STAT206

Andy Zhang

Fall 2014

Contents

Chapter 1

Introduction

1.1 Statistics

1.1.1 Definitions

Statistics Collection, organization, analysis, interpretation and presentation of data. It is also defined as the quantification of uncertainty.

Unit A single element, usually a person or object, whose characteristics are of interest. Ex: A student enrolled in the course.

Population The set of all units which are of interest. Ex: All students enrolled in the course

Variable A measurement of the characteristic of interest from a unit. Ex: Number of Canadian provinces visited by a student

Sample A subset of units from the population for which measurements of the desired variable are actually made. Ex: 29 students chosen from the class

Descriptive Statistics Summarize the data in the sample, both graphically and numerically

Inferential statistics USe the sample data to estimate an attribute of the population. Include a quantification of uncertainty

Sampling Error An error which occurs due to the uncertainty in randomly selecting a sample.

Study error A systematic error which occurs because the sample does not accurately represent the population

1.1.2 Process

Identify the problem of interest

- Who or what do you want to learn about?
  - Define the **population** of interest

- Individual elements of the population are called **units**

- What research question would you like answered?
  - Define your **hypothesis**

Plan the data collection

- How will you select a subset of **units** from the **population** to be in your **sample**?
  - How large will the **sample** be?
- What is (are) the **variable (s)** of interest?
  - How will you measure it (them)?

Analyze the data

- Graph the data — histogram, scatter-plot, etc
- Compute **Descritive statistics** — e.g. sample mean, sample variance, etc.
- Compute **Inferential statistics** — e.g. confidence intervals, hypothesis tests about population **parameters**
  - Inferential statistics include a quantification of the sampling error

Draw conclusions

- Use the results of your analysis to address the original research question
- Address limitations of the study, especially any potential systematic **study errors**

1.1.3 Data Types

Categorical Variable A qualitative measure. Each unit belongs to **one of K** possible classes.

Discrete variable A quantitative measure. Each unit's measurement can take on one of a **countable** number of possible values

Continuous variable A quantitative measure. Each unit's measurement can take on an **uncountable** number of possible values, usually some interval of real numbers

1.1.4 (Grouped) Frequency Tables

- Display the number of units which are in each class
- Discrete / Continuous variables are grouped into classes
- In the case of numerical variables, there is a loss of information

See more: `http://en.wikipedia.org/wiki/Stem-and-leaf_display`

1.1.5 Stem and Leaf Plot

- A **stem-and-leaf plot** is a way to summarize a relatively **small** data set, without the loss of information that occurs with a frequency table
- Left is possible **first** digits, right is remaining digits in ascending order

See more: `http://en.wikipedia.org/wiki/Stem-and-leaf_display`

1.1.6 Bar Chart

- Bar charts are used to graphically display information from categorical variables

See more: `http://en.wikipedia.org/wiki/Bar_chart`

1.1.7 Histogram

- A histogram is similar to a bar chart, but it's for numerical data
- The range is divided in distinct classes, and each observation is assigned to exactly one class
- Histogram shows frequency of observations in each class

See more: `http://en.wikipedia.org/wiki/Histogram`

- If class ranges are not same length, we can use density histogram instead
- When interpreting a density histogram, it is the area that is meaningful
- Height is $height = \frac{relativefrequency}{width} = \frac{frequency}{width*n}$

See more `http://en.wikipedia.org/wiki/Histogram`

1.1.8 Measures of Centrality

- The **sample mean** of a set of n values, $x_1, x_2, x_3, \ldots, x_n$ denoted by $\bar{x}$ is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- The **median** is the number x^* such that half of the observed values are below x^* and half are above
- If after writing our values in ascending order, we donte the i^{th} value as $x_{(i)}$, then

$$x^* = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})} & \text{if } n \text{ is even} \end{cases}$$

1.1.9 Measures of Variability

Measures of variability

- The **sample variance** of a set of values $x_1, x_2, x_3, \ldots, x_n$ denoted by s^2 is

