STAT 231 Notes

Angel Zhang

Winter 2022

1.2 Data Collection

Variate

A **variate** is a characteristic of a unit.

Continuous Variate

Height, weight and age are examples of continuous variates.

Discrete variate

The number of defective smartphones and the number of aphids on a tree are examples of discrete variates.

Categorical Variates

- **Categorical variates** do not take on numerical values.
- Examples are hair colour, university program or marital status of a person (the unit).

Ordinal Variate

- If a categorical variate has a natural ordering, then it is called an **ordinal variate**.
- Example: The size of a unit is an ordinal variate if the categories for size are: large, medium, and small.

Attribute

- An **attribute** of a population or process is a *function* of the variates over the population or process.
- Examples of attributes are average, variability and proportion.

Types of Empirical Studies

1. **Sample surveys:** Information about the population is obtained by selecting a representative sample of units from the population and determining the variates of interest for each unit in the sample.
2. **Observational studies:** Data are collected about a population or process without any attempt to change the value of one or more variates for the sampled units. A distinct between a sample survey and an observational study is that for observational studies the population of interest is usually infinite or conceptual.
3. **Experimental studies:** The experimenter intervenes and changes or sets the values of one or more variates for the units in the sample.

Note: These three types of emprical studies are *not* mutually exclusive, and many studies involve aspects of all of them.

1.3 Data Summaries

Measures of location

The sample mean, median and mode describe the center of the distribution of variate values in a data set. The units for mean, median and mode are the same as for the original variate

Sample Mean

The **sample mean** is:

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

Sample Median

The **sample median** $\hat{m}$, is the middle value when n is odd and the sample is ordered from the smallest to largest, or the average of the two middle values when n is even.

Sample Mode

The **sample mode** is the value of y which appears in the sample with the highest frequency (not necessarily unique).

Measures of dispersion or variability

The sample variance and sample standard deviation measure the variability or spread of the variate values in a data set. The units for standard deviation. range, and IQR are the same as for the original variate.

Sample Variance

The **sample variance** is:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]$$

The **sample standard deviation** is

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

Range

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

Measures of Shape

Measures of shape generally indicate how the data, in terms of a relative frequency histogram, differ from the Normal bell-shaped curve. Sample skewness and sample kurtosis have no units.

Sample Skewness

- The **sample skewness** is:

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right]^{3/2}}$$

- The sample skewness is a measure of the (lack of) symmetry in the data.
- If the relative frequency histogram of the data is approximately symmetric, then the sample skewness will be approximately zero.
- If the relative frequency histogram of the data has a long right tail, then the sample skewness will be positive.
- If the relative frequency histogram of the data has a long left tail, then the sample skewness will be negative.

Sample Kurtosis

- The **sample kurtosis** is:

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right]^2}$$

- The sample kurtosis measures the heaviness of the tails and the peakedness of the data relative to data that are Normally distributed.
- Since the term $(y_i - \overline{y})^4$ is always positive, the sample kurtosis is always positive.
- If the sample kurtosis is greater than 3 then this indicates heavier tails (and more peaked center) than data that are Normal distributed.
- For data that arise from a model with no tails, for example the Uniform distribution, the sample kurtosis will be less than 3.

Sample Quantile and Percentile

For $0 < p < 1$, the p**th (sample) quantile** (also called the $100p$**th (sample) percentile**), is a value, called $q(p)$, determined as follows:

- Let $k = (n+1)p$ where n is the sample size.
- If k is an integer and $1 \leq k \leq n$, then $q(p) = y_{(k)}$.
- If k is not an integer but $1 < k < n$ then determine the closet integer j such that $j < k < j+1$ and then $q(p) = \frac{1}{2}\Big[y_{(j)} + y_(j+1)\Big]$

Lower Quartile, Median, Upper Quartile

The quantiles $q(0.25)$, $q(0.5)$ and $q(0.75)$ are called the **lower** or **first quartile**, the **median**, and the **upper** or **third quartile** respectively.

Interquantile Range

$$IQR = q(0.75) - q(0.25)$$

Five Number Summary

The **five number summary** of a data set consists of the smallest observation, the lower quartile, the median, the upper quartile and the largest observation.

Sample Correlation

The **sample correlation**, denoted by r, for data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} y_i \right)^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)$$

- The sample correlation, which takes on values between -1 and 1, is a measure of the *linear relationship* between the two variates x and y.
- If the value of r is close to 1, then there is a strong positive linear relationship.
- If the value of r is close to -1, then there is a strong negative linear relationship.
- If the value of r is close to 0, then there is no linear relationship.

