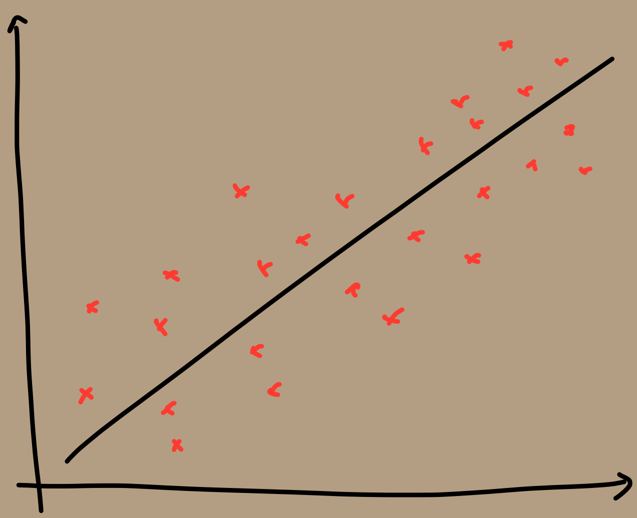
STAT 241
Personal Notes

☆ 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)

1. An "empirical study" is one where we learn by observation and/or experimentation.
2. Note these involve uncertainty – repeated experiments generate different results.
3. 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)

1. 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

2. 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)

1. "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.

2. 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!

3. 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)

1. "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

2. 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)

1 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

2 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 (C1S56)

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

MEASURES OF CENTRAL TENDENCY / LOCATION (C1S58)

1. 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".
2. We also use lower-case for constants, and upper-case for random variables.

ORDERED SAMPLE / ORDER STATISTICS (C1S59)

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

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

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

SAMPLE MEAN/AVERAGE: $\bar{y}$ (C1S58)

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

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

* the keyword "sample" is important!

SAMPLE MEDIAN: $\hat{m}$ (C1S59)

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

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

2. 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). (C1S66)

SAMPLE MODE (C1S61)

1. The "sample mode" is just the most common value(s) in a set of data.
2. In this case, the "sample modal class" is the group/class with the highest frequency.

MEASURES OF VARIABILITY / DISPERSION (C1S67)

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

ROBUST [MEASURE] (C1S80)

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 (C1S69)

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

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

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

"68-95" RULE FOR GAUSSIAN ESTIMATION (C1S70)

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 (C1S73)

The "range" is defined as

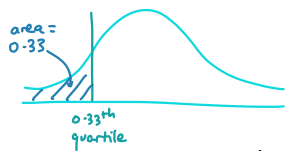
$$\text{range} = y_{(n)} - y_{(1)}.$$

* the range is very susceptible to outliers!

QUANTILES & PERCENTILES (C1S74)

1. The "p^{th} quantile", 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} quantile / 50^{th} percentile.
2. In other words, the p^{th} quantile of a distribution is the value q such that

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



3. We can calculate quantiles in R using the code

```
> quantile(c(y1,..., yn), p)
```

QUARTILES: $q(0.25)$, $\hat{m}$, $q(0.75)$ (C1S79)

1. The "lower quartile", or "first quartile", denoted by $q(0.25)$, is the 25^{th} percentile.
2. The "upper quartile", or "third quartile", denoted by $q(0.75)$, is the 75^{th} percentile.
3. The "second quartile" is just the median $\hat{m}$.

INTERQUARTILE RANGE / IQR (C1S80)

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)

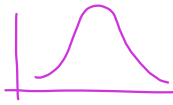
1. "Sample skewness" measures the asymmetry of the data, and is equal to

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

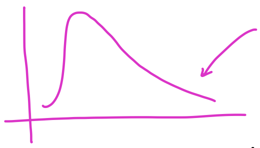
2. Interpretation of sample skewness's value:

① If $ss = 0$ ⇒ distribution is symmetric;

eg Gaussian, uniform



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



The data has a long right tail.

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

The data has a long left tail.

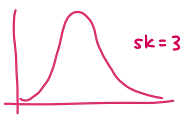
SAMPLE KURTOSIS (CIS96)

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

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

2. Interpretation of sample kurtosis' value:

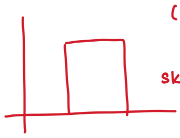
① $sk = 3$ ⇒ distribution looks "Gaussian" (bell-shaped);



sk = 3

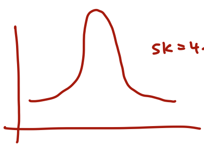
② $sk < 3$ ⇒ distribution has shorter tails (more concentrated in the peak)

(uniform dist^n)



sk = 1.8

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



sk = 4.2

ASSUMING A MODEL IS GAUSSIAN (CIS102)

1. 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)

1. 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)

1. 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 $y_{(n)}$.

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

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

GRAPHICAL SUMMARIES (C1S112)

💡 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 (C1S116)

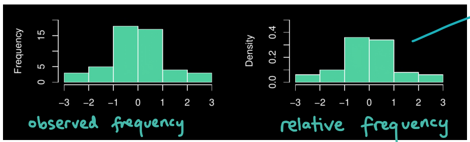
💡₁ Essentially, histograms create a graphical summary of our data that we can use to compare with a pdf for crvs, or a pmf for a drv.

💡₂ Let our data be $y_1, \dots, y_n$.
Partition the range of the y's into k non-overlapping intervals

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

Let f_j = # of values from $\{y_1, \dots, y_n\}$ in I_j.
The f_j's are called the "observed frequencies".

💡₃ Then, draw a rectangle above each of the intervals with height proportional to the observed / relative frequency.



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

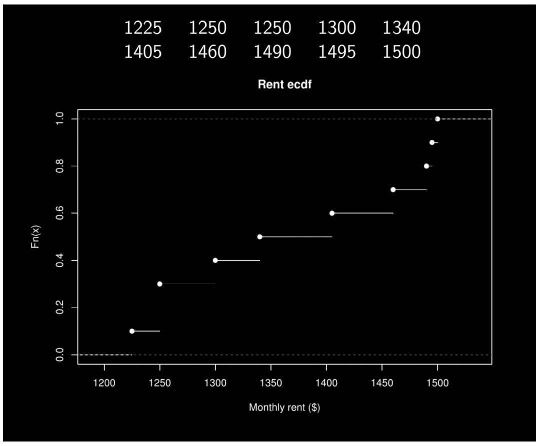
(standard histogram) (rel. freq. histogram)

EMPIRICAL CDF (C1S124)

💡₁ An "empirical cdf" lets us compare the distribution of a dataset with a cdf of a random variable.

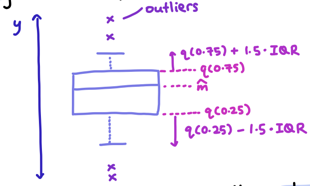
💡₂ Mathematically, the empirical cdf is defined by

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

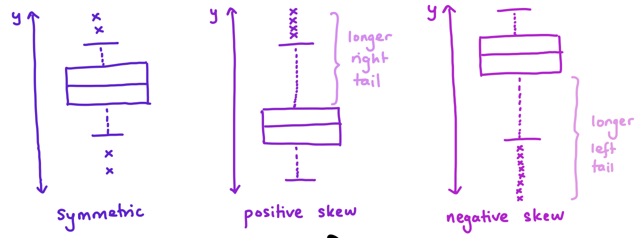


BOX-PLOT (C1S139)

💡₁ "Box-plots" give a graphical summary of the shape of a dataset's distribution in a similar way to the five number summary.

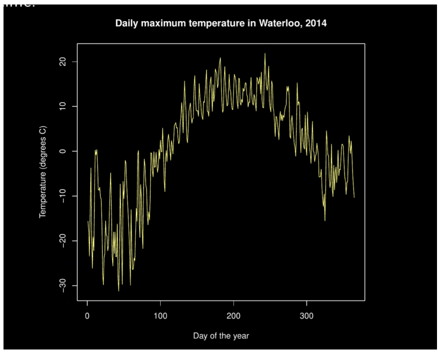


💡₂ Box-plots can also show the skewness of a distribution:



RUN CHART (C1S154)

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

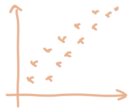
SAMPLE CORRELATION: r (CIS162)

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

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

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

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

eg in the survey above,

relative risk of liking hockey among those w/ a Canadian hometown $= \dfrac{\text{prop. of Canada hometown who like hockey}}{\text{prop. of non-Canada hometown who dislike hockey}}$

$$= \frac{(33/55)}{(9/52)}$$

$$= 3.467$$

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!

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 (C2S191)

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)

1. 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" (C2S199)

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 (C2S200)

1. 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. $f(y;\theta)$ (C2S205)

1. In particular, for each "family" of distributions, we get a different model for each value of θ.
2. Thus, we say the random variable is "parameterized" by θ.
   If the r.v. is Y, we write the pf/pdf of Y as
   $$f(y;\theta) \quad \text{for} \quad y \in A = \text{range}(Y)$$
   to emphasize the dependence of the model on θ.

ESTIMATION OF UNKNOWN PARAMETERS (C2S206)

1. To determine how well the model fits the data, we need a value of θ obtained from the data.
2. We usually denote this value $\hat{\theta}$.
   * don't confuse θ & $\hat{\theta}$!
     - θ = the underlying "true" value
     - $\hat{\theta}$ = our own estimate
3. This process is referred to as "estimating" the value of θ.