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

- The **sample standard deviation** denoted s, is the square root of the sample variance
- The **range** of the set is the difference between the maximum and minimum value

$$range = x_{(n)} - x_{(1)}$$

1.1.10 Box Plot

- The box indicates the middle 50% of the observations, i.e. the second and third quartiles
- The line through the box indicates the median observation
- The whiskers indicate the highest and lowest observations

See more: `http://en.wikipedia.org/wiki/Box_plot`

Chapter 2

Probability

2.1 Definitions

Probability measure the uncertainty associated with an event. An event is something that might occur

- Classical: $\frac{Number of ways event can occur}{Total number of equally likely outcomes}$
- Relative Frequency: Proportion of times the event occurs, as the number of trials approaches infinity
- Subjective: Estimates of probability that the event occurs, based on subjective opinion

Experiment is a repeatable phenomenon or process

Trial is a single repetition of an experiment

Sample Space , S, is the set of distinct outcomes for an experiment or process

Discrete A sample space is discrete if it has a finite or countably infinite number of simple events. Otherwise it is non discrete or continuous

Mutually Exclusive means two events never occur simultaneously

Complement of an event and an event are always mutually exclusive

Uniform distribution The total probability is uniformly distributed among all possible outcomes

Permutation is the number of ways to arrange r out of n objects: $n^{(r)} = \frac{n!}{(n-r)!} = n(n-1)(n-2)...(n-r+1)$

Combinations If we don't care about the order of objects, but just which objects are chosen, the number of ways to choose r out of n items is $\binom{n}{r} = \frac{n!}{r!(n-r)!}$

Set Operations

- AB or $A \cap B$ is the intersection of two events.
- $A \cup B$ is the union of two events.
- $\bar{A}$ is the complement of A, not event A

Conditional The probability of event A, conditional on the occurence of event B, denoted by $P(A|B)$ is $P(A|B) = \frac{P(A \cap B)}{P(B)}$, $P(B) \neq 0$

Independent Two events are said to be independent iff $P(A \cap B) = P(A)P(B)$. This implies that $P(A|B) = P(A)$, $P(B|A) = P(B)$. In other words, events A and B are independent if whether B occurs does not influence whether A occurs, and vice versa

Bayes' Theorem Suppose A and B are any two events in S, then $P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|\bar{B})P(\bar{B})}$

Chapter 3

Random Variables

3.1 Definitions

Random Variable X is a function from the sample space S to the real numbers $X : S \to \mathbb{R}$. Its range $R(X)$ is the set of possible real values it can take. We use X to denote a random variable and x to denote its observed value

- Discrete if takes on a countable number of possible values
- Continuous if it takes on all values in some interval of the real line

Indicator or binary variables take values of 0 or 1

Probability Function (pf) of X, denoted $f(x)$, denotes the probability that X takes on the value x. $f(x) = P(X = x)$ defined for all x in the range of X

Probability Distribution The set of pairs $\{(x, f(x))|x \in R(X))\}$ is called the probability distribution of X

- $0 \leq f(x) \leq 1$ for any x
- $\sum f(x) = 1$

Cumulaative distribution function (cdf) of X, denoted $F(x)$ denotes the probability that X rtakes on a value $\leq x$: $F(x) = P(X \leq x)$. Very useful for continuous random variables

Mean or Expected Value of X is defined as $\mu = E(X) = \sum_x xf(x)$. We can understand $E(X)$ as the average value that X would assume over a theoretically infinite number of trials. $E(X)$ is not a random variable, it is a constant.
We also define the expected value as $E[g(X)] = \sum_x g(x)f(x)$

Variance of a random variable X is the expected squared difference from the mean, that is $Var(X) = E[(X - E(X))^2] = \sum_{allx} f(x)(x - \mu)^2$. An alternative would be $Var(X) = E[X^2] - (E(X))^2$

Chapter 4

Discrete Probability Distributions

4.1 Definitions

Bernoulli Distribution Repeated trials of an experiment

- Each trial can be a success or a failure
- The probability of a success is the same for each trial
- The outcomes of different trials are **independent**
- Let X record success or failure

