Review of Statistics — Data

Data is a collection of data objects and their attributes. Each attribute is a property or characteristic of the object, a collection of attributes describe an object.

Data Types

We have following types of data:

Nominal: e.g. ID numbers, eye color, zip codes

Ordinal: e.g. rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}

Interval:e.g. calendar dates, temperatures in Celsius or Fahrenheit

Ratio: Examples: temperature in Kelvin, length, time, counts

We classify nominal and ordinal data as **Categorical Qualitative** data, classify interval and ratio data as **Numeric Quantitative** data.

For some statistical methods, you must specify numerical variables as either being discrete or continuous.

Discrete numerical data: counting process, like number of something.

Continuous numerical data: measuring process, like time, length.

Four types of data properties:

Distinctness: =, ≠

Order: <, >

Addition: +, -

Multiplication: *, /

Also, the type of data depends on which of the following properties it possesses:

Nominal data: distinctness

Ordinal data: distinctness & order

Interval data: distinctness, order & addition

Ratio data: all 4 properties

Question 1: Is it physically meaningful to say that a temperature of 10° degrees twice that of 5° on the Celsius scale? the Kelvin scale?

Question 2: If Bill's height is three inches above the average and Bob's height is six inches above the average, then would we say that Bob is twice as tall as Bob?

Ratio vs Interval: with or without a true zero point.

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

Sometimes, Only presence (a non-zero attribute value) is regarded as important, in this case we use **Asymmetric** data. e.g. words present in documents, dummy categories.

Data Quality

Poor data quality negatively affects many data processing efforts.

Examples of data quality problems:

Noise: noise refers to modification of original values

Outliers: outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set.

Mostly, outliers are noise that affect data analysis and decision, but sometimes outliers are the goal of our analysis, e.g. credit card fraud, intrusion detection.

Missing values: information is not collected, or may not be applicable (e.g., annual income is not applicable to children).

Usually, we have 3 ways to deal with missing values: eliminate data objects, estimate missing values, or ignore the missing value during analysis.

Duplicate data: data set may include data objects that are duplicates, or almost duplicates of one another

---

Review of Statistics — Random Variables

Random Variables

Random variables represent outcomes from random phenomena. They are specified by two objects. The **range** R of possible values and the **frequency** $f(x)$ with which values from within the range can occur. When the range is a discrete set we have a **discrete random variable** and when the range is continuous we have a **continuous random variable**. See several examples below. For notational convenience we often extend the range to a larger set where outcomes outside the original range are assigned zero frequency. The terms: a random variable X, the distribution $f(x)$, or the Range R, is discrete (respectively continuous) are equivalent.

The cumulative frequency or distribution function, of a random variable X, is defined as

$$F(x) = \begin{cases} \sum_{-\infty}^{x} f(y), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{x} f(y)dy, & \text{if } X \text{ is continuous} \end{cases}$$

Expectation and Moments

The expected value (or first moment) of a random variable X is a constant $E(X)$ defined by

$$E(x) = \int_R x dF(X) = \begin{cases} \sum_{-\infty}^{\infty} xf(x), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} xf(x)dx, & \text{if } X \text{ is continuous} \end{cases}$$

The n-th moment of a random variable X is a constant $E(X^n)$ defined by

$$E(x^n) = \int_R x^n dF(X) = \begin{cases} \sum_{-\infty}^{\infty} x^n f(x), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n f(x) dx, & \text{if } X \text{ is continuous} \end{cases}$$

The variance of a random variable X is a constant $Var(X)$, or $\sigma^2(X)$, defined by

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

Joint Distributions - Independent Random Variables

Often two random variables X and Y, have a joint distribution defined by a two dimensional frequency function $f(X, Y)$. In the discrete case

$$f(x, y) = P(X = x, Y = y).$$

In the continuous case

$$P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x', y') dx' dy'.$$

The marginal frequency, in the discrete case, is defined

$$f(x) = P(X = x) = \sum_{y=-\infty}^{\infty} f(x, y).$$

In the continuous case it is easier to define the marginal distribution:

$$F(x) = P(X \leq x) = \sum_{y=-\infty}^{\infty} f(x, y).$$

The covariance of two jointly distributed random variables X and Y is a constant $Cov(X, Y)$ defined by