STEPS IN CHOOSING A MODEL (C2S208)

1. 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.
2. To choose a model for this experiment, we use the following steps:
   ① Collect / examine the data;
   ② Propose a model;
      eg $G(\mu, \sigma)$
   ③ Fit the model;
      eg find $\hat{\mu}$, $\hat{\sigma}$
   ④ Check the model;
   ⑤ If required, propose a revised model and return to ②,
   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)

1. 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).
2. 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).
3. 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

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

where Ω is the parameter space of θ.

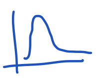
* L is technically a function of θ & y, but for brevity we usually just write $L(\theta)$.

MAXIMUM LIKELIHOOD ESTIMATE / m.l. ESTIMATE: $\hat{\theta}$ (C2S225)

1. 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 $\hat{\theta}$.
2. In particular, generally $\hat{\theta}$ satisfies

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

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)

1. Let $\hat{\theta}$ be the MLE of $L(\theta)$. Then, the "relative likelihood function" is

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

2. 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:

$$R(\theta) = \frac{\theta^y(1-\theta)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}}, \quad \hat{\theta} = \frac{y}{n} \quad \text{(C2S235)}$$

For binomial data, necessarily

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

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

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

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

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

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

* 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)

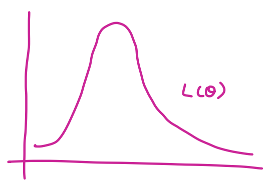
1. The "log likelihood function" is defined to be

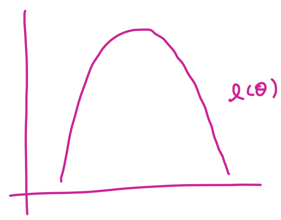
$$\boxed{\ell(\theta) = \log L(\theta) \quad \forall \theta \in \Omega.}$$

* $\log = \ln$ for this course!

2. Note that $\ell(\theta)$ is maximized for the same value of θ as the regular likelihood function.
   * ie $L'(\hat{\theta}) = 0 \iff \ell'(\hat{\theta}) = 0$.
3. $\ell(\theta)$ 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).
4. 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

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

RELATIVE LIKELIHOOD FOR POISSON DATA:

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

For Poisson data, necessarily

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

Proof First, see that

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

Therefore

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Let $y_1,\dots,y_n$ be observations from $Y\sim G(\mu,\sigma)$.

Then necessarily

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

Proof.

$$\begin{aligned}L(\theta) &= \prod_{i=1}^{n} f(y_i;\mu,\sigma)\\ &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right]\\ &= (2\pi)^{-\frac{n}{2}}\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right].\end{aligned}$$ ☒

In particular, the MLE is

$$\boxed{\hat\mu=\bar y, \qquad \hat\sigma = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar y)^2\right]^{\frac{1}{2}}.}$$

Proof. First, see that

$$\ell(\theta) = \log(L(\theta)) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Then

$$\frac{\partial\ell}{\partial\mu} = \frac{n}{\sigma^2}(\bar y-\mu) \quad \& \quad \frac{\partial\ell}{\partial\sigma} = -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Thus

$$\frac{\partial\ell}{\partial\mu}=0 \Rightarrow \hat\mu=y \quad \&$$

$$\frac{\partial\ell}{\partial\sigma}=0,\ \hat\mu=y \Rightarrow \hat\sigma = \frac{1}{n}\left[\sum_{i=1}^{n}(y_i-\bar y)^2\right]^{\frac{1}{2}}.$$ ☒

INVARIANCE PROPERTY OF MLEs (C2S272)

Let $\hat\theta$ be the MLE of a parameter θ.

Then $g(\hat\theta)$ is necessarily the MLE of $g(\theta)$.

eg Suppose $Y\sim \text{Poi}(\theta)$, $\hat\theta=3$.

Then $P(Y\geq 3) = 1-P(Y\leq 2) = 1-\sum_{y=0}^{2}\frac{\theta^y e^{-\theta}}{y!}$.

*in R, we calculate this via "1 - ppois(2,3)"

But this is a function of θ, so the MLE of $P(Y\geq 3)$ is

$$1-\sum_{y=0}^{n}\frac{\hat\theta^y e^{-\hat\theta}}{y!}.$$

We should always clarify when/where we use the invariance property.

CHECKING MODEL FIT (C2S276)

COMPARING OBSERVED VS EXPECTED FREQUENCIES [FOR DRV] (C2S277)

To check whether a model fits a given set of data, we can compare the observed frequencies & the expected frequencies using a table.

eg Suppose a hockey team scored the following # of goals in these # of games:

Goals	0	1	2	3	4	5	6	7
Games	2	17	21	18	15	7	1	1

Let's say we assume the data can be modelled by a Poisson dist^n, say $Y \sim Poi(\theta)$.

Then, we estimate θ using the MLE of θ, aka $\hat{\theta}$;

$$\hat{\theta} = \bar{y} = \frac{1}{82}(2(0) + 17(1) + \dots + 1(7)) = 2.695.$$

Next, we calculate the expected frequencies.

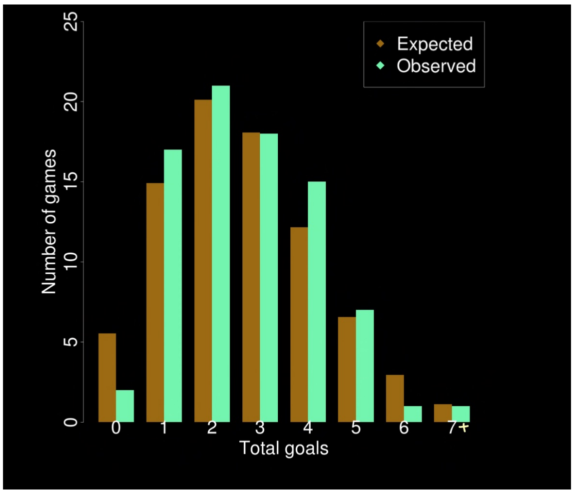
Since the range of Poi is technically $0, 1, 2, \dots$, we need to account for the "right tail" by grouping all the values ≥ 7 into "one" value:

Goals	0	1	2	3	4	5	6	≥ 7
Obs.	2	17	21	18	15	7	1	1
Exp.	5.54	14.93	20.11	18.07	12.17	6.56	2.95	1.67

where

$$\text{exp value for } i = nP(Y=i) = 82 \, \frac{e^{-2.695}(2.695)^i}{i!}.$$

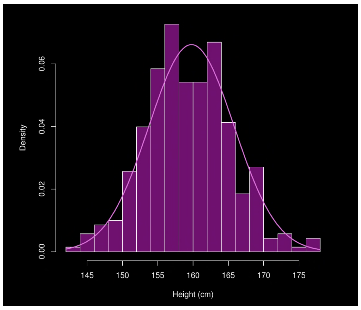
We may also plot the expected/observed frequencies via a bar plot.



COMPARING OBSERVED VS EXPECTED FREQUENCIES [FOR CRV] (C2S300)

We can do something similar for continuous random variables.

eg Consider the dataset



We have
$\bar{y} = 159.77$,
$s^2 = 36.36$,
$s = 6.03$.
Let Y be the data.

How reasonable is it to model the data via a Gaussian distribution?

Suppose it is; ie $Y \sim G(\mu, \sigma)$.

We estimate

$\mu \approx$ the sample mean (MLE); &

$\sigma \approx$ the sample sd (not the MLE),

so $Y \sim G(159.77, 6.03)$.

We then can estimate the exp. probabilities Y falls into one of the intervals of the histogram outlined above; eg

$$P(160 \leq Y \leq 162) = P\left(\frac{160 - 159.77}{6.03} < Z < \frac{162 - 159.77}{6.03}\right)$$

$$= 0.129$$

* in R, we can calculate this via

```
> pnorm(0.370) - pnorm(0.038)
```

or

```
> pnorm(162, 159.77, 6.03) - pnorm(160, 159.77, 6.03).
```

and thus calculate the exp. # of values to fall within the given interval; eg

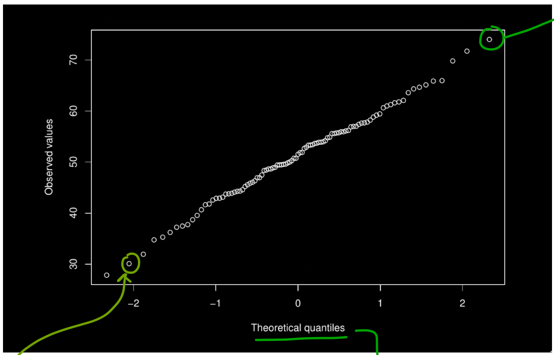
$$e_j = 351 p_j, \quad \text{where } I_j \text{ is the } j^{th} \text{ interval,}$$

and compare this with the observed values.

QQ / QUANTITY - QUANTITY PLOTS (C2S311)

A "QQ plot" plots the observed values / quantiles from the sample data on the y-axis over the theoretical values obtained by fitting a model to said data on the x-axis.

In particular, we may standardize the theoretical values and plot that on the x-axis instead.



after standardization.

each dot corresponds to a quantile; ie the q^{th} quartile.

① The y-value corresponds to the value y such that "q" of the data is $\leq y$.

② The x-value corresponds to the value x such that if we fit the rv Y to a model, and standardize said model to be Z, then $P(Z \leq x) = q$.

