STAT 241 Personal Notes

* These notes are strictly my own interpretation of the course materials.

Marcus Chan

Taught by Michael Wallace

UW CS '25

Chapter 1: Introduction to Statistical Science

C # S #
chapter number, slide number

Statistical science is the science of "empirical studies".

EMPIRICAL STUDY (C1S24)

An "empirical study" is one where we learn by observation and/or experimentation.

Note these involve uncertainty - repeated experiments generate different results.

But we model these uncertainties using probability models.

UNIT (C1S25)

A "unit" is an individual which we can take measurement(s).

POPULATION (C1S26)

A "population" is a collection of units.

eg - all current UW undergrad students
- all donuts in Tim Hortons right now

*note: we need to be precise when defining populations, or any other terms!
eg if we said "all UW students" this is ambiguous, since it might include grads, alumni, etc

PROCESS (C1S27)

A "process" is a system by which units are produced.

eg - hits on a particular website are units in a process
- claims made by insurance policy holders are units in a process

Note that although populations & processes are collections of units:

① Populations are "static" (defined at one point in time), but

② Processes usually occur over time.

VARIATES (C1S32)

"Variates" are characteristics of the units.

* we usually represent these by letters x, y & z.

CONTINUOUS VARIATES (C1S33)

"Continuous variates" are those that can be measured (at least theoretically) to an infinite degree of accuracy.

eg height, weight, lifetime of a fuse, etc

DISCRETE VARIATES (C1S33)

"Discrete variates" are those that can only take finitely or countably many values.

eg # of car accidents on a certain stretch of highway / yr, etc.

Note that depending on how we measure a continuous variate, it may become discrete.

eg if we measure weight w/ a scale that only goes to 2dp, the resulting variate is discrete!

Ultimately the distinction affects

① our assumptions of the data; and

② the probability models we use.

- for discrete variates, we usually use discrete prob models (eg Poisson)
- for cts variates, we usually use cts prob models (eg Gaussian)
- but there are exceptions. (C1S43)

CATEGORICAL VARIATES (C1S35)

"Categorical variates" are those where the units fall into non-numeric categories, without any implied order.

eg hair color, university program

ORDINAL VARIATES (C1S35)

"Ordinal variates" are those where an ordering is implied, but not necessarily from a numeric measure.

eg strongly disagree, ..., strongly agree;
small, medium, large;
etc

COMPLEX VARIATES (C1S37)

"Complex variates" are those that are more unusual, and don't fall neatly into the other variate types.

eg open-ended responses to a survey question

We usually need processing to convert these into one of the other types.

eg text processing to convert a tweet's content into "positive", "negative" or "neutral"

ATTRIBUTES [OF A POPULATION/PROCESS] (C1S48)

"Attributes" of a population/process are functions of a variate which is defined for all units in said population/process.

eg (STAT 231 asmts) – mean # of completed asmts
– prop. of asmts subbed in last 24 hrs
(KW Humane Society) – prop. of dogs that arrive in good health
– mean # of owners of dogs in their care

TYPES OF EMPIRICAL STUDIES (C1S50)

SAMPLE SURVEY (C1S52)

A "sample survey" is where information is obtained about a finite population by
① selecting a "representative" sample of units from the population; and
② determining the variates of interest for each unit in the sample.

eg – poll to predict who will win an election
– survey of potential consumers to compare products & state their preference (eg Coke vs Pepsi)

OBSERVATIONAL STUDY (C1S53)

An "observational study" is where information about a population/process is collected without any change to the sampled units' variates.

eg a study of blood alcohol levels for students at a 8:30am Mon lecture

Usually, the following are true:

Observational	Survey
① Pop^n of interest is infinite/conceptual	Pop^n is finite/real
② Data collected routinely over time	Data collected once
③ More passive (sit and see)	More "aggressive" (specific questions asked)

* but these are just guidelines – there are exceptions. (C1S55)

EXPERIMENTAL STUDY (C1S54)

An "experimental study" is one where the experimenter intervenes and modifies some of the variates for the units in a study.

eg same example as above, but some students are warned beforehand, whereas some are not.