We say that X follows a **Bernoulli distributionS** ($X \sim Bernoulli(p)$), where p is the probability of sucesss

$$f(x) = \begin{cases} p & \text{if } x = 1 \\ (1-p) & \text{if } x = 0 \end{cases}$$

$E(X) = p$
$Var(X) = E[X^2] - (E(X))^2 = p - p^2 = p(1-p)$

Binomial Distribution Physical setup: We perform a sequence of n independent Bernoulli trials

- Each trial has two possible outcomes: success or failure
- Trials are independent
- Each trial has probability of success equal to p

$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, 2, ...n$
$E(X) = np$
$Var(X) = np(1-p)$

Poisson Process Physical setup: Events occur randomly in time (or space) according to the following conditions

- Independence: The number of occurrences in disjoint (non overlapping) intervals are independent
- Individuality: Events occur singly i.e $P(\text{two or more events occur simultaneously}) = 0$
- Homogeneity: Events occur according to a uniform (constant) rate or intensity (λ)

If events occur with an average rate of λ per unit of time and X is the number of events which occur in t units of time, then $X \sim Poisson(\lambda t)$
$f(x) = \frac{e^{\lambda t}(\lambda t)^x}{x!}$, $x = 0, 1, ...$

$E(X) = \lambda t$
$Var(X) = \lambda t$

Hypergeometric Distribution Physical setup: We have a collection of N objects which can be classified into two distinct types, called success and failure. There are r and $N - r$ failures. A sample of n objects is selected without replacement.
Let X be the number of successes selected, then X is said to follow a hypergeometric distribution ($X \sim Hyper(N, r, n)$). To compute the probability functino, note that, for X = x

- There are $\binom{N}{n}$ points in the sample space (if we do not consider the order of selection)
- There are $\binom{r}{x}$ ways too select the x success objects from the r available
- There are $\binom{N-r}{n-x}$ ways to select the remaining $n - x$ failure objects from the $N - r$ available

$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$, $x = 0, 1, ...min(r, n)$
$E(X) = \frac{nr}{N}$
$Var(X) = \frac{nr(N-r)(N-n)}{N^2(N-1)}$

Geometric Distribution Physical Setup: Bernoulli Trials are repeated until the first success. Let X be the number of independent Bernoulli (p) trials until the first success (including the first success), then X follows a geometric distribution, $X \sim Geom(p)$
$f(x) = p(1-p)^{x-1}$, $x = 1, 2...$
$E(X) = \frac{1}{p}$
$Var(X) = \frac{1-p}{p^2}$

Chapter 5

Continuous Probability Distributions

5.1 Definitions

Continuous Random Variable X is a function from the sample space to the real numbers: $X : S \to \mathbb{R}$
The range $R(X)$ is continuous. Individual points in $\mathbb{R}$ must have a 0 probability since the interval length is 0

Probability Density Function (pdf) of a continuous random variable X, denoted $f(x)$ assigns a probability to an $x \in R(X)$. If we have the probability density function for X, then we can define the probability that X takes a value in an interval $(a, b) \subseteq R(X)$ as $P(A < X < b) = \int_a^b f(x)dx, (a, b) \subseteq R(X)$
Properties of the probability density function:
$f(x) \geq 0, \forall x \in R(X)$
$\int_{x \in R(X)} f(x)dx = 1$

Cumulative Distribution Function of a continuous brandom variable X, denoted F(x), gives the probability that X takes on a value less than or equal to x
$F(x) = P(X \leq x) = P(X < x)$
Properties

- $F(-\infty) = 0$
- $F(\infty) = 1$
- $F(x)$ is non decreasing

Continuous Uniform Distribution Physical setup: The probability of any subinterval of the range is proportional to the length of the interval. For $a < b$, if X is uniformly distributed on the interval (a, b) then we write $X \sim U(a, b)$
$f(x) = \frac{1}{b-a}$, $a \leq x \leq b$

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } b < x \end{cases}$$

Mean or Expected Value of a continuous random variable X is defined as
$E(X) = \int_x xf(x)dx = \mu$
Properties:

- $E[g(X)] = \int_x g(x)f(x)dx$
- $E[aX + bY] = aE(X) + bE(Y)$

Variance is the expected squared difference from the mean, that is $Var(X) = E[(X - E(X))^2]$. If X_1 and X_2 are independent random variables, and $a, b \in \mathbb{R}$, then $Var[aX_1 + bX_2] = a^2 Var(X_1) + b^2 Var(X_2)$

Exponential Distribution Physical setup: Events occur according to a Poisson process, and we measure the inter arrival times between events. If X is the amount of time until the next event in a Poisson process, then $X \sim Exp(\theta)$, where $\theta = \frac{1}{\lambda}$
$f(x) = (\frac{1}{\theta})e^{-\frac{x}{\theta}}$, $x > 0$
$F(x) = 1 - e^{-\frac{x}{\theta}}$, $x > 0$
$E(X) = \theta$
$Var(X) = \theta^2$

Simulation The most common use of the continuous uniform distribution to simulate other random variables.
Theorem: If $F(x)$ is an arbitrary cdf, and $Y \sim U(0, 1)$, then $X = F^{-1}(Y)$ has cdf $F(x)$.
We can use Y to generate an observation of the random variable X:

- Generate an observation y from $Y \sim U(0, 1)$ using your favourite software
- Compute $x = F^{-1}(y)$

Chapter 6

Normal Distribution

6.1 Definitions

Normal Distribution A continuous random variable X with range $(-\infty, \infty)$ has a normal distribution denoted $X \sim N(\mu, \sigma^2)$, if its pdf has the form $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$, $x \in \mathbb{R}$
where the mean μ and variance σ^2 are parameters.
$E(X) = \mu$, $Var(X) = \sigma^2$

Properties Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ be independent
$Y = aX_1 + bX_2 + c \sim N(a\mu_1 + b\mu_2 + c, a^2\sigma_1^2 + b^2\sigma_2^2)$
If $X \sim N(\mu, \sigma^2)$, then $Z = (\frac{1}{\sigma})X - (\frac{\mu}{\sigma}) = \frac{X-\mu}{\sigma} \sim N(0, 1)$
We also have $P(Z > z) = P(Z < -z)$ for any $z \in \mathbb{R}$

Central Limit Theorem : We use the normal distribution to apporximate probabilities for non normal distributions. This is possible because the normal distribution tends to approximate sums of randm variable. Although this is a Theorem about limits, we will use it when n is large, but finite to approximate the distribution of $\sum_i X_i$, or $\bar{X}$ by a normal distribution

Independent We say that X and Y are independent if, for all x and y, we have
$f(x, y) = P(X = x \cap Y = y) = P(X = x)P(Y = y) = f_X(x)f_Y(y)$

Central Limit Theorem — Sum Let X_1, X_2, X_3,...X_n be independent variables all having the same distribution with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$
As $n \to \infty$, the cumulative distribution function of the random variable $\sum_i X_i$ approaches the cumulative distribution function for $N(n\mu, n\sigma^2)$
The cumulative distribution function of the random variable $\frac{\sum_i X_i - n\mu}{\sigma\sqrt{n}}$ approaches the cumulative distribution function for $N(0, 1)$

Central Limit Theorem — Average Let X_1, X_2, X_3,...X_n be independent variables all having the same distribution with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$
As $n \to \infty$, the cumulative distribution function of the random variable $\bar{X}$ approaches the cumulative distribution function for $N(\mu, \frac{\mu^2}{n})$

The cumulative distribution function of the random variable $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ approaches the cumulative distribution function for $N(0,1)$

Continuity Correction can improve the approximation to a sum or average of discrete random variables using a normal random variable. We think of the center of a bar with width 1 as an integer value and the bar actually covering $(x-0.5, x+0.5)$. So instead of integrating from $(0,5)$ for example, we integrarte on $(-0.5, 5.5)$

Chapter 7

Confidence Intervals

7.1 Definitions

Introduction to Estimation Suppose that a probability distribution which serves as a model for some random process depends on an unknown parameter θ. In order to use the model we have to estimate or specify a value for θ using some data sets collected for the random variable.