Empirical Cumulative Distribution Function

For a data set $\{y_1, y_2, \ldots, y_n\}$, the empirical cumulative distribution function or e.c.d.f is defined by

$$\hat{F}(y) = \frac{\text{number of values in the data set } \{y_1, y_2, \ldots, y_n\} \text{ which are } \leq \text{y}}{n} \text{ for all } y \in \mathbb{R}$$

1.5 Data Analysis and Statistical Inference

Descriptive Statistics

Descriptive statistics is the portrayal of the data, or parts of it, in numerical and graphical ways so as to show features of interest.

Statistical Inference

Statistical inference is using the data obtained in the study of a process or population to draw general conclusions about the process or population itself.

2.2 Maxiumum Likelihood Estimation

Point Estimate

- A **point estimate** of a parameter is the value of a function of the observed data $y_1, y_2, \ldots, y_n$ and other known quantities such as the sample size n.
- We use $\hat{\theta}$ to denote an estimate of the parameter θ.
- Note that $\hat{\theta} = \hat{\theta}(y_1, y_2, \ldots, y_n) = \hat{\theta}(\mathbf{y})$ depends on the sample $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ drawn.
- A function of the data which does not involve any unknown quantities such as unknown parameters is called a **statistic**.
- A point estimate is a statistic.

Likelihood Function

- Let the discrete (vector) random variable $\mathbf{Y}$ represent potential data that will be used to estimate θ.
- Let $\mathbf{y}$ represent the actual observed data that are obtained in a specific application.
- It is usually assumed here that the data set consists of measurements on a *random sample* of units from a population or process.
- The **likelihood function** for θ is defined as:

$$L(\theta) = L(\theta; \mathbf{y}) = P(\mathbf{Y} = \mathbf{y}; \theta) \text{ for } \theta \in \Omega$$

  where the parameter space Ω is *the set of possible values* of θ.
- The likelihood function is the probability that we observe the data $\mathbf{y}$, considered as a function of the parameter θ.

Maximum Likelihood Estimate

- The value of θ which maximizes $L(\theta)$ for given data $\mathbf{y}$ is called the **maximum likelihood estimate (m.l. estimate)** of θ.
- It is the value of θ which *maximizes* the probability of observing the data $\mathbf{y}$.
- This value is denoted $\hat{\theta}$.

Relative Likelihood Function

The **relative likelihood function** is defined as

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \text{ for } \theta \in \Omega$$

Note that $0 \leq R(\theta) \leq 1$ for all $\theta \in \Omega$, and that $R(\hat{\theta}) = 1$.

Log Likelihood Function

The **log likelihood function** is defined as

$$l(\theta) = \ln L(\theta) = \log L(\theta) \text{ for } \theta \in \Omega$$

Likelihood Function for a Random Variable

- In many applications the data $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ are *independent and identically distributed* random variables each with probability function $f(y; \theta), \theta \in \Omega$.
- We refer to $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ as a random sample from the distribution $f(y; \theta)$.
- In this case the observed data are $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ and

$$L(\theta) = L(\mathbf{y}; \theta) = \prod_{i=1}^{n} f(y_i; \theta) \text{ for } \theta \in \Omega$$

- Note that if $Y_1, Y_2, \ldots, Y_n$ are independent random variables, then their joint probability function is the *product* of their individual probability functions.

2.3 Likelihood Functions for Continuous Distributions

Suppose $y_1, y_2, \ldots, y_n$ are the observations from a random sample from the distribution with probability density function $f(y; \theta)$ which have been rounded to the nearest Δ which is assume to be small. Then

$$P(\mathbf{Y} = \mathbf{y}; \theta) \approx \prod_{i=1}^{n} \Delta f(y_i; \theta) = \Delta^n \prod_{i=1}^{n} f(y_i; \theta)$$

Likelihood Function

If $y_1, y_2, \ldots, y_n$ are the observed values of a random sample from a distribution with probability density function $f(y; \theta)$, then the **likelihood function** is defined as

$$L(\theta) = L(\theta; \mathbf{y}) = \prod_{i=1}^{n} f(y_i; \theta) \text{ for } \theta \in \Omega$$