DATA SUMMARIES (CIS56)

These are used for
① the estimation of attributes; and
② checking fit for a model.

MEASURES OF CENTRAL TENDENCY/ LOCATION (CIS58)

We usually represent our data using the notation $\{y_1, \dots, y_n\}$, where each $y_i \in \mathbb{R}$ and "n" is called the "sample size".

We also use lower-case for constants, and upper-case for random variables.

ORDERED SAMPLE / ORDER STATISTICS (CIS59)

We call the "ordered sample" or "order statistics" of the data to be

$$y_{(1)}, \dots, y_{(n)},$$

where $y_{(1)} \leq \dots \leq y_{(n)}$, $y_{(1)} = \min\{y_1, \dots, y_n\}$ & $y_{(n)} = \max\{y_1, \dots, y_n\}$.

SAMPLE MEAN/AVERAGE: $\bar{y}$ (CIS58)

The "sample mean", denoted by "$\bar{y}$", is equal to

$$\bar{y} := \frac{y_1 + \dots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

* the keyword "sample" is important!

SAMPLE MEDIAN: $\hat{m}$ (CIS59)

The "sample median", denoted as "$\hat{m}$", is defined by

$$\hat{m} := \begin{cases} y_{(\frac{n+1}{2})}, & n \text{ is odd} \\ \frac{1}{2}\left(y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)}\right), & n \text{ is even} \end{cases}$$

Note that
① In symmetrical distributions, $\bar{y} \approx \hat{m}$; but
② In skewed distributions, $\bar{y} \not\approx \hat{m}$ (there may be a significant gap between them). (CIS66)

SAMPLE MODE (CIS61)

The "sample mode" is just the most common value(s) in a set of data.

In this case, the "sample modal class" is the group/class with the highest frequency.

MEASURES OF VARIABILITY/ DISPERSION (CIS67)

"Measures of variability" convey how "spread out" the data is.

ROBUST [MEASURE] (CIS80)

We say a measure is "robust" if it is not (significantly) affected by extreme values.

eg IQR is robust, range is not

SAMPLE VARIANCE & STANDARD DEVIATION: s^2, s (CIS69)

We define the "sample variance", denoted "s^2", of the data $\{y_1, \dots, y_n\}$ to be

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

The "sample standard deviation", denoted "s", is just the square root of the sample variance.

"68-95" RULE FOR GAUSSIAN ESTIMATION (CIS70)

Suppose the data $\{y_i\}$ is from a Gaussian distribution $G(\bar{y}, s)$. * $G(\mu, \sigma) = N(\mu, \sigma^2)$

Then necessarily
① 68% of the sample lies in $[\bar{y}-s, \bar{y}+s]$; and
② 95% of the sample lies in $[\bar{y}-2s, \bar{y}+2s]$.

* this can be verified in R using the code

```
> pnorm(1) - pnorm(-1)
> pnorm(2) - pnorm(-2)
```

RANGE (CIS73)

The "range" is defined as

$$\text{range} = y_{(n)} - y_{(1)}.$$

* the range is very susceptible to outliers!

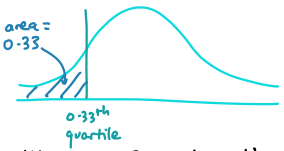
QUANTILES & PERCENTILES (CIS74)

The "p^{th} quartile", also called the "$(100p)^{th}$ percentile", is the value such that a fraction p of the data fall at or below said value.

* the median is the 0.5^{th} quartile / 50^{th} percentile.

In other words, the p^{th} quartile of a distribution is the value q such that

$$p = \int_{-\infty}^{q} f(y)\,dy.$$



We can calculate quartiles in R using the code

```
> quantile(c(y_1, ..., y_n), p)
```

QUARTILES: $q(0.25)$, $\hat{m}$, $q(0.75)$ (CIS79)

The "lower quartile", or "first quartile", denoted by $q(0.25)$, is the 25^{th} percentile.

The "upper quartile", or "third quartile", denoted by $q(0.75)$, is the 75^{th} percentile.

