STA304/1003 H1 F - Summer 2014: Surveys, Sampling, and Observational Data

Lecture 2 - Part II: Review of Statistics/Probability used in Sampling Introduction to Probability Sampling

Ramya Thinniyam

May 20, 2014

Statistics Review - Sampling from an Infinite Population

Graphical Data Summaries:

- Relative Frequency Histogram : symmetry, shape, outliers, patterns, spread, central tendency
- Boxplot : symmetry, outliers, quartiles, etc.
- QQ-Plot: normality, symmetry, outliers

Example: Old Faithful Geyser

Old Faithful is a cone-type geyser in Yellowstone Park, USA. Eruptions can shoot 3,700 to 8,400 US gallons of boiling water to a height of 106 to 185 feet (32 to 56 m) lasting from 1.5 to 5 minutes. Intervals between eruptions can range from 45 to 125 minutes.

In R:

1. Make a directory called Rdata on your C drive
2. Save data file as a *.csv file in Rdata
3. Read data in using read.csv command:
   $>$ oldfaithful $<-$ read.csv("C:/Rdata/oldfaithful.csv")
4. Columns are called oldfaithful$eruptions and oldfaithful$waiting

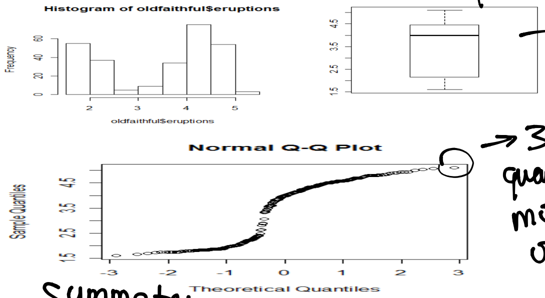
Graphs

```
> hist(oldfaithful$eruptions)
> boxplot(oldfaithful$eruptions)
> qqnorm(oldfaithful$eruptions
```

What features are apparent from the graphs?

box plot

median



→ 3 or -3 quantile or more, outlier

Symmetry

Numerical Data Summaries / Statistics

Data: $y_1, y_2, \ldots, y_n$

Sample Mean: $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

- Measures location
- Estimates population mean, μ

Sample Variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$

- Measures spread
- Estimates population variance, σ^2

AIM of Statistics: Estimate parameters of interest and quantify error not random

unknown quantity

take a sample to estimate

Parameter:

- Usually denoted θ
- A characteristic of the population - fixed , unknown

Estimator: (random variable)

- Usually denoted $\hat{\theta}$
- A statistic (function of sample data) used to estimate a parameter (most common is eg. $\bar{y}$)

Estimate:

- Numeric value of an estimator

Confidence Interval : 95%

- $100(1 - \alpha)\%$ of samples generate a CI that covers the true parameter
- Sample sizes selected to ensure error of estimation is less than B,

$$P(|\hat{\theta} - \theta| < B) = 1 - \alpha$$

typically is 95% CI

Probability Framework

A (Random) Experiment is a process that can be repeated resulting in a single outcome that cannot be predicted with certainty.
Ex. tossing a die, flipping a coin, picking a card from a deck, randomly picking a ball from an urn with 2 red balls and 4 blue balls

A sample point is a single outcome of an experiment.
Ex: Toss a die - 6 or 1 or 2 ... etc
Ex: Flip a coin - H or T

Sample Space and Events

Sample space, Ω / S is the set of all possible sample points of an experiment.
Ex: Toss a die and observe the up face - $\Omega = \{1, 2, 3, 4, 5, 6\}$
Ex: Flip two coins and observe the up faces - -
$\Omega = \{HH, HT, TH, TT\}$

An event is a specific collection of sample points (subset of the sample space).
Events are denoted by capital letters like A, B, etc.

Ex: Toss a die and observe the up face. A is the event "even number" $A = \{2, 4, 6\}$, B is the event "multiple of 3" $B = \{3, 6\}$
Ex: Flip two coins and observe up faces. A: at least one head, B: exactly one head

$S = \{HH, HT, TH, TT\}$
$A = \{HH, HT, TH\}$
$B = \{HT, TH\}$

Compound Events

Union, $(A \cup B)$ of two events A and B contains all outcomes in A or B (or both)

Intersection, $(A \cap B)$ of two events A and B contains all outcomes which are in both A and B

Complement, $(A^c$ / $\bar{A}$ / $A')$ of an event A contains all the outcomes that are NOT in A.

Two events are called Mutually Exclusive / Disjoint if they have no outcomes in common. ie their intersection is the empty set.

Example: Toss two coins and observe the up faces.

A: at least one head, B: exactly one head, C: head on the first toss, D: tail on the first toss

Find $A \cup B$, $C \cup D$, A^c , D^c. Which events are mutually exclusive?

Probability

- Probability is a number between 0 and 1 assigned to each of the outcomes of a random experiment. Probability of an event A is denoted $P(A)$.

  If a sample space has k possible outcomes that are equally likely, then the probability of any one outcome is $\frac{1}{k}$. Then,

  $P(A) = \frac{number\ \ of\ \ outcomes\ \ in\ \ A}{total\ \ number\ \ of\ \ outcomes}$.