2.4 Likelihood Functions For Multinomial Models

The *likelihood function* for $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ based on data $y_1, y_2, \ldots, y_k$ is given by

$$L(\boldsymbol{\theta}) = L(\theta_1, \theta_2, \ldots, \theta_k) = \frac{n!}{y_1! y_2! \cdots y_k!} \prod_{i=1}^{k} \theta_i^{y_i}$$

or more simply

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{k} \theta_i^{y_i}$$

The *log likelihood function* is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{k} y_i \log \theta_i$$

If y_i represents the number of times outcome i occurred in n "trials", $i = 1, 2, \ldots, k$, then

$$\hat{\theta}_i = \frac{y_i}{n} \text{ for } i = 1, 2, \ldots, k$$

are the *maximum likelihood estimates* of $\theta_1, \theta_2, \ldots, \theta_k$

Theorem 16 (Invariance Property of Maximum Likelihood Estimates)

If $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ is the *maximum likelihood estimate* of $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ then $g(\hat{\boldsymbol{\theta}})$ is the *maximum likelihood estimate* of $g(\boldsymbol{\theta})$

2.6 Checking the Model

Qqplots for checking the Gaussian model

- Suppose that we want to check if a $G(\mu, \sigma)$ model fits the set of data $\{y_1, y_2, \ldots, y_n\}$
- Let $\{y_{(1)}, y_{(2)}, \ldots, y_{(n)}\}$ represent the data ordered from smallest to largest.
- Let $Q(p)$ be the pth quantile for the $G(\mu, \sigma)$ distribution, and $q(p)$ be the $p$the sample quantile.
- If the Gaussian model is appropriate for the data then we would expect $Q\left(\frac{i}{n+1}\right)$ to be close in value to $q\left(\frac{i}{n+1}\right)$, for $i = 1, 2, \ldots, n$.
- If we plot the point $\left(Q\left(\frac{i}{n+1}\right), q\left(\frac{i}{n+1}\right)\right)$ for $i = 1, 2, \ldots, n$, then we should see a set of points that lie reasonably along a straight line.
- If μ and σ are unknown, let $Q_Z(p)$ be the pth quantile for the $G(0, 1)$ distribution. We have $Q(p) = \mu + \sigma Q_Z(p)$. Therefore if we plot the points $\left(Q_Z\left(\frac{i}{n+1}\right), q\left(\frac{i}{n+1}\right)\right)$ for $i = 1, 2, \ldots, n$, we should still see a set of points that lie reasonably along a straight line if a Gaussian model is reasonable for the data.

- The line in the qqplot is the line joining the lower and upper quartile of the empirical and Gaussian distribution, that is the line joining $(Q_Z(0.25), q(0.25))$ and $(Q_Z(0.75), q(0.75))$

3.1 Empirical Studies

PPDAC

- **Problem:** a clear statement of the study's objectives, usually involving one or more questions
- **Plan:** the procedures used to carry out the study including how the data will be collected
- **Data:** the physical collection of the data, as described in the Plan
- **Analysis:** the analysis of the data collected in light of the Problem and the Plan
- **Conclusion:** the conclusions that are drawn a bout the Problem and their limitations.

3.2 The Steps of PPDAC

Types of Problems

- **Descriptive:** The problem is to determine a particular attribute of a population or process.
- **Causative:** The problem is to determine the existence or non-zcexistence of a causal relationship between two variables.
- **Predictive:** The problem is to predict a future value for a variate of a unit to be selected from the process or population.

Target Population, Target Process

The **target population** or **target process** is the collection of units to which the experimenters conducting the empirical study which the conclusions to apply.

Study Population, Study Process

The **study population** or **study process** is the collection of units available to be included in the study. The study population is often but not always a subset of the target population.

Study Error

If the attributes in the study population/process differ from the attributes in the target population/process then the difference is called **study error**.

Sampling Protocol, Sample Size

The **sampling protocol** is the procedure used to select a sample of units from the study population/process. The number of units sampled is called the **sample size**.

Sample Error

If the attributes in the sample differ from the attributes in the study population/process the difference is called **sample error**.

Measurement Error

If the measured value and the true value of a variate are not identical the difference is called **measurement error**.

4.2 Estimators and Sampling Distributions

Point Estimator, Sampling Distribution

- A **(point) estimator** $\tilde{\theta}$ is a *random variable* which is a *function* $\tilde{\theta} = g(Y_1, Y_2, \ldots, Y_n)$ of the random variables $Y_1, Y_2, \ldots, Y_n$. The distribution of θ is called the **sampling distribution** of the estimator.
- If we know that sampling distribution of the estimator $\tilde{\theta}$ then we can use it to quantify the uncertainty in an estimate $\hat{\theta}$, that is, we can determine the probability that the estimator $\tilde{\theta}$ is "close" to the true but unknown value of θ.