If we model $Y \sim G(\mu, \sigma)$, then the QQ-plot of the points

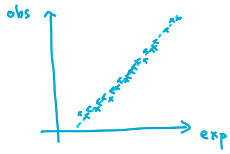
$$\left(\Phi^{-1}\left(\frac{i}{n+1}\right),\ y_{(i)}\right) \quad \text{for } 1 \leq i \leq n,$$

where $y_{(1)}, \dots, y_{(n)}$ is the observed data, should be approximately a straight line if the normal distribution is a good fit for said data.

USING QQ-PLOTS TO INFER SHAPE OF DISTRIBUTION (C2S341)

We can use QQ-plots to infer the underlying shape of a distribution:

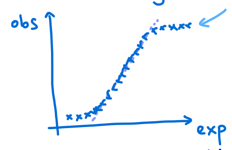
① If the points are along a straight line, then this indicates the data is normal.



② If the data is S-shaped, this indicates symmetry (ie low skewness).

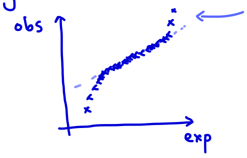
→ then, the relative abundance of points in the "center" vs. tails implies the magnitude of the kurtosis.

low kurtosis, symmetric



"leveling out"
⇒ very little data in the tails
⇒ data concentrated in the peak
⇒ ie low kurtosis

high kurtosis, symmetric

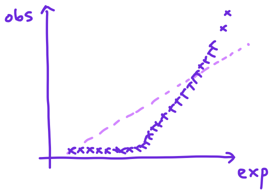


"increasing near the tails"
⇒ data concentrated more in the tails
⇒ ie high kurtosis

③ If the data is U-shaped, this indicates asymmetry.

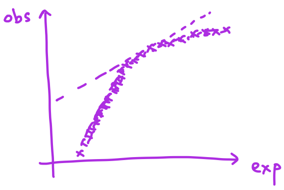
→ then, the relative abundance of points in the left vs right tails implies the magnitude and sign of the skewness.

positive skew, asymmetric



⇒ more data towards the right (ie long right tail)
⇒ data is skewed to the right
⇒ ie positive skew

negative skew, asymmetric



⇒ more data towards the left (ie long left tail)
⇒ data is skewed to the left
⇒ ie negative skew

NORMALITY CHECKING SUMMARY (C2S344)

To assume data is a good fit for a Gaussian model, we need to check:

① The sample mean & median are approximately equal;
② The sample skewness is close to 0;
③ The sample kurtosis is close to 3;
④ Approximately 95% of the observations lie in $[\bar{y} - 2s,\ \bar{y} + 2s]$;
⑤ Histograms & ecdfs should show agreement between the data & theoretical distribution;
⑥ The QQ-plot should roughly be a straight line.

UNBIASED ESTIMATOR: S^2 (C2S352)

The "unbiased estimator" for data is defined to be

$$\boxed{S^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.}$$

k^{th} POPULATION MOMENT: μ_k (C2S353)

1. The "k^{th} population moment" of Y is defined to be

$$\boxed{\mu_k = E[Y^k].}$$

2. In particular,
   1. $\mu_1 = E[Y]$;
   2. $\mu_2 = Var(Y) + E[Y]^2$.

k^{th} SAMPLE MOMENT: m_k (C2S355)

Let $y_1, \dots, y_n$ be a sample. Then, the "k^{th} sample moment" is defined to be

$$\boxed{m_k = \frac{1}{n}\sum_{i=1}^{n} y_i^k.}$$

METHOD OF MOMENTS FOR ESTIMATION (C2S358)

1. The "method of moments" allows us to estimate parameters for a model, based off the data we are using the model for.

2. Steps:
   1. Compute the first p sample moments, where $p = \#$ of parameters.
   2. Relate the population moments to the true parameter values.
   3. Use the sample moments to solve the resulting system of equations to estimate the parameters.

EXAMPLE 1: $G(\mu, \sigma)$ (C2S356)

1. Problem:

"Suppose $Y \sim G(\mu, \sigma)$. Use the sample $y_{(1)}, \dots, y_{(n)}$ to estimate μ and σ".

Sol^{n}. Since $Y \sim G(\mu, \sigma)$, and

$$\mu = E(Y) = \mu_1.$$

$$\sigma^2 = E(Y^2) - E(Y)^2 = \mu_2 - \mu_1^2.$$

we can estimate the values of μ & σ by

$$\hat{\mu} = m_1, \quad \hat{\sigma}^2 = m_2 - m_1^2.$$

Hence

$$\hat{\mu} = m_1 = \frac{1}{n}\sum_{i=1}^{n} y_i$$

&

$$\hat{\sigma}^2 = m_2 - m_1^2 = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \bar{y} = \frac{1}{n}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right).$$

2. Note: we use the "^" notation for both MLE and method of moments!

EXAMPLE 2: Unif(a, b) (C2S361)

Problem:

"Suppose $y_1, \dots, y_n$ are independently sampled from a continuous uniform distribution on (a, b). What are the method of moments estimates on (a, b)?"

Sol^{n}. We need to estimate $\underline{2}$ parameters, so we require

$$\mu_1 = E(Y), \quad \mu_2 = E(Y^2)$$

and hence we need to use

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad m_2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2.$$

*remember m_1 & m_2 are both numbers (since they are based off the sample!)

Then, using LOTUS,

$$\mu_1 = \int_a^b \frac{y}{b-a}\,dy = \frac{1}{b-a}\left[\frac{y^2}{2}\right]_a^b = \frac{1}{2}\left(\frac{1}{b-a}\right)(b^2 - a^2) = \frac{1}{2}(b+a),$$

&

$$\mu_2 = \int_a^b \frac{y^2}{b-a}\,dy = \dots = \frac{1}{3}(a^2 + b^2 + ab).$$

We then estimate $\mu_1 \approx m_1$ & $\mu_2 \approx m_2$, so that

$$m_1 = \frac{1}{2}(\hat{a} + \hat{b}) \Rightarrow \hat{a} = 2m_1 - \hat{b},$$

&

$$m_2 = \frac{1}{3}(\hat{a}^2 + \hat{b}^2 + \hat{a}\hat{b}) \Rightarrow (\hat{b} - m_1)^2 = 3(m_2 - m_1^2)$$

using the appropriate subst^{ns}.

Solving for $\hat{a}$ & $\hat{b}$ yields the estimates

$$\hat{b} = m_1 + \sqrt{3(m_2 - m_1^2)}$$

&

$$\hat{a} = m_1 - \sqrt{3(m_2 - m_1^2)},$$

and by evaluating m_1 & m_2 we could then compute $\hat{a}$ & $\hat{b}$.

Moreover, note that

$$m_2 - m_1^2 = \frac{n-1}{n}s^2,$$

and so we could also write the above as

$$\hat{a} = m_1 - \sqrt{3\left(\frac{n-1}{n}\right)s^2}, \quad \hat{b} = m_1 + \sqrt{3\left(\frac{n-1}{n}\right)s^2}.$$

EXAMPLE 3: CONTEST & PRIZES (C2S372)

Problem:

"A contest awards prizes as follows:

- P(win \$1) = a;
- P(win \$10) = b;
- P(lose) = $1 - a - b$.

You buy five tickets and win three times, including one \$10 win.

Use MM to estimate a & b."

Sol^{n}. Again, we have two parameters, so we need

$$\mu_1 = E(Y), \quad \mu_2 = E(Y^2)$$

and use the sample moments

$$m_1 = \frac{1}{n}\sum y_i, \quad m_2 = \frac{1}{n}\sum y_i^2.$$

Since Y is a drv, we use

$$E(Y^k) = \sum_{y \in A} y^k f(y),$$

and for this example

$$A = \{0, 1, 10\}, \quad f(0) = 1 - a - b, \quad f(1) = a, \quad f(10) = b.$$

Hence

$$\mu_1 = E(Y) = 0(1-a-b) + 1(a) + 10(b) = a + 10b;$$

&

$$\mu_2 = E(Y^2) = 0^2(1-a-b) + 1^2(a) + 10^2(b) = a + 100b.$$

Then, we estimate μ_i with m_i to get

$$m_1 = \hat{a} + 10\hat{b}, \quad m_2 = \hat{a} + 100\hat{b}.$$

Solving for $\hat{a}$ & $\hat{b}$ yields that

$$\hat{b} = \frac{m_2 - m_1}{90}, \quad \hat{a} = m_1 - \frac{m_2 - m_1}{9}.$$

Finally, for our sample we observed $\{0, 0, 1, 1, 10\}$ and so

$$m_1 = 2.4, \quad m_2 = 20.4$$

and so

$$\hat{a} = 0.4, \quad \hat{b} = 0.2. \quad \blacksquare$$

Chapter 3: Planning and Conducting Empirical Studies

1. Recall "empirical studies" are those where data collected can be used to learn about a population / process.

* we use this "Pfizer vs. Moderna" study for examples:

www.nejm.org/doi/full/10.1056/NEJMoa2115463

so it might be helpful to have the study open whilst reading this chapter.

PPDAC (C3S384)

1. We can design an empirical study using "PPDAC".
2. In particular, this stands for
   ① Problem — a clear statement of the study's objectives;
   ② Plan — the procedures in the study, how the data is collected
   ③ Data — the physical collection of the data
   ④ Analysis — analysis of said data
   ⑤ Conclusion — conclusions drawn from said analysis, and their limitations

PROBLEM (C3S393)

The "problem" addresses
1. what group of things/people do we want the conclusions to apply?
2. what variates can we define?
3. what are the questions we are trying to answer?
4. what conclusions are we trying to draw?