The "second quartile" is just the median $\hat{m}$.

INTERQUARTILE RANGE / IQR (CIS80)

The "interquartile range" is defined as

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

* IQR is robust — it is not affected by extreme values.

* if considering discrete data, the interpretation of IQRs can vary depending whether we consider the "interval" from $q(0.25)$ to $q(0.75)$ to be open, semi-open or closed.

MEASURES OF SHAPE (CIS84)

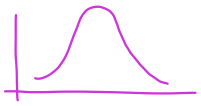
SAMPLE SKEWNESS (CIS88)

"Sample skewness" measures the asymmetry of the data, and is equal to

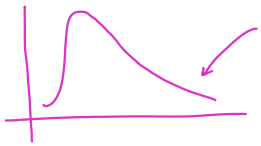
$$\text{sample skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{3/2}}$$

Interpretation of sample skewness's value:

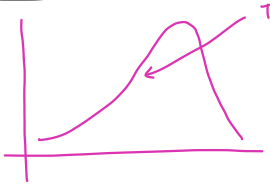
① If ss = 0 ⇒ distribution is symmetric; eg Gaussian, uniform



② If ss > 0 ⇒ distribution is positively skewed / skewed to the right;



③ If ss < 0 ⇒ distribution is negatively skewed / skewed to the left.



SAMPLE KURTOSIS (CIS96)

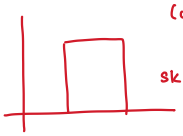
"Sample kurtosis" measures whether data is concentrated in the central "peak" or in the tails, and is calculated by

$$\text{sample kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^2}.$$

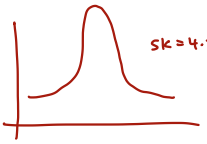
Interpretation of sample kurtosis' value:

① sk = 3 ⇒ distribution looks "Gaussian" (bell-shaped);

② sk < 3 ⇒ distribution has shorter tails (more concentrated in the peak)



③ sk > 3 ⇒ distribution has longer tails (less concentrated in the peak)



ASSUMING A MODEL IS GAUSSIAN (CIS102)

To assume data can be reasonably modelled by a Gaussian distribution, we must have the following:

① The sample mean & median should be approximately equal;

② The sample skewness should be close to 0;

③ The sample kurtosis should be close to 3; and

④ ~95% of the observations should lie in the interval $[\bar{y}-2s,\ \bar{y}+2s]$.

IN STATISTICS, WE DON'T PROVE THINGS! (CIS103)

In statistics, we don't prove assumptions are true, but instead find evidence against an assumption.

① If there is sufficient evidence against the assumption, then we say the data is "not consistent" with said assumption.

② Otherwise, we say the data is "consistent" with the assumption.

FIVE NUMBER SUMMARY (CIS108)

The "five number summary" for a set of data is

① The minimum value $y_{(1)}$;

② $q(0.25)$;

③ $q(0.5)$;

④ $q(0.75)$; &

⑤ The maximum value $y_{(n)}$.

In R, we can find the five number summary via the code

```
> summary(...)
```

GRAPHICAL SUMMARIES (CIS112)

When displaying graphs, note that

① All graphs should be displayed at an appropriate size;

② Graphics should have clear titles which are fairly self-explanatory;

③ Axes should be labelled & units given where appropriate;

④ Choice of scales should be made with care; and

* ⑤ Graphics should not be used without thought, especially if there are better ways of displaying the information.

HISTOGRAMS (CIS116)

Essentially, histograms create a graphical summary of our data that we can use to compare with a pdf for crvs, or a pmf for a drv.

Let our data be $y_1, \dots, y_n$.

Partition the range of the y's into k non-overlapping intervals

$I_j = [a_{j-1}, a_j)$, $j = 1, 2, \dots, k$.

Let f_j = # of values from $\{y_1, \dots, y_n\}$ in I_j.

The f_j's are called the "observed frequencies".

Then, draw a rectangle above each of the intervals with height proportional to the observed/relative frequency.