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$

Two jointly distributed random variables X and Y are independent if:

$$F(x, y) = F(x)F(y) \text{ for all } x, y$$

Properties of Expectation and Variance

$$E(aX+b) = aE(X) + b,$$

$$E(X+Y) = E(X) + E(Y),$$

$$E(XY) = E(X)E(Y), \text{ for independent } X \text{ and } Y,$$

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

$$Var(aX+b) = a^2 Var(X),$$

$$Var(X+Y) = Var(X) + Var(Y), \text{ for independent } X \text{ and } Y,$$

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y),$$

$$Cov(X,Y) = E(XY) - E(X)E(Y),$$

$$Cov(X,X) = Var(X),$$

Discrete Random Variables & Distribution

- Bernoulli Random Variable

The Bernoulli random variable represents experiments (trails) with only two possible outcomes - success (1) or failure (0). We write $X \sim Bernoulli(p)$.

where $P(X=1) = p$ and $P(X=0) = 1 - P(X=1) = 1-p$.

$$E(X) = p,\ Var(X) = p(1-p).$$

- Binomial Random Variable

The Binomial random variable represents the number of successes in n independent Bernoulli trials each with success probability p. We write: $X \sim Binomial(n,p)$.

Probability mass function (p.m.f.): $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$.

$$E(X) = np,\ Var(X) = np(1-p).$$

- Geometric Random Variable

The Geometric random variable represents the number of trials until you get the first success in n independent Bernoulli trials each with success probability p. We write: $X \sim G(p)$.

$$\text{p.m.f. } P(X = k) = \binom{n}{k} p(1-p)^{k-1}.$$

$$E(X) = \frac{1}{p},\ Var(X) = \frac{1-p}{p^2}.$$

Memoryless Property. An important property of the geometric distribution is:

$$P(X > s + t \,|\, X > t) = P(X > s) \text{ for all } s, t > 0.$$

- Poisson Random Variable

The Poisson random variable typically represents the number of events that occur in a fixed time interval, when two events are unlikely to occur simultaneously during a short time period. We write: $X \sim Poisson(\lambda)$. The parameter λ represents the rate, i.e., the average number of events per unit of time or area.

$$\text{p.m.f. } P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

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

- Discrete Uniform Random Variable

The Discrete Uniform random variable typically represents outcomes from a finite set that each can occur with equal probability. We write: $X \sim U(1,...,n)$

$$\text{p.m.f. } P(X = k) = \frac{1}{n}$$

$$E(X) = \frac{n+1}{2},\ Var(X) = \frac{n^2-1}{12}$$

- Discrete Uniform Random Variable

Continuous Random Variables & Distribution

- Exponential Random Variable

The Bernoulli random variable typically represents the random time between events that happen at a constant average rate. We use the notation $X \sim exp(\lambda)$.

Probability density function (p.d.f.): $f(x) = \lambda e^{-\lambda x}$,

Cumulative distribution function (c.d.f): $F(x) = 1 - \lambda e^{-\lambda x}$

The exponential distribution may be viewed as a continuous counterpart of the geometric distribution, which describes the number of Bernoulli trials necessary for the first success. The exponential distribution models the time until the first arrival of a Poisson processes. Specifically, suppose that $N(t) = \#$ of rare events in $[0,t] \sim Poisson(\lambda t)$. Let $X =$ time of first event (or $X =$ time of next event, or $X =$ time between two events). Then $X \sim exp(\lambda)$.

$$E(X) = \frac{1}{\lambda}, Var(X) = \frac{1}{\lambda^2}$$

Memoryless Property. An important property of the exponential distribution is:

$$P(X > s + t \,|\, X > t) = P(X > s)$$

This means, for example, that the conditional probability that we need to wait, 5 more seconds before the first arrival, given that the first arrival has not yet happened after 30 seconds, is the same as the initial probability that we need to wait more than 5 seconds for the first arrival i.e., $P(X > 35 \,|\, X > 30) = P(X > 5)$. The exponential distribution is the only memoryless continuous distribution.

- Uniform Random Variable

The Uniform random variable typically represents outcomes such that values in all intervals of the same length are equally probable. We use the notation $X \sim U(a, b)$, where $b > a$. The main application of uniform r.v.s is in the simulation of random variables.