TARGET POPULATION/PROCESS (C3S394)

The "target population/process" is the collection of units to which the experimenters conducting the empirical study wish the conclusions to apply.

In the problem, the units & target population/process must be defined.

eg in the vaccine study, possible target popns/processes:
1. people in VA health-care system now and in future
2. unvaccinated people in the VA health care system now in the future
3. ① & ② but limiting the time period to the duration of the COVID-19 pandemic.

VARIATES [IN EMPIRICAL STUDIES] (C3S398)

A "variate" is a characteristic of a unit.

To determine the variates, look at what is measured or recorded on each unit.

eg for the vaccine study, the variates include
- which vaccine each participant took; (ie Pfizer / Moderna)
- outcome indicators such as COVID-19 infection, symptoms, hospitalization, and death;
- age, sex, race, residence, geographic location;
- etc

ATTRIBUTES [IN EMPIRICAL STUDIES] (C3S402)

"Attributes" are functions of variates over a population.

In the problem step, the questions of interest are specified in terms of the attributes of the target population.

eg in the vaccine study, possible attributes include
1. the proportion of people in the target popn who would contract COVID-19 after receiving the Pfizer vaccine within 24 weeks;
2. the proportion of people in the target popn who would contract COVID-19 after receiving the Moderna vaccine within 24 weeks;
3. the difference in the preceding two numbers.

TYPES OF PROBLEMS (C3S405)

Types of problems an empirical study can solve:

1. "Descriptive" — determine a particular attribute of the population.

   eg
   - the national unemployment rate
   - estimating the relative efficacy of the two vaccines among all those who received it at the time of the study

2. "Causative" — determine the existence (or lack of) of a causal relationship between two variates.

   eg
   - does a new hockey helmet reduce the risk of concussion
   - whether giving someone the Moderna vaccine instead of the Pfizer vaccine reduces their risk of COVID-19

3. "Predictive" — predict the response for a given unit.

   eg
   - predict e-cig weekly sales if sales tax on them is doubled
   - estimating relative efficacy of Pfizer & Moderna

Note that we usually cannot answer causative problems from observational studies.

Causative & descriptive problems are also hard to distinguish.

PLAN (C3S411)

In the plan, we available for
① decide what units are available for study;
② what units will be examined; &
③ what variates will be collected and how.

STUDY POPULATION / PROCESS (C3S411)

1. The "study population / process" is the collection of units available to be included in the study.
2. Note the study population is a strict subset of the target population.
   eg - veterans of age ≥18 years, no previous COVID infection, etc (in vaccine study)

STUDY ERROR (C3S423)

1. "Study error" occurs when the attributes in the study population differ from said attributes in the target population.
   eg Say $Y_T \sim Bin(n, \theta_T)$ represents the # of people in a random sample of size n from the target pop^n who support Brexit.
   Say $Y_S \sim Bin(n, \theta_S)$ represents the # of people in a random sample of size n from the study pop^n who support Brexit.
   We might be concerned $\theta_T \neq \theta_S$. (C3S422)
2. Note that just the study / target populations being different is not study error — the difference must be in their attributes.
3. Moreover, note study error concerns populations; we do not care about the study / target samples.
4. Hence, we must be careful when thinking about the attributes of interest in a study.
5. In particular, as the values of the target or study populations' attributes are unknown, the study error cannot be quantified.
6. Instead, we generally rely on expertise from other sources to determine whether conclusions derived from the study population may apply to the target population.
   eg whether studies on mice (study) apply to humans (target).

SAMPLING PROTOCOL (C3S430)

1. The "sampling protocol" is the procedure used to select a sample of units from the study population.
2. In practice, obtaining a (truly) random sample is difficult / impossible / expensive, so less rigorous sampling methods are usually used.
   eg "matching" in the vaccine study

SAMPLE SIZE (C3S430)

1. The "sample size" is the number of units sampled from the sampling protocol.

SAMPLE ERROR (C3S435)

1. "Sample error" occurs when the attributes in the sample differ from the attributes in the study population.
   * again, it must be a difference in the attributes, not just because the two groups differ!
2. Note sample error does not care about the target population & sample!

MEASUREMENT ERROR (C3S442)

1. "Measurement error" occurs if the measured and true values of a variate are not identical.
   eg - measuring blood pressure
   - patients more stressed in doctor's office
   - so reading is higher
   - "white coat hypertension"

STEPS IN THE PLAN (C3S445)

1. Flowchart:

eg STAT 231 example (C3S447)

Target population → Current & future UW math faculty undergraduate students
↓ study error
Study population → Students enrolled in STAT 231 in the F2018 semester
↓ sample error
Sample → Students who gave their permission for responses to be used
↓ measurement error
Measured variate values → Responses to survey of whether hybrid learning was effective

DATA (C3S454)

1. The "data" step concerns collecting data according to the plan.
2. To do this, the
   ① variates must be clearly defined; &
   ② satisfactory methods of measuring them must be used.

RECORDING DATA (C3S455)

1. Note mistakes can occur in recording data into a PB, and so for more complex investigations it is useful to put checks in place to avoid these mistakes & detect those that are made.
2. Moreover, when lots of data is used, database design and management is important.
3. Also, if data is recorded longitudinally (ie over a period of time), departures from the plan might occur; these must be recorded.
   eg persons might drop out of a long-term medical study because of adverse reactions to a treatment.
4. Such departures will affect the Analysis & Conclusion steps.

ANALYSIS (C3S456)

1. In the "analysis" step, we analyze the data collected.
2. This includes
   ① numerical & graphical summaries of the data;
   ② selecting an appropriate model; &
   ③ checking if said model is a good fit.
3. We usually formulate these questions in terms of the model parameters.
   eg "if $Y \sim Bin(n, \theta)$ & $\theta = P($new drug cures a disease$)$, what is θ?"
4. Departures from the plan that affect the analysis must also be noted.

CONCLUSION (C3S458)

1. In the "conclusion" step, the questions posed in the problem are answered to the extent permitted by the data.
2. In other words, the conclusion is directed by the problem.
3. The conclusion must also feature
   ① a discussion and/or quantification of potential study, sample & measurement errors;
   ② departures from the plan that affect the analysis;
   ③ the limitations of the study.

Chapter 4: Estimation

STATISTICAL MODELS & ESTIMATION (C4S463)

In choosing a model for the analysis step of PPDAC we need to consider:

① Model A: a model for variation in the population / process being studied which includes the attributes which are to be estimated; and

② Model B: a model which takes into account how the data were collected & which is constructed in conjunction with model A.

eg (See C4S466 for more details)

Y = # of ♀ in a randomly chosen sample from the target pop^n who have had COVID-19.

For model A, we may assume

$Y \sim Bin(n, \theta_T)$,

where θ_T = proportion of target pop^n who have had COVID-19

(not "probability person has COVID"!)

For model B, we take into account target & study populations are not the same.

We assume

$Y \sim Bin(n, \theta)$

where θ = proportion of study pop^n who have had COVID-19.

* If $\theta_T \neq \theta$, this represents study error.

For this course, we assume

① data arises from a random sample from the study population; &

② variates are measured without error.

This only means we are able to estimate attributes of interest about the study population, not the target population.

* if we make any inferences about the target pop^n, we have to state our assumptions.

ESTIMATORS & SAMPLING DISTRIBUTIONS (C4S474)

Note that sampling is an inherently random process.

RANDOM VARIABLE ASSOCIATED WITH $\bar{y}$: $\bar{Y}$ (C4S491)

1. Let $Y_1, \dots, Y_n$ be iid. Then we define

$$\boxed{\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i}.$$

2. In particular, if $Y_i \sim G(\mu, \sigma)$, then

$$\boxed{\bar{Y} \sim G(\mu, \tfrac{\sigma}{\sqrt{n}})}.$$

ESTIMATOR (C4S498)

An "estimator" is a rule that tells us how to process the data to obtain an estimate of an unknown parameter θ.

POINT ESTIMATORS : $\tilde{\theta}$ (C4S496)

Let $Y_1, \dots, Y_n$ be potential observations in a random sample.

Consider the point estimate

$$\boxed{\hat{\theta} = g(y_1, \dots, y_n)}.$$

Then, we can associate $\hat{\theta}$ with a random variable

$$\boxed{\tilde{\theta} = g(Y_1, \dots, Y_n)}.$$

eg the random variable associated w/ $\hat{\theta} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is $\tilde{\theta} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

* $\hat{\theta}$ = estimate (single value) ; & $\tilde{\theta}$ = random variable !

SAMPLING DISTRIBUTION [OF AN ESTIMATOR] (C4S500)

The "sampling distribution" of an estimator $\tilde{\theta}$ is the distribution of $\tilde{\theta}$.

GAUSSIAN SAMPLING DISTRIBUTION (C4S511)

1. Let $Y_1, \dots, Y_n \sim G(\mu, \sigma)$, so $\bar{Y} \sim G(\mu, \frac{\sigma}{\sqrt{n}})$

(this is the sampling distribution of the sample mean.)

2. Vary parameters, and how does it affect the sampling distribution:

	↑ sample size, n	↑ std dev, σ	↑ mean, μ
location	does not change	does not change	moves to the right
spread	decreases	increases	does not change
shape	does not change	does not change	does not change

3. Thus, the probability we draw a sample that yields an estimate $\tilde{\mu}$ close to μ

① increases as n increases;
② decreases as σ increases; &
③ does not change with μ.