(standard histogram) (rel. freq. histogram)

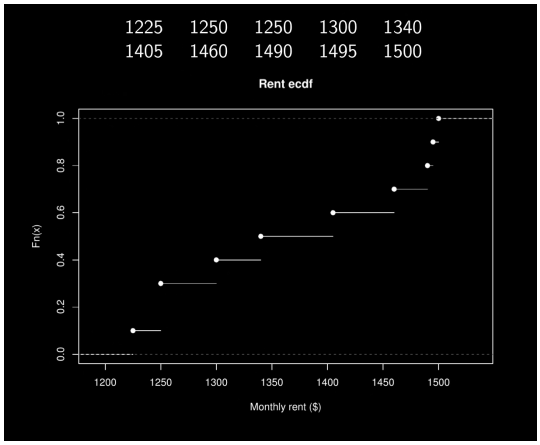
$$\text{height} = \frac{f_j/n}{a_j - a_{j-1}}$$

EMPIRICAL CDF (CIS124)

An "empirical cdf" lets us compare the distribution of a dataset with a cdf of a random variable.

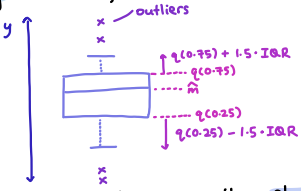
Mathematically, the empirical cdf is defined by

$$\hat{F}(y) = \frac{\#\text{ of values in } \{y_1, \dots, y_n\} \text{ which are } \le y}{n} \quad \forall y \in \mathbb{R}.$$

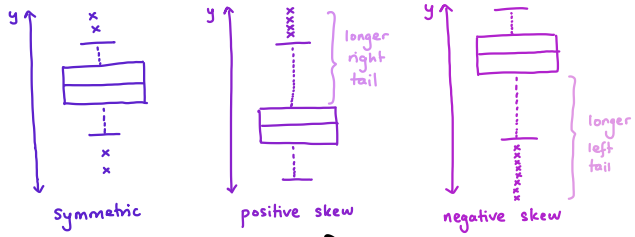


BOX-PLOT (CIS139)

"Box-plots" give a graphical summary of the shape of a dataset's distribution in a similar way to the five number summary.

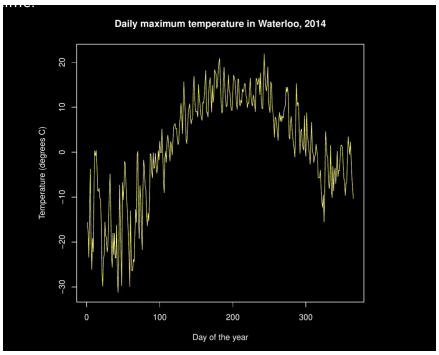


Box-plots can also show the skewness of a distribution:



RUN CHART (CIS154)

A "run-chart" gives a graphical summary of data which are varying over time.



SCATTERPLOTS (CIS157)

BIVARIATE VS UNIVARIATE DATA (CIS157)

1. "Bivariate data" is of the form $\{(x_1,y_1), \dots, (x_n,y_n)\}$, where $x_i, y_i \in \mathbb{R}$ $\forall i$.
2. In constrast, "univariate data" is of the form $\{y_1, \dots, y_n\}$ for $y_i \in \mathbb{R}$.

SCATTER-PLOT (CIS158)

A "scatter-plot" for bivariate data is simply a plot of the (x_i,y_i)'s.

SAMPLE CORRELATION: r (CIS162)

1. The "sample correlation", denoted "r", gives us a numerical summary of a bivariate dataset.
2. For data $\{(x_1,y_1), \dots, (x_n,y_n)\}$,

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}.$$

3. In particular, $r \in [-1,1]$, and measures the linear relationship between x & y.
   1. If $r \approx -1$, we say there is a "strong negative linear relationship" between the two variates.
   2. If $r \approx +1$, we say there is a "strong positive linear relationship" between the two variates.

* $|r| \approx 1$ does not imply a causal relationship (correlation does not imply causation!)