4.3 Interval Estimation Using the Likelihood Function

100p% Likelihood Interval

A **100p% likelihood interval** for θ is the set $\{\theta : R(\theta) \geq p\}$, where $R(\theta)$ is the relative likelihood function.

Guidelines for Interpreting Likelihood Intervals

- Values of θ inside a 50% likelihood interval are *very plausible* in light of the observed data.
- Values of θ inside a 10% likelihood interval are *plausible* in light of the observed data.
- Values of θ outside a 10% likelihood interval are *implausible* in light of the observed data.
- Values of θ outside a 10% likelihood interval are *very implausible* in light of the observed data.

Log Relative Likelihood function

- The **log relative likelihood function** is

$$r(\theta) = \log(\theta) = \log\left[\frac{L(\theta)}{L(\hat{\theta})}\right] = l(\theta) - l(\hat{\theta})$$

  where $l(\theta) = \log L(\theta)$ is the log likelihood function.
- If the likelihood function $R(\theta)$ is unimodal, then the log likelihood function $r(\theta)$ is also unimodal.
- The log relative likelihood function can also be used to obtain a $100p\%$ likelihood interval since $R(\theta) \geq p$ if and only if $r(\theta) \geq \log p$.

4.4 Confidence Intervals and Pivotal Quantities

100p% Confidence Interval

- Suppose the interval estimator $[L(\mathbf{Y}, U(\mathbf{Y})]$ has the property that

$$P\{\theta \in [L(\mathbf{Y}, U(\mathbf{Y})]\} = P[L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})] = p$$

  is constructed for the parameter θ based on observed data $\mathbf{y}$. The interval estimate $[L(\mathbf{y}), U(\mathbf{y})]$ is called a **100p% confidence interval** for θ and p is called the **confidence coefficient**.
- $P\{\theta \in [L(\mathbf{Y}, U(\mathbf{Y})]$ is called the **coverage probability** of the interval estimator $[L(\mathbf{Y}, U(\mathbf{Y})]$.
- $P\{\theta \in [L(\mathbf{y}, U(\mathbf{y})]\} = p$ is an *incorrect* statement. The parameter θ is a constant, not a random variable.

Pivotal Quantity

- A **pivotal quantity** $Q = Q(\mathbf{Y}; \theta)$ is a function of the data $\mathbf{Y}$ and unknown parameter θ such that the distribution of the random variable Q is *fully known.*
- That is, probability statements such as $P(Q \leq b)$ and $P(Q \geq a)$ depend on a and b, but not on θ or any other unknown information.

4.5 The Chi-squared Distribution and t Distributions

Chi-squared Distribution

- The $\chi^2(k)$ distribution is a continuous family of distributions on $(0, \infty)$ with probability density function of the form

$$f(x, k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

  where $k \in \{1, 2, \dots\}$ is a parameter of the distribution.
- We write $X \approx \chi^2(k)$. The parameter k is referred to as the "degrees of freedom" (d.f.) parameter.
- For $k = 2$, the probability density function is the Exponential(2) probability density function.
- For $k > 2$, the probability density function is unimodal with maximum value at $x = k - 2$.
- For values of $k \geq 30$, the probability density function resembles that of a $N(k, 2k)$ probability density function.

Expected Value and Variance for a Chi-squared Distribution

If $X \sim \chi^2(k)$, then

$$E(X) = k$$

and

$$Var(X) = 2k$$

Theorem 29

Let $W_1, W_2, \dots, W_n$ be independent random variables with $W_i \sim \chi^2(k_i)$. Then

$$S = \sum_{i=1}^{n} W_i \sim \chi^2\Big(\sum_{i=1}^{n} k_i\Big)$$

Theorem 30

If $Z \sim G(0,1)$ then the distribution of $W = Z^2$ is $\chi^2(1)$.

Corollary 31

If $Z_1, Z_2, \ldots, Z_n$ are mutually independent $G(0,1)$ random variables and $S = \sum_{i=1}^{n} Z_i^2$, then $S \sim \chi^2(n)$.