* in particular, because

$$P(|\tilde{\mu} - \mu| \le \varepsilon) = P(\mu - \varepsilon \le \tilde{\mu} \le \mu + \varepsilon) = P\left(\frac{-\varepsilon\sqrt{n}}{\sigma} \le Z \le \frac{\varepsilon\sqrt{n}}{\sigma}\right) \quad (\text{as } \tilde{\mu} \sim G(\mu, \tfrac{\sigma}{\sqrt{n}})).$$

4. Moreover, see that $sd(\bar{Y}) \approx \frac{\sigma}{\sqrt{n}}$, and so

① $sd(\bar{Y})$ decreases as n increases, and so more of our sample estimates will be closer to μ;
② $sd(\bar{Y})$ increases as $\tilde{\sigma}$ increases, and so less of our sample estimates will be closer to μ;
③ $sd(\bar{Y})$ does not change with $\tilde{\mu}$. (C4S540)

NORMAL APPROXIMATIONS (C4S525)

1. If $Y_1, \dots, Y_n$ are iid with mean μ & variance σ^2, then by the CLT for large enough samples we have

$$\boxed{\frac{\bar{Y} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = Z_n \to G(0,1)}.$$

2. Particular examples:

① Binomial — If $Y \sim Bin(n, \theta)$, then

$$\boxed{\frac{\frac{Y}{n} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \sim G(0,1)}.$$

② Exponential — If $Y_i \sim Exp(\theta)$, then for large n

$$\boxed{\frac{\bar{Y} - \theta}{\frac{\theta}{\sqrt{n}}} \sim G(0,1)}.$$

③ Poisson — If $\theta \ge 5$ then if $Y \sim Poi(\theta)$ then

$$\boxed{Y \approx G(\theta, \sqrt{\theta})}.$$

If $Y_i \sim Poi(\theta)$, then for large n

$$\boxed{\frac{\bar{Y} - \theta}{\sqrt{\frac{\theta}{n}}} \sim G(0,1)}.$$

COMPARING ESTIMATORS (C4S551)

BIAS [OF AN ESTIMATOR]:

Bias ($\tilde{\theta}$) (C4S553)

1. The "bias" of an estimator $\hat{\theta}$ is given by

$$\boxed{\text{Bias}(\tilde{\theta}) = E(\tilde{\theta}) - \theta.}$$

2. If the bias is zero, we say the estimator is "unbiased".

EXAMPLE: $Y \sim Bin(n, \theta)$, $\tilde{\theta} = \frac{Y}{n}$ (C4S554)

Problem:

" Suppose $Y \sim Bin(n, \theta)$. Show $\tilde{\theta} = \frac{Y}{n}$ is unbiased. "

Solⁿ.
$$\begin{aligned} \text{Bias}(\tilde{\theta}) &= E(\tilde{\theta}) - \theta \\ &= E\left(\frac{Y}{n}\right) - \theta \\ &= \frac{1}{n}E(Y) - \theta \\ &= \frac{1}{n}(n\theta) - \theta = 0. \ \# \end{aligned}$$

EXAMPLE: $\tilde{\sigma}^2$ IN $G(\mu, \sigma)$ (C4S555)

Problem:

" Consider $Y_1, \dots, Y_n$ iid $G(\mu, \sigma)$.
What is the bias of the ML estimator
$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$? "

Solⁿ. Note $\text{Bias}(\tilde{\sigma}^2) = E(\tilde{\sigma}^2) - \sigma^2$.

Then
$$\begin{aligned} E[\tilde{\sigma}^2] &= E\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right) \\ &= \frac{1}{n}E\left(\sum_{i=1}^{n} Y_i^2 - 2Y_i\bar{Y} + \bar{Y}^2\right) \end{aligned}$$

Since $\bar{Y} = \frac{1}{n}\sum Y_i$, thus $\sum Y_i = n\bar{Y}$. So

$$\begin{aligned} E[\tilde{\sigma}^2] &= \frac{1}{n}E\left[\sum_{i=1}^{n}(Y_i^2) - 2n\bar{Y}\bar{Y} + n\bar{Y}^2\right] \\ &= \frac{1}{n}E\left[\sum_{i=1}^{n}(Y_i^2) - n\bar{Y}^2\right] \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}E[Y_i^2] - nE(\bar{Y}^2)\right) \end{aligned}$$

Then, note $Var(Y) = E(Y^2) - E(Y)^2$.
For $Y \sim G(\mu, \sigma)$, thus
$\therefore \sigma^2 = E(Y^2) - \mu^2$, & so $E(Y^2) = \mu^2 + \sigma^2$.
Since $\bar{Y} \sim G(\mu, \frac{\sigma}{\sqrt{n}})$, we can show
$$E(\bar{Y}^2) = \frac{\sigma^2}{n} + \mu^2.$$

Thus
$$\begin{aligned} E[\tilde{\sigma}^2] &= \frac{1}{n}\left(\sum_{i=1}^{n}E[Y_i^2] - nE[\bar{Y}^2]\right) \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] \\ &= \frac{1}{n}(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2) \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

So the bias is
$$\text{Bias}(\tilde{\sigma}^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$
which is not zero!

* the MLE & MoM estimator of the variance slightly underestimates the true variance.

* note this bias decreases as n increases.

MEAN SQUARED ERROR / MSE [OF AN ESTIMATOR] (C4S562)

1. The "mean squared error" of an estimator is

$$\boxed{E[(\hat{\theta} - \theta)^2].}$$

2. We prefer estimators with a smaller MSE.

$$E[(\tilde{\theta} - \theta)^2] = Var(\tilde{\theta}) + \text{Bias}(\tilde{\theta})^2$$

<< BIAS - VARIANCE DECOMPOSITION OF THE MSE >> (C4S570)

Notes:
① Large variance & small bias are both undesirable.
② If the estimator is unbiased, then the MSE is just the variance.

Proof.
$$\begin{aligned} E[(\tilde{\theta} - \theta)^2] &= E[(\tilde{\theta} - E[\tilde{\theta}] + E(\tilde{\theta}) - \theta)^2] \\ &= E((\tilde{\theta} - E(\tilde{\theta}))^2 + 2(\tilde{\theta} - E(\tilde{\theta}))(E(\tilde{\theta}) - \theta)) + (E(\tilde{\theta}) - \theta)^2) \\ &= Var(\tilde{\theta}) + \text{Bias}(\tilde{\theta})^2 + 2E((\tilde{\theta} - E(\tilde{\theta}))(E(\tilde{\theta}) - \theta)) \end{aligned}$$

Then
$$\begin{aligned} E((\tilde{\theta} - E(\tilde{\theta}))(E(\tilde{\theta}) - \theta)) &= \underbrace{(E(\tilde{\theta} - E(\tilde{\theta})))}_{\text{constant}}(E(\tilde{\theta}) - \theta) \\ &= 0 \times (E(\tilde{\theta}) - \theta) \\ &= 0. \end{aligned}$$

Proof follows. ☑

EFFICIENCY (C4S575)

SCORE [OF A PARAMETER]: $U(\theta;Y)$ (C4S578)

1. The "score" of an unknown parameter θ is the gradient of the log-likelihood function; ie

$$\boxed{U(\theta;Y) = \frac{\partial}{\partial\theta}\ell(\theta;\underline{Y}).}$$

2. Notes:

① $U(\theta;Y)$ is a random variable;

② $$\boxed{U(\theta;Y) = \frac{\partial}{\partial\theta}\log L(\theta;Y) = \frac{1}{L(\theta;Y)}\frac{\partial}{\partial\theta}L(\theta;Y);}$$

③ $E[U(\theta;Y);\theta] = 0$.

EXAMPLE: $E(U|\theta)$ OF $Exp(\theta)$ (C4S580)

Problem:

"Suppose $Y_1,\dots,Y_n$ are iid $Exp(\theta)$ r.v. Show expected value of the score is 0."

Soln. First, see that

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

Thus

$$\ell(\theta) = -n\log(\theta) - \frac{1}{\theta}\sum_{i=1}^{n}y_i.$$

Hence

$$U(\theta;Y) = \frac{\partial}{\partial\theta}\left[-n\log(\theta) - \frac{1}{\theta}\sum_{i=1}^{n}Y_i\right] = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n}Y_i.$$

So

$$\begin{aligned}E(U|\theta) &= E\left(-\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n}Y_i\right)\\ &= -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n}E(Y_i)\\ &= -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n}\theta\\ &= -\frac{n}{\theta} + \frac{n\theta}{\theta^2} = 0,\end{aligned}$$

as expected. ■

FISHER INFORMATION [OF A PARAMETER]: $\mathcal{I}(\theta)$ (C4S584)

1. The "Fisher Information" of an parameter θ is the variance of its score; ie

$$\boxed{\mathcal{I}(\theta) = E\left(\left[\frac{\partial}{\partial\theta}\log L(\theta;Y)\right]^2 \,\middle|\, \theta\right).}$$

2. We can also write

$$\boxed{\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta;Y) \,\middle|\, \theta\right].}$$

Proof. Note that

$$\frac{\partial^2}{\partial\theta^2}\log L(\theta;Y) = \frac{\frac{\partial^2}{\partial\theta^2}L(\theta;Y)}{L(\theta;Y)} - \left(\frac{\partial}{\partial\theta}\log L(Y;\theta)\right)^2$$

Then $E\left(\frac{\partial^2}{\partial\theta^2}L(\theta;Y)\big/L(\theta;Y)\right) = 0$, and so taking expectations of both sides, yields that

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta;Y)\,\middle|\,\theta\right]$$

as needed.