   3. If $r \approx 0$, we say there is "no linear relationship" between the two variates.

* $r \approx 0$ does not imply x & y are unrelated — it just implies they are not linearly correlated.

eg

Here $r=0$ but obviously the data is related quadratically.

RESPONSE & EXPLANATORY VARIATES (CIS171)

In an experiment, the "explanatory variate" is the variate that attempts to explain / determine the distribution of the "response variate".

* explanatory variate = "independent" variable
response variate = "dependent" variable.

BIVARIATE CATEGORICAL DATA (CIS172)

1. We use the following survey as motivation:
   1. Hometown in Canada, like hockey
   2. Hometown not in Canada, like hockey
   3. Hometown in Canada, dislike hockey
   4. Hometown not in Canada, dislike hockey
2. Sample results from that survey:

	Canada ✓	Canada ✗	Σ
Hockey ✓	33	9	42
Hockey ✗	22	43	65
Σ	55	52	107

RELATIVE RISK (CIS176)

Let $A \subseteq X$ & $B \subseteq Y$ be events in bivariate data "$X \times Y$". Then the "relative risk" of "A with B" is equal to

$$\text{relative risk} = \frac{P(AB \mid A)}{P(A\neg B \mid A)}.$$

eg in the survey above,

$$\text{relative risk of liking hockey among those w/ a Canadian hometown} = \frac{\text{prop. of Canada hometown who like hockey}}{\text{prop. of non-Canada hometown who dislike hockey}} = \frac{(33/55)}{(9/52)} = 3.467$$

DATA ANALYSIS & STATISTICAL INFERENCE (CIS182)

DESCRIPTIVE STATISTICS (CIS183)

"Descriptive statistics" is the portrayal of data (or parts of it) in numerical & graphical ways.

* all our previous work falls under this category!

STATISTICAL INFERENCE (CIS184)

"Statistical inference" is the process of drawing general conclusions for a population/process based off of data obtained in a study about said population/process.

eg "based off my sample, I expect 90% of asmts this term to be submitted within the final 24 hrs of the deadline"

INDUCTIVE VS DEDUCTIVE REASONING (CIS185)

"Inductive reasoning" occurs when we reason from the "specific" (observed data about a sample) to the "general" (the target population/process).

In contrast, "deductive reasoning" occurs when we use general results to prove theorems.

* proof by induction = deductive reasoning!

ESTIMATION PROBLEMS (CIS187)

In "estimation problems", we are concerned about estimating one or more attributes of a population/process.

eg
- estimate the prop. of STAT 231 students who like poutine
- "fitting" a probability distribution for a process.

HYPOTHESIS TESTING PROBLEMS (CIS188)

In a "hypothesis testing problem", we use the data to assess the truth of some question/hypothesis.

eg is it true a higher proportion of math majors than CS majors like poutine?

PREDICTION PROBLEMS (CIS189)

In a "prediction problem", we use the data to predict a future value of a variate for a unit to be selected from the population/process.

eg given the past performance of a stock/other data, predict the value of the stock at some point in the future.

Chapter 2: Statistical Models and Maximum Likelihood Estimation

STATISTICAL MODELS (C25191)

1. A "statistical model" is a mathematical model that incorporates probability.
2. These are useful since they can describe many different processes.
   eg
   - the daily closing value of CAD
   - when catastrophic events occur (eg pandemics)
   - the effect of drinking alcohol on your health
3. We use random variables to represent a variate / characteristic of a randomly selected unit from the population / process.
   eg let Y = how long I need to wait for the next game on an online video game.
4. Statistical models can also be used to quantify any uncertainties obtained when drawing conclusions from data.
   eg how the observed mean / variance of data differs from the actual mean / variance of data (eg goals scored in hockey)
5. In particular, we can formulate questions of interest as parameters of the statistical model.
   eg In the last example, say
   X = # of hockey goals in a particular game
   and suppose
   $X \sim Po(\theta)$.
   We can then estimate θ (ie the mean # of goals scored).
6. We can then make decisions based on the results of our models, and use computers to simulate the processes.