Standard Uniform distribution: $X \sim U(0,1)$.

$$\text{p.d.f. } f(x) = \frac{1}{b-a}, \text{ c.d.f. } F(X) = \frac{x-a}{b-a}$$

$$E(X) = \frac{a+b}{2}, Var(X) = \frac{(b-a)^2}{12}$$

- Normal Random Variable

The Uniform random variable typically represents outcomes such that values in intervals close to the mean are more probable than values in intervals of the same length that are far from the mean. It is often called the bell curve because the graph of its probability density resembles a bell. The standard normal distribution is the normal distribution with a mean of zero and a standard deviation of one; its density and its distribution is denoted with $\phi(x)$ and $\Phi(x)$, respectively. We use the notation $X \sim N(\mu, \sigma^2)$.

$$\text{p.d.f. } \phi(x) = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}, \text{ c.d.f. } \Phi(X) = F(X) = \int_{-\infty}^{x} f(y)dy$$

Some properties of the Normal distribution

$P(\mu - \sigma < X < \mu + \sigma) = 0.68$

$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$

$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$

If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, then $Y \sim N(a\mu + b, a^2\sigma^2)$

If $X \sim N(\mu, \sigma^2)$ and $Y = (X - \mu)/\sigma$ (Z score), then $Y \sim N(0,1)$

Review of Statistics — Sampling

Sample

Population: A population contains all the items or individuals of interest that you seek to study.

Sample: A sample contains only a portion of a population of interest.

Parameter: A parameter summarizes the value of a population for a specific variable

Statistic: A statistic summarizes the value of a specific variable for sample data.

Types of sampling model:

Probability sample & **Non-probability sample**: Select items with or without knowing their probabilities of selection or distribution.

Non-probability samples:

Judgement sample: collect the opinions of preselected experts in the subject matter.

Convenience sample: select items that are easy, inexpensive or convenient to sample.

Probability samples:

Simple random sample: every item has the same chance of selection as every others, with or without replacement.

Systematic sample: partition the N items into n groups of k items, where $k = N/n$. Then randomly select items in each group.

Stratified sample: subdivide the N items into separate subpopulations or strata. A stratum is defined by some common characteristic, such as gender or year in school. Then randomly select items in each strata.

Cluster sample: divide the N items into clusters that contains several items. Clusters are often naturally occurring groups, such as counties or election districts.

Measures

Population Mean: $\mu = \dfrac{\sum_{i=1}^{N} X_i}{N}$

Sample Mean: $\bar{X} = \dfrac{\sum_{i=1}^{N} X_i}{n}$

Population Variance $\sigma^2 = \dfrac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$

Population Standard Deviation $\sigma = \sqrt{\dfrac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$

Sample Variance: $S^2 = \dfrac{\sum_{i=1}^{N} (X_i - \mu)^2}{n-1}$, degree of freedom $n-1$

Sample Standard Deviation: $S = \sqrt{\dfrac{\sum_{i=1}^{N} (X_i - \mu)^2}{n-1}}$

Median: $Median = \dfrac{n+1}{2}$ ranked value

Z score: $Z = \dfrac{X - \bar{X}}{S}$

Sample Covariance: $cov(X,Y) = \dfrac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$, degree of freedom $n-1$

Sample Coefficient of Correlation: $r = \dfrac{cov(X,Y)}{S_X S_Y}$, or we use the denotation ρ.

Important property: $-1 \leq \rho \leq 1$. (why?)

Sampling Distributions from Normally Distributed Population

In many applications, you want to make inferences that are based on statistics calculated from samples to estimate the values of population parameters.

Standard Error of the Mean: $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$

Z for the Sampling Distribution of the Mean: $Z = \dfrac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$

e.g. If you randomly select a sample of 25 boxes from the many thousands that are filled in a day, and a mean weight is computed for this sample. If the values in the population are normally distributed, the mean is 168 grams, the std is 15 grams. Calculate the probability that the sample mean below 365 grams.

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{365 - 368}{\frac{15}{\sqrt{25}}} = \frac{-3}{3} = -1$$

The are corresponding to $Z = -1$ in table is 0.1587.

e.g. If you select a sample of 100 boxes, what is the prob. that the sample mean is below 365 grams?