3. Hence, the information tells us about the shape of the log-likelihood.

4. In turn, the shape of $\ell(\theta)$ near the maximum likelihood tells us how many values of θ lead to similar values of the log-likelihood itself.

5. If we have iid rv $Y_1,\dots,Y_n$, then if

$$\boxed{\mathcal{I}_1(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta;Y_1)\,\middle|\,\theta\right]}$$

then

$$\boxed{\mathcal{I}(\theta) = n\mathcal{I}_1(\theta).}$$

EXAMPLE: $\mathcal{I}(\theta)$ OF $Exp(\theta)$ (C4S589)

Problem:

"Let $Y_1,\dots,Y_n$ be iid $Exp(\theta)$ rv. Find the Fisher Information."

Soln. Earlier we showed that

$$\frac{\partial}{\partial\theta}\log L(\theta;y) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n}Y_i.$$

Thus

$$\frac{\partial^2}{\partial\theta^2}\log L(\theta;y) = \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n}Y_i.$$

Since $E(Y_i) = \theta$, thus

$$\begin{aligned}\mathcal{I}(\theta) &= -E\left[\frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n}Y_i \,\middle|\, \theta\right]\\ &= -E\left[\frac{n}{\theta^2} - \frac{2n\theta}{\theta^3}\right]\end{aligned}$$

$$\boxed{\therefore\ \mathcal{I}(\theta) = \frac{n}{\theta^2}.}$$

$Var(\tilde{\theta}) \geq \frac{1}{\mathcal{I}(\theta)}$ <<CRAMER-RAO LOWER BOUND>> (C4S594)

Let $\tilde{\theta}$ be an unbiased estimator. Then necessarily

$$\boxed{Var(\tilde{\theta}) \geq \frac{1}{\mathcal{I}(\theta)}.}$$

MINIMUM-VARIANCE UNBIASED ESTIMATOR / MVUE (C4S594)

A "minimum-variance unbiased estimator" is $\tilde{\theta}$ such that

$$\boxed{Var(\tilde{\theta}) = \frac{1}{\mathcal{I}(\theta)}.}$$

EFFICIENCY [OF AN UNBIASED ESTIMATOR]: $e(\theta)$ (C4S600)

1. The "efficiency" of an unbiased estimator $\tilde{\theta}$ of a parameter θ is

$$\boxed{e(\tilde{\theta}) = \frac{1}{\mathcal{I}(\theta)Var(\theta)}.}$$

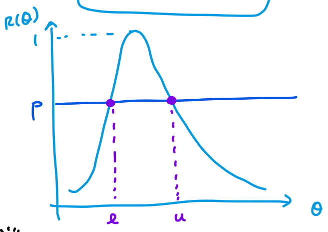
2. If $Var(\tilde{\theta}) = \frac{1}{\mathcal{I}(\theta)}$, then $e(\tilde{\theta}) = 1$, and in this case we say the estimator is "efficient".

INTERVAL ESTIMATION (C4S602)

LIKELIHOOD INTERVAL (C4S614)

1. A "$100p\%$ likelihood interval" for the parameter θ is the set

$$\boxed{\{\theta \mid R(\theta) \geq p\}.}$$



⇒ $100p\%$ confidence interval is $[l, u]$ (ie $[L(y), U(y)]$).

2. Interpretations: if the data is in a
   1. 50% likelihood interval → very plausible
   2. 10% likelihood interval → plausible
   3. 5% likelihood interval → implausible
   4. 1% likelihood interval → very implausible

3. In particular, increasing the sample size n decreases the width of the likelihood intervals.
   → distⁿ becomes narrower as $n \uparrow$

LOG RELATIVE LIKELIHOOD FUNCTION: $r(\theta)$ (C4S627)

1. The "log relative likelihood function" of θ is

$$\boxed{r(\theta) = \log R(\theta) = \ell(\theta) - \ell(\hat{\theta}).}$$

2. To obtain a $100p\%$ likelihood interval, we plot $r(\theta)$ and draw a line at $r(\theta) = \log(p)$.

CONFIDENCE INTERVALS & PIVOTAL QUANTITIES (C4S632)

COVERAGE PROBABILITIES [OF INTERVAL ESTIMATORS] (C4S633)

1. Let $Y = (Y_1, \dots, Y_n)$ be the potential data to be collected.
2. Let $[L(Y), U(Y)]$ be an "interval estimator" which can be used to construct the possible values θ can take.
3. Then, the "coverage probability" for the interval estimator $[L(Y), U(Y)]$ is equal to

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

4. We choose $L(Y)$, $U(Y)$ such that
   ① The coverage probability is large; &
   - 0.90, 0.95, 0.99

   ② The interval is as narrow as possible.

   but these conflict!
5. Usually, we fix the coverage probability and try to find the narrowest interval.

CONFIDENCE INTERVAL & COEFFICIENT (C4S640)

1. A "$100p\%$ confidence interval" for a parameter θ is an interval estimate $[L(y), U(y)]$ such that

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

2. p is called the "confidence coefficient" of the interval.
3. Note that
   ① θ is an unknown constant of the population, not a random variable.
   ② So, we cannot say "the probability θ lies in between $L(\theta)$ & $U(\theta)$ is p".
4. But, we can say we are "$100p\%$ confident" that the interval contains the true (and unknown) value of θ.
5. Note greater confidence corresponds to a wider confidence interval!

PIVOTAL QUANTITY (C4S652)

1. A "pivotal quantity" $Q = Q(Y;\theta)$ is a function of the data Y & the unknown parameter θ such that the distribution of the random variable Q is completely known.

   eg $\underbrace{\dfrac{\bar{Y} - \mu}{\sigma}}_{\text{the PQ}} \sim G(0,1)$
2. To use a pivotal quantity to construct a confidence interval:
   ① Determine numbers a, b such that $P(a \leq Q(Y, \theta) \leq b) = p$;

   * not necessarily symmetric. But try to find narrowest interval.

   ② Re-express $a \leq Q(Y;\theta) \leq b$ in the form $L(Y) \leq \theta \leq U(Y)$;
   ③ Then

$$p = P(L(Y) \leq \theta \leq U(Y)) = P(a \leq Q(Y;\theta) \leq b).$$

   ④ For observed data y, the interval $[L(y), U(y)]$ is a $100p\%$ confidence interval for θ.

GAUSSIAN DATA (C4S673)

Let $Z \sim G(0,1)$. Then, a $100p\%$ confidence interval for a sample size of n is

$$\boxed{\left(\bar{y} - a\frac{\sigma}{\sqrt{n}},\ \ \bar{y} + a\frac{\sigma}{\sqrt{n}}\right),}$$

where

$$\boxed{P(Z \leq a) = \frac{1+p}{2}.}$$

in R: qnorm((1+p)/2)

TWO-SIDED, EQUAL-TAILED CIS FOR μ (C4S674)

A $100p\%$ confidence interval for μ is of the form

$$\boxed{\text{point estimate} \pm \text{distribution quantile} \times \text{sd(estimator)}.}$$

* note not all CIs are symmetric in general!

ASYMPTOTIC / APPROXIMATE PIVOTAL QUANTITY (C4S682)

1. An "asymptotic/approximate pivotal quantity" is a set of random variables $Q_n = Q_n(Y_1, \dots, Y_n; \theta)$ such that as $n \to \infty$, the distribution of Q_n ceases to depend on θ or other unknown information.
2. These can be used to construct approximate CIs for θ.

EXAMPLE: $\text{Bin}(n, \theta)$, $\tilde{\theta} = \frac{Y}{n}$ (C4S686)

Problem:

> "Let $Y \sim \text{Bin}(n, \theta)$, $\tilde{\theta} = \frac{Y}{n}$. Find an approximate 95% CI for θ."

Sol^{n}. By CLT,

$$\frac{\tilde{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \sim G(0,1) \text{ approximately.}$$

Moreover,

$$Q_n = \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \sim G(0,1) \text{ approximately.}$$

Then since

$$0.95 = P(-1.96 \leq Z \leq 1.96)$$

Hence

$$0.95 = P\left(-1.96 \leq \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \leq 1.96\right)$$

thus

$$0.95 \approx P\left(\tilde{\theta} - 1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}} \leq \theta \leq \tilde{\theta} + 1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}\right)$$

and so an approximate 95% CI is

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

INTERVAL ESTIMATION (C4S687)

Ways of finding interval estimates for an unknown parameter:

① Use a $100p\%$ likelihood interval.

② Use a $100p\%$ confidence interval if an exact pivotal quantity exists; or

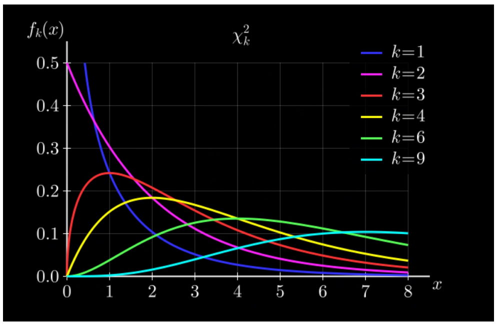
③ Use a $100p\%$ approximate confidence interval based on an approximate pivotal quantity (usually using CLT).

SAMPLE SIZE CALCULATION (C4S695)

1. Suppose we want to estimate θ, the proportion of units in a large population who have a specific characteristic, and we plan to select n units at random.
2. Suppose we use the $100p\%$ CI
$$\boxed{\hat{\theta} \pm a\frac{s}{\sqrt{n}}.}$$
3. We can specify we want a CI of width $\leq 2\ell$; ie
$$\boxed{a\frac{s}{\sqrt{n}} \leq \ell,}$$
or
$$\boxed{n \geq \left(\frac{a}{\ell}\right)^2 s^2,}$$
which tells us the minimum value of n needed for the CI to be of width at most 2ℓ.