CHOOSING A PROBABILITY MODEL (C2 S198)

When choosing a probability model, we use some or all of the following:

① Background knowledge / assumptions about the population/process that lead to certain distributions;

② Past experience with data sets from the population/process which show certain distributions are suitable;

③ Mathematical convenience (ie the tradeoff between complexity & accuracy), or

④ A current data set which the model can be assessed.

"ALL MODELS ARE WRONG, BUT SOME ARE USEFUL" (C2 S199)

Note that no statistical model is ever perfect, but that does not mean we cannot learn anything from imperfect ones.

(Quote from John Box)

FAMILIES OF PROBABILITY DISTRIBUTIONS (C2 S200)

Recall the following probability distributions:

① Poisson(θ)

② Exponential(θ)

* "θ" = mean of the distribution (not $\frac{1}{\text{mean}}$).

③ Binomial(n, θ)

④ Gaussian(θ) = Gaussian(μ, σ)

⑤ Multinomial(n, $\theta_1, \ldots, \theta_n$)

⑥ Geometric(θ)

Y IS PARAMETERIZED BY θ · f(y; θ) (C2 S205)

In particular, for each "family" of distributions, we get a different model for each value of θ.

Thus, we say the random variable is "parameterized" by θ.

If the r.v. is Y, we write the pf/pdf of Y as

$f(y; \theta)$ for $y \in A = \text{range}(Y)$

to emphasize the dependence of the model on θ.

ESTIMATION OF UNKNOWN PARAMETERS (C2 S206)

To determine how well the model fits the data, we need a value of θ obtained from the data.

We usually denote this value $\hat{\theta}$.

* don't confuse θ & $\hat{\theta}$!
  - θ = the underlying "true" value
  - $\hat{\theta}$ = our own estimate

This process is refered to as "estimating" the value of θ.

STEPS IN CHOOSING A MODEL (C2 S208)

Suppose we have an experiment which involves collecting data to increase knowledge about a certain phenomena or to answer questions about a phenomena that has been carefully designed.

To choose a model for this experiment, we use the following steps:

① Collect/examine the data;

② Propose a model;
   eg $\mathcal{N}(\mu, \sigma)$

③ Fit the model;
   eg find $\hat{\mu}, \hat{\sigma}$

④ Check the model;

⑤ If required, propose a revised model and return to ③,

⑥ Lastly, draw conclusions using the chosen model & the observed data.

* more about this in Chap 3.

MAXIMUM LIKELIHOOD ESTIMATION (C2S210)

POINT ESTIMATE [OF A PARAMETER]: $\hat{\theta}$ (C1S215)

A "point estimate" of a parameter, say θ, is the value of a function of the observed data y and the other known quantities (eg the sample size n).

We denote this estimate by "$\hat{\theta}$", where $\hat{\theta} = \hat{\theta}(y)$.

* note $\hat{\theta}$ is a function on y, and so $\hat{\theta}$ depends on the value of y (the observed data).

For example:

① $G(\mu, \sigma)$: estimate μ by $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$ (the sample mean)

② $Bin(n, \theta)$: estimate θ by $\hat{\theta} = \frac{y}{n}$ (the sample proportion)

PARAMETER SPACE: Ω

The "parameter space" Ω of a parameter θ is the set of all possible values θ can take.

LIKELIHOOD FUNCTION [FOR DRV] $L(\theta)$ (C2S224)

Let Y be potential data that will be used to estimate θ, and let y be the actual observed data.

Suppose Y is a drv.

Then, the "likelihood function for θ" is defined to be

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

where Ω is the parameter space of θ.

* L is technically a function of θ & y, but for brevity we usually just write $L(\theta)$.

MAXIMUM LIKELIHOOD ESTIMATE / M.L. ESTIMATE: $\hat{\theta}$ (C2S225)

The "maximum likelihood (ie m.l.) estimate" for given data y is the value of θ which maximizes $L(\theta)$, and we denote it by $\hat{\theta}$.

In particular, generally $\hat{\theta}$ satisfies

$$\left.\frac{dL(\theta)}{d\theta}\right|_{\theta=\hat{\theta}} = 0.$$

Why?
- most distributions look like  with a single "max" peak
- so the only place the derivative will be 0 is at the peak, which we want.