Estimate of a parameter θ is the value of a function of the observed data $y_1, y_2, ..., y_n$ and other known quantities such as the sample size n.

Likelihood function for θ is defined as $L(\theta) = L(\theta; y) = P(Y = y; \theta)$ for $\theta \in \omega$
where the parameter space ω is the set of possible values for θ

Suppose that θ is the success probability in a binomial model, so that $Y \sim Bi(n, \theta)$. Suppose that we ran the experiment once and recorded y successes in n trials. Then
$L(\theta) = P(Y = y; \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$

- Hold y constant and vary θ.
- This makes $L(\theta)$ into a function of θ.
- Then we can use calculus to choose θ to maximize $L(\theta)$
- This choice for θ is what we call $\hat{\theta}$, the maximum likelihood estimate for θ.
- Here we would get $\hat{\theta} = \frac{y}{n}$ which agrees with our intuition

Maximum Likelihood Estimate The value of θ which maximizes $L(\theta)$ for given data y is called the maxmimum likelihood estimate of θ. It is denoted as $\hat{\theta}$.

Relative Likelihood Function is defined as $R(\theta) = \frac{L(\theta)}{L(\hat{\theta}}$ for $\theta \in \omega$.
Note that $0 \leq R(\theta) \leq 1$ for all $\theta \in \omega$

Log Likelihood Function is $l(\theta) = \ln L(\theta)$ for $\theta \in \omega$. Note that $\hat{\theta}$ maximizes $R(theta)$ and $l(\theta)$

Estimator We call $\tilde{\theta}$ the estimator of θ corresponding to $\hat{\theta}$. We will always use

- $\hat{\theta}$ to denote an estimate, that is, a numerical value
- $\tilde{\theta}$ to denote the corresponding estimator

An estimator $\tilde{\theta}$ is a random variable which is a function $\tilde{\theta} = g(Y_1, ...Y_n)$ of the random variables $Y_1, ..., Y_n$. The distribution of $\tilde{\theta}$ is called the sampling distribution of the estimator.
List of estimators:

- $Y_i \sim Bernoulli(p)$, $\tilde{p} = \bar{Y}$
- $Y_i \sim Poisson(\lambda)$, $\tilde{\lambda} = \bar{Y}$
- $Y_i \sim Exponential(\theta)$, $\tilde{\theta} = \bar{Y}$
- $Y_i \sim Normal(\mu, \sigma^2)$, $\tilde{\mu} = \bar{Y}$
- $Y_i \sim Normal(\mu, \sigma^2)$, $\tilde{\sigma^2} = \bar{s^2}$

Unbiased An estimator is said to be unbiased if its expected value equals the parameter being estimated: $E(\tilde{\theta}) = \theta$.

The standard deviation of an estimator is called its standard error: $SE(\tilde{\theta}) = \sqrt{Var(\tilde{\theta})}$
If we have two unbiased estimators for a parameter, the one with the smaller standard error is preferred.

Interval Estimate for θ based on observed data y takes the form $[L(y), U(y)]$. We assume that the probability model chosen is correct and that θ is the true value of the parameter.

Coverage Probability To quantify the uncertainty in the interval estimate, we define the coverage probability for the interval estimator $[L(Y), U(Y)]$ as
$C(\theta) = P(L(Y) \leq \theta \leq U(Y))$

Confidence Intervals A 100p% confidence interval for a parameter θ is an interval estimate $[L(y), U(y)]$ for which
$P(L(Y) \leq \theta \leq U(Y)) = p$. We say that we are 100p% confident that the true parameter is in the interval

Pivotal Quantity is a function of the data Y and the unknown parameter θ such that the distribution of the random variable Q is completely known. We define it as: $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$.

To compute a 100p% confidence interval for μ, determine the threshold c such that $P(Z \leq c) = 1 - (\frac{1-p}{2}) = \frac{p+1}{2}$, where $Z \sim N(0, 1)$.
We then construct the interval $\bar{x} \pm c\frac{\sigma}{\sqrt{n}}$