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!

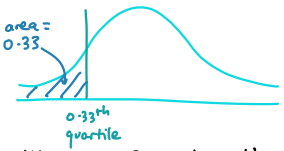
QUANTILES & PERCENTILES (CIS74)

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)

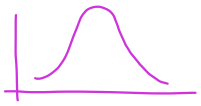
SAMPLE SKEWNESS (CIS88)

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

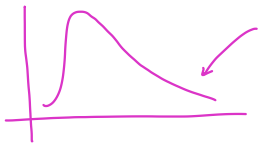
$$\text{sample skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{3/2}}$$

Interpretation of sample skewness's value:

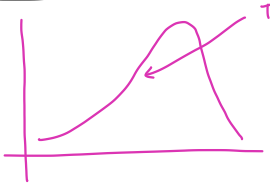
① If ss = 0 ⇒ distribution is symmetric; eg Gaussian, uniform



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)

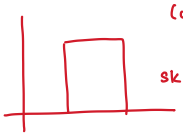
"Sample kurtosis" measures whether data is concentrated in the central "peak" or in the tails, and is calculated by

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

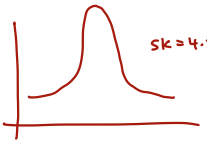
Interpretation of sample kurtosis' value:

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)

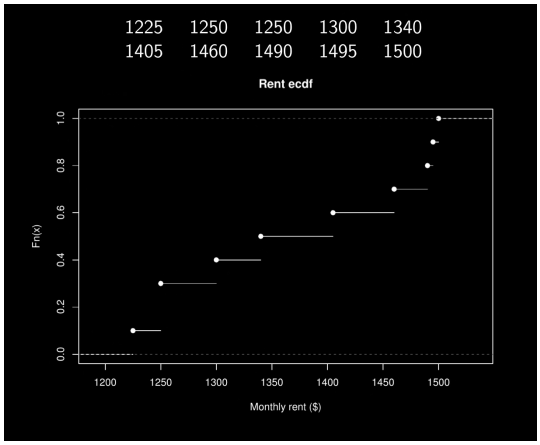
$$\text{height} = \frac{f_j/n}{a_j - a_{j-1}}$$

EMPIRICAL CDF (CIS124)

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

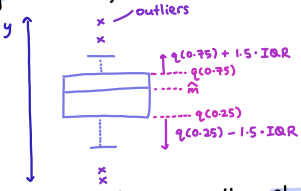
Mathematically, the empirical cdf is defined by

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

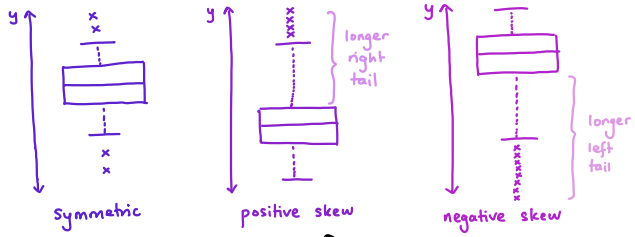


BOX-PLOT (CIS139)

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

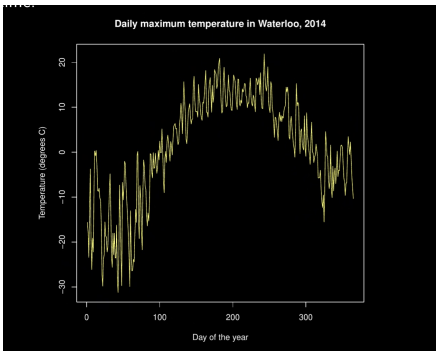


Box-plots can also show the skewness of a distribution:



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)

$_1$ "Inductive reasoning" occurs when we reason from the "specific" (observed data about a sample) to the "general" (the target population/process).

$_2$ In contrast, "deductive reasoning" occurs when we use general results to prove theorems.

* proof by induction = deductive reasoning!