RELATIVE LIKELIHOOD FUNCTION: $R(\theta)$ (C2S234)

Let $\hat{\theta}$ be the MLE of $L(\theta)$.

Then, the "relative likelihood function" is

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

Note that

① $0 \le R(\theta) \le 1$;

② $L(\hat{\theta})$ is a constant; and

③ $R(\hat{\theta}) = 1$, and so R is maximized at $\theta = \hat{\theta}$.

RELATIVE LIKELIHOOD FOR BINOMIAL DATA: $R(\theta) = \frac{\theta^y(1-\theta)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}}$, $\hat{\theta} = \frac{y}{n}$ (C2S235)

For binomial data, necessarily

$$R(\theta) = \frac{\theta^y(1-\theta)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}}, \quad \hat{\theta} = \frac{y}{n}.$$

Why? $\rightarrow L(\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$

$= \theta^y(1-\theta)^{n-y}$.

Then $L(\hat{\theta}) = \hat{\theta}^y(1-\hat{\theta})^{n-y}$.

($\hat{\theta} = \frac{y}{n}$ from earlier)

$\Rightarrow R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^y(1-\theta)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}}$.

* when computing relative likelihoods, we can ignore any constants wrt θ as they will cancel out in the computation of $R(\theta)$

LOG LIKELIHOOD FUNCTION: $\ell(\theta)$ (C2S237)

The "log likelihood function" is defined to be

$$\ell(\theta) = \log L(\theta) \quad \forall \theta \in \Omega.$$

* log = ln for this course!

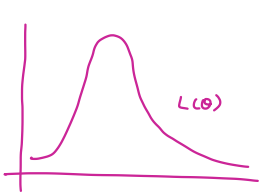
Note that $\ell(\theta)$ is maximized for the same value of θ as the regular likelihood function.

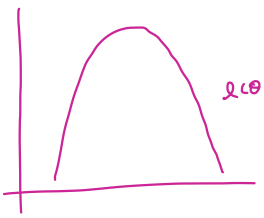
* ie $L'(\hat{\theta}) = 0 \iff \ell'(\hat{\theta}) = 0$.

$\ell(\theta)$ is also preferred over $L(\theta)$ because it is usually easier to take derivatives of ℓ (which typically involves sums) over L (which typically involves products).

However, note $\ell(\theta)$ has a different "shape" than $L(\theta)$ (it looks more "quadratic").

eg $L(\theta) = \theta^{10}(1-\theta)^{15}$





LIKELIHOOD FUNCTION FOR INDEPENDENT EXPERIMENTS (C2S244)

Suppose we observe data $Y = (Y_1, \dots, Y_n)$ that are iid each with p.f. $P(Y_i = y_i ; \theta)$.

Then the (combined) likelihood function for θ based on the data $(y_1, \dots, y_n)$ is

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta) = \prod_{i=1}^{n} P(Y_i = y_i ; \theta) \quad \forall \theta \in \Omega.$$

RELATIVE LIKELIHOOD FOR POISSON DATA:

$$R(\theta) = \frac{\theta^{n\bar{y}} e^{-n\theta}}{\hat{\theta}^{n\bar{y}} e^{-n\hat{\theta}}}, \quad \hat{\theta} = \bar{y} \qquad \text{(C2S254)}$$

For Poisson data, necessarily

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^{n\bar{y}} e^{-n\theta}}{\hat{\theta}^{n\bar{y}} e^{-n\hat{\theta}}}, \quad \hat{\theta} = \bar{y}$$

Proof First, see that

$$P(Y_i = y_i ; \theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!}.$$

Therefore

$$L(\theta) = \prod_{i=1}^{n} P(Y_i = y_i ; \theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!}$$

$$= \prod_{i=1}^{n} \frac{1}{y_i!} \prod_{i=1}^{n} \theta^{y_i} \prod_{i=1}^{n} e^{-\theta}$$

$$\propto \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta} \quad \text{(we ditch the constant)}$$

$$= \theta^{n\bar{y}} e^{-n\theta}, \quad (\because \bar{y} = \tfrac{1}{n}\textstyle\sum y_i)$$

and so

$$\ell(\theta) = \log L(\theta) = n\bar{y} \log(\theta) - n\theta.$$

Thus

$$\ell'(\theta) = \frac{n\bar{y}}{\theta} - n \quad (= 0)$$

and so ℓ (and thus L) is maximized when $\theta = \bar{y}$ $(= \hat{\theta})$.