Axioms of Probability

Properties of probabilities in finite sample spaces:

1. $P(\Omega) = 1$
2. For any event A, $0 \leq P(A) \leq 1$
3. $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$ if A_i are disjoint

Other Useful Rules:

1. $P(A) + P(A^c) = 1$
2. Additive Rule:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Conditional Probability

Conditional Probability of A given B is defined as the probability that the resulting outcome is one of the outcomes of A given that we know that it is one of the outcomes from B.

• Sample space is reduced for this probability. New sample space = sample space for B

Conditional Probability Formula:

$P(A|B) = \frac{P(A \cap B)}{P(B)}$, provided $P(B) \neq 0$

Independence

Events are called Independent if the occurrence of one event does not affect the probability of the other event.
Events are that are not independent are called Dependent.

If A and B are independent events, then $P(A|B) = P(A)$ and $P(B|A) = P(B)$

For any events using conditional probability formula, we have:
$P(A \cap B) = P(A|B)\ P(B) = P(B|A)\ P(A)$

This definition leads to:

If A and B are independent events, then $P(A \cap B) = P(A)\ P(B)$

General Result:

$$P(A_1 \cap A_2 \cap \ldots \cap A_k) = P(A_1|A_2 \cap \ldots \cap A_k) P(A_2|A_3 \cap \ldots \cap A_k) \ldots P(A_{k-1}|A_k) P(A_k)$$

(Useful formula for Cluster Sampling)

Connection to Sampling

In Sampling:

Population - N units Sample - n units

Think of N balls in a box labelled $1, 2, \ldots, N-1, N$ and draw n balls. Called Simple Random Sampling.

Do we sample with replacement or without replacement?

without,
b/c we don't want
to select the same
unit more than once!

But in some
cases...

Simple Random Sampling with Replacement

In Simple Random Sampling with Replacement (SRSWR), a unit is placed back into the population after being selected. ie. put ball back in the box, same population is used for each draw

- N^n possible samples
- Each sample has probability of $\frac{1}{N^n}$ of being selected
- Usually do not care about the order within a sample

Example: $N = 5, n = 2$

a) Find $P(\{4,5\}) = P(4 \textit{ and } 5 \textit{ are selected in the sample})$ $= P((4,5) \cup (5,4)) = \frac{2}{25}$

b) Let $A = \{4 \textit{ is selected on the first draw}\}$ and $B = \{4 \textit{ is selected on the second draw}\}$. Find $P(A \cap B)$. $= P((4,4)) = \frac{1}{25}$

c) Find $P(4 \textit{ is selected in the sample})$ $= P(A) + P(B) - P(A \cap B) = \frac{1}{5} + \frac{1}{5} - \frac{1}{25} = \frac{9}{25}$

Simple Random Sampling without Replacement

In Simple Random Sampling without Replacement (SRS), a unit cannot be selected again.

More efficient, use this most of the time.

- $\binom{N}{n} = \frac{N!}{(N-n)!n!}$ possible samples
- Each sample has probability of $\frac{1}{\binom{N}{n}}$ of being selected
- Order does not matter
- Successive draws are NOT independent

Random Variables

A Random Variable (RV) is a function that assigns a numerical value to each outcome in the sample space. (a variable whose value is determined by chance)

Random variable names: Upper-case letters (e.g. X, Y, Z, etc.)
Values they take on are called realization : corresponding lower-case letters (e.g. x, y, z, etc.)
The set of values that the RV can take on are often called the "support" of the random variable

Two Types of RVs:

1. Discrete: A RV that can take one of a countable or finite list of distinct values.Its support is a collection of isolated points on the number line.
2. Continuous: A RV that can take any value in an interval (or collection of intervals) on the real line.

Probability Distributions

Probability Distribution, denoted $p(x) = P(X = x)$ is a graph, table, or formula that specifies the probabilities associated with each value of the discrete random variable.

Requirements for Discrete Probability Distribution:

1. $0 \leq p(x)$ (≤ 1) for each value x
2. $\sum_x p(x) = 1$

Expected Values

Mean/Expected Value/Expectation, denoted μ / μ_X / $E(X)$:
expected average value of RV over the long run.

$$\mu_X = E(X) = \sum_x x\, p(x) ,$$

- $V(X) = E[(X-\mu_X)^2] = E(X^2) - \mu_X^2 = Cov(X,X)$:
  Variance - Spread
- $\sigma_X = STD(X) = \sqrt{V(X)}$: Standard Deviation
- $Cov(X,Y) = E[(X-\mu_X))(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y$:
  Covariance - How much two variables vary together (linear)
- $Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$: Correlation - Standardized
  covariance How much the 2 variance cluster about the line?

Properties of Expected Values

1. For any function g, $E[g(X)] = \sum_x g(x)\, p(x)$
2. For constants a and b, $E(aX + b) = aE(x) + b$
3. If X and Y are independent, $E(XY) = E(X)E(Y)$
   ie. $Cov(X, Y) = 0$
4. $$Cov\left[\sum_{i=1}^{n}(a_i X_i + b_i), \sum_{j=1}^{m}(c_j Y_j + d_j)\right] = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i c_j Cov(X_i, Y_j)$$
5. $V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$
6. $-1 \leq Corr(X, Y) \leq 1$