CHI-SQUARED DISTRIBUTION: χ^2_k (C4S701)

1. The "chi-squared distribution" is parameterized by its degrees of freedom 'k'.
2. k affects the shape of the resulting pdf:
$$\boxed{\text{pdf} = \frac{1}{2^{k/2}\Gamma(k/2)} y^{\frac{k}{2}-1} e^{-\frac{y}{2}}.}$$
* not needed to know.



3. Properties:
   ① If $W_1, W_2, \ldots, W_n$ are iid with $W_i \sim \chi^2_{k_i}$, then
   $$\boxed{S = \sum_{i=1}^{n} W_i \sim \chi^2_{\sum k_i}.}$$
   ② If $Z \sim G(0,1)$, then
   $$\boxed{Z^2 \sim W \sim \chi^2_1.}$$
   ③ If $Z_1, \ldots, Z_n \sim G(0,1)$, then
   $$\boxed{S = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n.}$$
   ④ Also,
   $$\boxed{W \sim \chi_2 = \text{Exp}(2).}$$

LIKELIHOOD RATIO STATISTIC: $\Lambda(\theta)$ (C4S716)

1. The "likelihood ratio statistic" is defined to be
$$\boxed{\Lambda(\theta) = -2\log\left(\frac{L(\theta)}{L(\tilde{\theta})}\right) = -2\log\left(\frac{L(\theta;Y)}{L(\tilde{\theta};Y)}\right).}$$
* Λ is a random variable!
2. For large enough n, we can show
$$\boxed{\Lambda \sim \chi^2_1.}$$

LIKELIHOOD BASED CONFIDENCE INTERVAL (C4S724)

1. Note a $100p\%$ likelihood interval is an approximate $100q\%$ confidence interval, where
$$\boxed{q = P(W \leq -2\log(p)), \quad W \sim \chi^2_1.}$$
See slides for proof.
2. In particular, since $\Lambda(\theta) \sim \chi^2_1$ for large n, the likelihood interval can be written like
$$\boxed{\{\theta : R(\theta) \geq p\} = \left\{\theta : -2\log\left[\frac{L(\theta)}{L(\tilde{\theta})}\right] \leq -2\log(p)\right\}.}$$
3. Hence, the confidence coefficient is
$$\boxed{\begin{aligned} P(\Lambda(\theta) \leq -2\log(p)) &\approx P(W \leq -2\log(p)) \\ &= P(|Z| \leq \sqrt{-2\log(p)}) \\ &= 2P(Z < \sqrt{-2\log(p)}) + 1. \end{aligned}}$$

* $P(W \leq c) = 2P(Z \leq \sqrt{c}) - 1.$

GAUSSIAN DATA: UNKNOWN μ & σ (C4S741)

1. Let $Y_1, \dots, Y_n \sim G(\mu, \sigma)$, where μ & σ are unknown.

2. We can use the MLE estimator for μ:

$$\boxed{\tilde{\mu} = \bar{Y}.}$$

3. We use the point estimator for σ^2

$$\boxed{S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.}$$

- we prefer S^2 since $E(S^2) = \sigma^2$.

t-DISTRIBUTION: t_k (C4S747)

1. A rv T is said to have a "(Student's) t distribution" if its pdf is

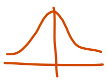
$$\boxed{f(t;k) = c_k \left(1 + \frac{t^2}{k}\right)^{-\frac{(k+1)}{2}}.}$$

where

$$\boxed{c_k = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)}.}$$

2. The parameter k is called the "degrees of freedom", and we write $T \sim t_k$ or $T \sim t(k)$.

3. Notes:
   ① The t distribution is unimodal and symmetric about 0;
   ② For large k, $t_k \approx G(0,1)$.



$Z \sim G(0,1)$, $U \sim \chi^2_k \Rightarrow \frac{Z}{\sqrt{U/k}} \sim t_k$ (C4S754)

1. Let $Z \sim G(0,1)$ & $U \sim \chi^2_k$ be independent. Then

$$\boxed{T \sim \frac{Z}{\sqrt{U/k}}}$$

has a t-distribution with k degrees of freedom.

$Y_1, \dots, Y_n \sim G(\mu, \sigma)$, μ, σ UNKNOWN $\Rightarrow$ $\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$ (C4S757)

1. First, see that if

$$\boxed{S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

then

$$\boxed{U = \frac{(n-1) S^2}{\sigma^2} \sim \chi^2_{n-1}.}$$

2. In particular, if $Y_1, \dots, Y_n \sim G(\mu, \sigma)$ where μ & σ are unknown, then

$$\boxed{\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.}$$

* this is a pivotal quantity!

CONFIDENCE INTERVAL FOR μ IF σ IS UNKNOWN (C4S760)

1. Let $Y_1, \dots, Y_n \sim G(\mu, \sigma)$ be iid where σ is unknown. Then necessarily a $100p\%$ CI for μ is

$$\boxed{\left(\bar{y} - a\frac{s}{\sqrt{n}},\ \bar{y} + a\frac{s}{\sqrt{n}}\right)}$$

* don't forget the $n-1$!!

where $P(T \le a) = \frac{1+p}{2}$, where $T \sim t_{n-1}$.

* in R, the command "pt(t, df)" returns $P(T \le t)$, where $T \sim t_{df}$.

* the command "qt(t, df)" returns t s.t. $P(T \le t) = q$, where $T \sim t_{df}$.

HOW PARAMETERS AFFECT WIDTH OF CI (C4S788)

1. The width of the CI is $2a\frac{s}{\sqrt{n}}$, where $P(T \le a) = \frac{1+p}{2}$, $T \sim t_{n-1}$.

2. Note that
   ① ↑ confidence level ⇒ new CI is wider
   ② ↑ sample size ⇒ new CI is narrower
      - as k ↑, t_k becomes less concentrated at peak
   ③ ↑ sample std dev ⇒ new CI is wider
   ④ ↑ (or ↓) sample mean ⇒ new CI's width is unchanged.

SAMPLE SIZE CALCULATION (C4S789)

1. In these, we assume σ is known.
   - since 's' depends on n.

2. Our CI is thus

$$\boxed{\bar{y} \pm a\frac{\sigma}{\sqrt{n}}, \quad P(Z \le a) = \frac{1+p}{2}.}$$

* we assume population std dev = sample std dev.

If we want this to have width 2ℓ, then we choose n such that

$$\boxed{n \approx \left(\frac{a\sigma}{\ell}\right)^2.}$$

CI FOR σ^2 (C4S796)

1. Let $Y_1, \dots, Y_n \sim G(\mu, \sigma)$ be iid, where μ & σ are unknown. Then we have shown

$$\boxed{W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}}$$

2. We pick a, b such that

$$\boxed{P(W \le a) = \frac{1-p}{2}, \quad P(W \ge b) = \frac{1-p}{2}}$$

or $P(W \le b) = \frac{1+p}{2}$

or in other words

$$P(a \le W \le b) = p.$$

* since χ^2 is not symmetric.

3. Thus, our coverage is

$$P\left(a \le \frac{(n-1)S^2}{\sigma^2} \le b\right) = p.$$

which can be rearranged to

$$\boxed{P\left(\frac{(n-1)S^2}{b} \le \sigma^2 \le \frac{(n-1)S^2}{a}\right) = p.}$$

4. Thus, a $100p\%$ CI for σ^2 is

$$\boxed{\left(\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right).}$$

* this is not symmetric!

★ in R, we use pchisq(w, df) & qchisq(w, df).

5. A $100p\%$ CI for σ is

$$\boxed{\left(\sqrt{\frac{(n-1)s^2}{b}}, \sqrt{\frac{(n-1)s^2}{a}}\right).}$$

Chapter 5: Tests of Hypothesis

NULL HYPOTHESIS: H_0 (CSS825)

1. The "null hypothesis" is a single "default" hypothesis.

   eg "the defendant is innocent"

2. Hypothesis testing is based on collecting data, and based on said data determining how plausible H_0 is.

ALTERNATIVE HYPOTHESIS: H_A/H_1 (CSS826)

1. The "alternative hypothesis" is the alternative to the null.
2. Often, H_A is just the negation of H_0.

   eg "the defendant is guilty"

TEST STATISTIC / DISCREPANCY MEASURE (CSS833)

1. A "test statistic" is a function of the data $D=g(Y)$ that is constructed to measure the "agreement" between the data Y & H_0.
2. Note:
   ① D is a random variable.
   ② If we observe $Y=y$, we use $d=g(y)$ to denote the observed value of D.
3. We usually define D so $d=0$ represents the best possible agreement between Y & H_0.
4. Note "large" values of d indicate poor agreement between Y & H_0.

P-VALUE (CSS845)

1. Suppose we use the test statistic $D=D(Y)$ to test the hypothesis H_0.
   Let $d=g(y)$ be the observed value of D.
   Then, the "p-value" of H_0 using D is

   $$\boxed{P(D \geq d;\ H_0).}$$

2. The p-value is the probability of observing a value of the test statistic greater than or equal to the observed value of the test statistic assuming H_0 is true.
3. In particular, a small p-value tells us that if H_0 were true, it would be unlikely to have observed data at least as surprising as the data we actually observed.