Therefore

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^{n\bar{y}} e^{-n\theta}}{\hat{\theta}^{n\bar{y}} e^{-n\hat{\theta}}}, \quad \hat{\theta} = \bar{y}. \qquad \blacksquare$$

RANDOM SAMPLE: $Y_1, \dots, Y_n$ (C2S256)

Suppose $Y_1, \dots, Y_n$ are iid with p.f $P(Y = y ; \theta) = f(y ; \theta)$.

We call $Y_1, \dots, Y_n$ a "random sample".

LIKELIHOOD FUNCTION FOR A RANDOM SAMPLE (C2S257)

Let $Y_1, \dots, Y_n$ be a random sample, with p.f. $P(Y = y ; \theta) = f(y ; \theta)$.

Let $y_1, \dots, y_n$ be a realization of (ie the observed data from) the random sample.

Then the likelihood function for θ based on the observed sample is

$$L(\theta) = \prod_{i=1}^{n} P(Y_i = y_i ; \theta) \quad \forall \theta \in \Omega.$$

Proof. $L(\theta) = P(\text{observing the data } y_1, \dots, y_n \text{ given } \theta)$

$$= P(Y_1 = y_1, \dots, Y_n = y_n ; \theta)$$

$$= P(Y_1 = y_1 ; \theta) \cdots P(Y_n = y_n ; \theta) \quad \text{(by independence)}$$

$$= \prod_{i=1}^{n} P(Y_i = y_i ; \theta). \qquad \blacksquare$$

LIKELIHOOD FOR CONTINUOUS RANDOM VARIABLES (C2S258)

LIKELIHOOD FUNCTION FOR CRV (C2S262)

Let $Y=(Y_1,\dots,Y_n)$ be a random sample from a continuous distribution with pdf $f(y;\theta)$ for $\theta\in\Omega$.

Let $y=(y_1,\dots,y_n)$ be a realization of Y.

Then, the "likelihood function for θ" based on the observed data $y=(y_1,\dots,y_n)$ is defined to be

$$L(\theta) = L(\theta;y) = \prod_{i=1}^{n} f(y_i;\theta) \quad \forall\theta\in\Omega.$$

MLE FOR Exp(θ): $\hat{\theta}=\bar{y}$ (C2S266)

Let $Y\sim \text{Exp}(\theta)$, and let $(y_1,\dots,y_n)$ be the observed data from a sample of size n.

Then the maximum likelihood estimate is necessarily $\hat{\theta}=\bar{y}$.

Proof. See that

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{y_i}{\theta}}$$

$$= \theta^{-n} e^{-n\bar{y}/\theta},$$

and so

$$\ell(\theta) = \log L(\theta) = -n\log\theta - \frac{n\bar{y}}{\theta} \quad (=0).$$

Hence

$$\ell'(\theta) = -\frac{n}{\theta} + \frac{n\bar{y}}{\theta^2} \quad (=0)$$

and it follows ℓ (and so L) is maximized when $\theta=\bar{y}$ $(=\hat{\theta})$. ∎

LIKELIHOOD FUNCTION FOR $\mathcal{N}(\mu,\sigma)$:

$$L(\theta) = (2\pi)^{-n/2}\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \quad \text{(C2S267)}$$

Let $y_1,\dots,y_n$ be observations from $Y\sim\mathcal{N}(\mu,\sigma)$.

Then necessarily

$$L(\theta) = (2\pi)^{-n/2}\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right].$$

Proof. $L(\theta) = \prod_{i=1}^{n} f(y_i;\mu,\sigma)$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{1}{}\right.$$