$Z = \dfrac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \dfrac{365 - 368}{\frac{15}{\sqrt{100}}} = \dfrac{-3}{1.5} = -2$, according to the table, the area less than

$Z = -2$ is 0.0228.

Sample Proportion: $p = \dfrac{X}{n} = \dfrac{\text{Number of items having the characteristic of interest}}{\text{Sample size}}$

Standard Error of the Proportion: $\sigma_p = \sqrt{\dfrac{\pi(1-\pi)}{n}}$, $\pi -$ population proportion

Z for the Sampling Distribution of the Proportion: $Z = \dfrac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$

You can use the normal distribution to approximate the binomial distribution when $n\pi$ and $n(1-\pi)$ are each at least 5.

e.g. 46% of American workers said that they work during nonbusiness hours. Suppose you select a random sample of 200 American workers and you want to determine the probability that more than 50% of them stated that they worked during nonbusiness hours.

In this case, $n\pi$ and $n(1-\pi)$ are larger than 5, the sample size is large enough to assume that the sampling distribution of the proportion is approximately normally distributed.

$Z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.50-0.46}{\sqrt{\frac{0.46\times 0.54}{200}}} = \frac{0.04}{0.0352} = 1.14$, according to the table, the area under the normal curve greater than 1.14 is 1-0.8729=0.1271. Therefore, if the population proportion is 0.46, the prob. is 12.71% that more than 50% of the 200 workers in the sample will say that they work during non-business hours.

Central Limit Theorem

As the sample size gets large enough, the sampling distribution of the mean is approximately normally distributed. This is true regardless of the shape of the distribution of the individual values in the population.

Review of Statistics — Confidence Interval

A point estimate is the value of a single sample statistic, such as a sample mean.
A confidence interval estimate is a range of numbers, called an interval, constructed around the point estimate.
You might find that a 95% confidence interval for the mean GPA at your university is $31.5 \leq \mu \leq 3.25$, you can interpret this interval estimate by stating that you are 95% confident that the interval states that the mean GPA at your university is between 3.15 and 3.25, [3.15,3.25] is an interval that includes the population mean.

Estimate For the Mean (σ known)

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or $\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$Z_{\alpha/2}$ is the value for an upper-tail probability of $\alpha/2$ from the standardized normal distribution.

$Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is the sampling error.

The value of $Z_{\alpha/2}$ needed for construction a confidence interval is called the critical value for the distribution. 95% confidence corresponds to an α value of 0.05. The critical Z value corresponding to a cumulative area of 0.975 is 1.96 because there is 0.025 in the upper tail of the distribution, and the cumulative area less than Z=1.96 is 0.975.

There is a different critical value for each level of confidence, $1 - \alpha$. A level of confidence of 95% leads to a Z value of 1.96. 99% confidence corresponds to an α value of 0.01. The Z value is approximately 2.58 because the upper-tail area is 0.005 and the cumulative area less than Z=2.58 is 0.995.

e.g. To determine whether the mean weight is consistent with the expected amount of 368 grams, managers periodically select a random sample of 100 filled boxes from the large number of boxes filled. Past experience states that the std of the fill amount is 15

grams. One random sample of 100 filled boxes they selected has a sample mean of 369.27 grams. Construct a 95% confidence interval estimate of the mean fill amount.

Using $Z_{\alpha/2} = 1.96$ for 95% confidence.

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 369.27 \pm (1.96)\frac{15}{100} = 369.27 \pm 2.94$$

$$366.33 \leq \mu \leq 372.21$$

Thus, with 95% confidence, the population mean is between 366.33 and 372.21 grams. Because the interval includes 368, the value indicating that the filling process is working properly, there is no evidence to suggest that anything is wrong with the filling process.

e.g. Construct a 99% confidence interval estimate for the population mean fill amount.

Using $Z_{\alpha/2} = 2.58$ for 99% confidence.

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 369.27 \pm (2.58)\frac{15}{100} = 369.27 \pm 3.87$$

$$365.40 \leq \mu \leq 373.14$$

Again, 368 is included within this wider interval, there is no evidence to suggest that anything is wrong with the filling process.