Useful Results

1. If $W \sim \chi^2(1)$, then $P(W \geq w) = 2[1 - P(Z \leq \sqrt{w})]$ where $Z \sim G(0,1)$
2. If $W \sim \chi^2(2)$, then $W \sim Exponential(2)$ and $P(W \geq w) = e^{-w/2}$

Student's t Distribution

- **Student's t distribution** (or more simply the t **distribution**) has probability density function
$$f(t;k) = c_k\left(1 + \frac{t^2}{k}\right)^{-(k+1)/2}$$
where the constant c_k is given by
$$c_k = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}$$
- The parameter k is called the **degrees of freedom**.
- We write $T \sim t(k)$ to indicate that the random variable T has a T distribution with k degrees of freedom.

Properties of the t Distribution

- The t probability density function is similar to that of the $G(0,1)$ distribution in several respects.
- It is symmetric about the origin and unimodal.
- For large values of k, the graph of the probability density function $f(t;k)$ is indistinguishable from that of the $G(0,1)$ probability function.
- For small k, the t probability density functions has more area in the extreme left and right tails.

Theorem 32

Suppose $Z \sim G(0,1)$ and $U \sim \chi^2(k)$ independently. Let

$$T = \frac{Z}{\sqrt{U/k}}$$

Then T has a Student's t distribution with k degrees of freedom

4.6 Likelihood-Based Confidence Intervals

Likelihood Ratio Statistic

Define the random variable $\Lambda(\theta)$

$$\Lambda(\theta) = -2\log\left[\frac{L(\theta)}{L(\tilde{\theta})}\right]$$

where L is the maximum likelihood function and $\hat{\theta}$ is the maximum likelihood estimator The random variable $\Lambda(\theta)$ is called the **likelihood ratio statistic**. It is an *asymptotic pivotal quantity.*

Theorem 33

If $L(\theta)$ is based on $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, a random sample of size n, and if θ is the true value of the scalar parameter, then (under mild mathematical conditions) the distribution of $\Lambda(\theta)$ converges to a $\chi^2(1)$ distribution as $n \to \infty$.

Theorem 33 Notes

This theorem means that $\Lambda(\theta)$ can be used as a pivotal quantity for sufficiently large n in order to obtain approximate confidence intervals for θ.

Theorem 34

A $100p\%$ likelihood interval is an approximate $100q\%$ confidence interval where $q = 2P(Z \leq \sqrt{-2\log p}) - 1$ and $Z \sim N(0,1)$.

Theorem 35

If a is a value such that $p = 2P(Z \leq a) - 1$ where $Z \sim N(0,1)$, then the likelihood interval $\{\theta : R(\theta) \geq e^{-a^2/2}\}$ is an approximate $100p\%$ confidence interval.

4.7 Confidence Intervals for Parameters in the $G(\mu, \sigma)$ Model

Theorem 36

Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from the $G(\mu, \sigma)$ distribution with sample mean $\overline{Y}$ and sample variance S^2. Then

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

The random variable T is a pivotal quantity since it is a function of the data $Y_1, Y_2, \ldots, Y_n$ and the unknown parameter μ whose distribution $t(n-1)$ is completely known.

Theorem 37

Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from the $G(\mu, \sigma)$ distribution with sample variance S^2. Then

$$U = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{\sigma}\right)^2 \sim \chi^2(n-1)$$

Prediction Interval for a Future Observation

- Suppose that $y_1, y_2, \ldots, y_n$ is an observed random sample from a $G(\mu, \sigma)$ population and that Y is a new observation which is to be drawn at random from the same $G(\mu, \sigma)$ population. We want to estimate Y and obtain an interval of values for Y. Then
$$\frac{\frac{Y-\overline{Y}}{\sigma\sqrt{1+\frac{1}{n}}}}{\sqrt{S^2/\sigma^2}} = \frac{Y - \overline{Y}}{S\sqrt{1+\frac{1}{n}}} \sim t(n-1)$$
is a pivotal quantity which can be used to obtain an interval of values for Y.
- Let a be the value such that
$$P(-a \leq T \leq a) = p \text{ or } P(T \leq a) = \frac{1+p}{2}$$
where $T \sim t(n-1)$. Then
$$\left[\overline{y} - as\sqrt{1+\frac{1}{n}}, \overline{y} + as\sqrt{1+\frac{1}{n}}\right]$$
is an interval of values f or the future observation Y with confidence coefficient p.
- The interval is called a $100p\%$ **prediction interval** instead of a confidence interval since Y is not a parameter but a *random variable*.