $x_1, \ldots, x_n$ we have

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sampling from the Normal Distribution

Theorem 5.3.1: Distributions of $\overline{X}$ and S^2

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ and let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

then

(a) $\overline{X}$ and S^2 are independent

(b) $\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$

(c) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

The student's t distribution

Definition: The t distribution

Let $U \sim N(0, 1)$ and $V \sim \chi_p^2$, U and V independent. The distribution of

$$T = \frac{U}{\sqrt{V/p}}$$

is called the ***t* distribution with *p* degrees of freedom, or t_p**

The T statistic

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Snedecor's F distribution

Def: $F_{p,q}$-distribution

Let $X \sim \chi^2_p$ and $Y \sim \chi^2_q$ be independent random variables. The distribution of

$$U = \frac{X/p}{Y/q}$$

is called the *F distribution with p and q degrees of freedom*

The F statistic

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu_X, \sigma^2_X)$.
Let $Y_1, Y_2, \ldots, Y_m$ be a random sample from $N(\mu_Y, \sigma^2_Y)$. Then

$$F = \frac{S^2_X/\sigma^2_X}{S^2_Y/\sigma^2_Y} \sim F_{n-1,m-1}$$

Order Statistics

Let $X_1, X_2, \ldots, X_n$ be a random sample. The statistics

- $X_{(1)} = \min\{X_1, X_3, \ldots, X_n\}$
- $X_{(2)} = \text{second smallest}\{X_1, X_3, \ldots, X_n\}$
- $\ldots$
- $X_{(n-1)} = \text{second largest}\{X_1, X_3, \ldots, X_n\}$
- $X_{(n)} = \max\{X_1, X_3, \ldots, X_n\}$

are called **order statistics**.

Examples:

- Weight of smallest kitten in a litter
- Highest score on an exam

We want to find the *(sampling) distributions* of order statistics

Distributions of order statistics

Theorem: distribution of order statistics

Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution, with pdf $f(x)$ and cdf $F(x)$.

Then the cdf and pdf of the jth order statistic $X_{(j)}$ are

$$F_{(j)}(x) = \sum_{k=j}^{n} \binom{n}{k} F(x)^k (1 - F(x))^{n-k}$$

$$f_{(j)}(x) = \frac{n!}{(j-1)!(n-j)!} f(x)\, F(x)^{j-1}\, (1 - F(x))^{n-j}$$