STEPS OF A HYPOTHESIS TEST (CSS849)

Steps:
① Specify H_0 to be tested using data Y.
② Define a test statistic $D(Y)$, where large values of D imply the data is less "consistent" with H_0.
③ Let $d=D(y)$; ie the 'observed value' of D.
④ Calculate the p-value $P(D \geq d;\ H_0)$
⑤ Draw a conclusion based on the p-value.

INTERPRETING THE p-VALUE (CSS852)

1. If $d=D(y)$ is large, and thus the p-value $P(D\geq d;\ H_0)$ is small, then either
   ① H_0 is true, but by chance we observed an event that is very unlikely when H_0 is true; or
   ② H_0 is false.
2. What does a "small" p-value mean?

p-value	There is — evidence against H_0 based on the data.
$0.1 < p$	no
$0.05 < p \leq 0.1$	weak
$0.01 < p \leq 0.05$	some
$0.001 < p \leq 0.01$	strong
$p \leq 0.001$	very strong

* these are only guidelines!

3. Depending on the p-value, we may state we 'reject', or fail to 'reject', the null hypothesis.

* we never accept H_0 (or H_1)!

TYPE I & II ERRORS: α, β (CSS861)

1. A "type I error" is the probability we reject H_0 when it is actually true.
   (ie false positive)
2. A "type II error" is the probability we fail to reject H_0 when it is actually false.
   (ie false negative)

POWER [OF A TEST] (CS864)

1. The "power" of a test is $1-\beta$, where β is the corresponding type 2 error.
2. A more powerful test is more desirable.
3. In particular,

   $$\boxed{\text{power} = P(\text{reject } H_0 \mid H_0 \text{ is false}).}$$

STEPS ON COMPUTING THE POWER OF A TEST (CSS871)

Steps to finding a test's power of $H_0: \theta=\theta_0$ against an alternative of $\theta \neq \theta_0$ at significance level α:
① Identify the "rejection region"; ie the test statistics that would lead us to reject H_0.
② For a specified value of $\theta \neq \theta_0$, compute the probability a sample would yield a test statistic in the rejection region.

P-HACKING (CSS878)

"p-hacking" is repeating experiments or being selective with one's results to falsely engineer a "significant" result.

ONE-SIDED TEST (CSS881)

$_1$ A "one-sided test" is a hypothesis test where

$$\boxed{\begin{array}{l} H_0: \ \theta = \theta' \\ H_1: \ \theta > \theta' \quad (\text{or} \quad \theta < \theta'). \end{array}}$$

$_2$ We may use the test statistic

$$\boxed{D = \max\{Y - \theta', \ 0\}.}$$

* symmetric for case where $\theta < \theta'$.

Our p-value is thus $\boxed{P(D \geq d)}$, where $d = y - \theta'$.

HYPOTHESIS TESTING FOR $G(\mu, \sigma)$ PARAMETERS

TESTING $H_0: \mu = \mu_0$, σ UNKNOWN (CSS895)

1. Let $Y_1, \dots, Y_n \sim G(\mu, \sigma)$ be a random sample, where σ is unknown.

   Recall
$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

2. To test $H_0: \mu = \mu_0$, we use the test statistic
$$\boxed{D = \frac{|\bar{Y} - \mu_0|}{S/\sqrt{n}}}.$$

   Why? Notice $E[\bar{Y}] = \mu_0$ if H_0 is true.
   Our question is "is $D = d$ surprisingly large"?
   ie our p-value is
$$P(D \geq d ; H_0) = P\left(\frac{|\bar{Y} - \mu_0|}{S/\sqrt{n}} \geq d ; H_0\right)$$
$$= P(|T| \geq d), \quad T \sim t_{n-1}$$
$$= 1 - P(-d \leq T \leq d).$$

3. Our p-value is thus
$$\boxed{p = 1 - P(-d \leq T \leq d)}$$
   where $T \sim t_{n-1}$.

4. In R, we may use
```
t.test(y, mu=1)
```

TESTS OF HYPOTHESIS & CIs (CSS914)

1. The p-value for testing $H_0: \theta = \theta_0$ is $\geq p$ iff the value $\theta = \theta_0$ is inside a $100(1-p)\%$ CI (using the same pivotal quantity).

   eg $H_0: \mu = \mu_0$ for $G(\mu, \sigma)$ data.
   Then p-value ≥ 0.05
$$\iff P\left(\frac{|\bar{Y} - \mu_0|}{S/\sqrt{n}} \geq \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}}\right) \geq 0.05$$
$$\iff P\left(|T| \geq \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}}\right) \geq 0.05, \quad T \sim t_{n-1}$$
$$\iff P\left(|T| \leq \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}}\right) \leq 0.95$$
$$\iff \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}} \leq a, \quad 0.95 = P(|T| \leq a)$$
$$\iff \mu_0 \in \left[\bar{y} - a\frac{s}{\sqrt{n}}, \ \bar{y} + a\frac{s}{\sqrt{n}}\right]$$
   which is a 95% CI for μ.

POWERING A STUDY (CSS919)

1. Usually, we fix α to a certain value and ask what sample size we require to attain a specified level of power.

2. For $\alpha = 0.05$, the power is
$$P\left(\bar{Y} < \mu_0 - 1.96\frac{S}{\sqrt{n}}\right) + P\left(\bar{Y} > \mu_0 + 1.96\frac{S}{\sqrt{n}}\right).$$
   So we should find n s.t.
$$1 - \beta = P\left(\bar{Y} < \mu_0 - 1.96\frac{\sigma}{\sqrt{n}}\right) + P\left(\bar{Y} > \mu_0 + 1.96\frac{S}{\sqrt{n}}\right).$$

3. More generally, to obtain power $1-\beta$ we seek n s.t.
$$\boxed{P\left(\bar{Y} < \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \beta,}$$
   where $\bar{Y} \sim G(\mu, \frac{\sigma}{\sqrt{n}})$, μ & σ are known, and $z_{\alpha/2}$ is such that $P(Z < a) = \frac{\alpha}{2}$ for $Z \sim G(0,1)$.

4. This is equivalent to
$$P\left(Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < \frac{\mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} - \mu}{\sigma/\sqrt{n}}\right) = 1 - \beta$$
   which can be rearranged to
$$\boxed{n \geq \left(\frac{\sigma(z_{\alpha/2} + z_{1-\beta})}{\mu_0 - \mu}\right)^2, \qquad P(Z < a) = 1 - \beta.}$$

   Proof. $P\left(Z < \frac{\mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} - \mu}{\sigma/\sqrt{n}}\right) = 1 - \beta \Rightarrow z_{1-\beta} < \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{\alpha/2}$
$$\Rightarrow \frac{\sqrt{n}}{\sigma} < \frac{z_{\alpha/2} + z_{1-\beta}}{\mu_0 - \mu}$$
$$\Rightarrow n \geq \left[\frac{\sigma(z_{\alpha/2} + z_{1-\beta})}{\mu_0 - \mu}\right]^2.$$

5. Note the result holds for $\mu_0 < \mu$ or $\mu_0 > \mu$.

TESTING $H_0: \sigma^2 = \sigma_0^2$, μ UNKNOWN (CSS932)

1. Recall
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

2. We can use the pivotal quantity
$$\boxed{U = \frac{(n-1)S^2}{\sigma_0^2}}.$$
   If $H_0: \sigma = \sigma_0$ is true, then it follows that
$$\boxed{U \sim \chi^2_{n-1}}.$$

3. Note large and small values of U provide evidence against H_0.

4. The p-value is then approximately
$$\boxed{p = \begin{cases} 2P(U \leq u), & P(U \leq u) < 0.5 \\ 2P(U \geq u), & P(U \geq u) \leq 0.5, \end{cases}}$$
   where
$$\boxed{U \sim \chi^2_{n-1}, \qquad u = \frac{(n-1)s^2}{\sigma_0^2}.}$$

☆ END of content for MT2.

LIKELIHOOD RATIO TEST STATISTIC

1. Recall the likelihood ratio statistic

$$\Lambda = -2\log\left(\frac{L(\theta; Y)}{L(\hat{\theta}; Y)}\right)$$

where $\hat{\theta}$ = MLE of θ.

2. For large n, we showed $\Lambda \sim \chi^2_{n-1}$.

TESTING $H_0: \theta = \theta_0$ USING Λ (C5S958)

To test $H_0: \theta = \theta_0$ using Λ:

① Find $L(\theta)$ & $R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}$ from the sample.

② Compute

$$\boxed{\lambda(\theta_0) = -2\log(R(\theta_0)) = -2\log\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right)}.$$

ie the observed value of Λ.

③ Find the p-value

$$\boxed{p = P(W \geq \lambda(\theta_0)) = 2\left[1 - P\left(Z \leq \sqrt{-2\log(R(\theta_0))}\right)\right]}$$

where $Z \sim G(0,1)$ & $W \sim \chi^2_1$.

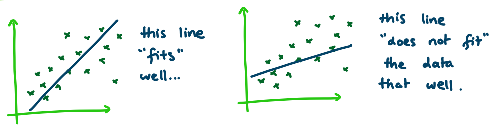
Chapter 6: Gaussian Response Models

1. **Idea:** we want to study the relationships between variates (in bivariate data).
2. One possible method: sample correlation.

$$\boxed{r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \quad S_{\alpha\beta} = \sum_{i=1}^{n}(\alpha_i-\bar{\alpha})(\beta_i-\bar{\beta}), \quad \alpha,\beta \in \{x,y\}}$$

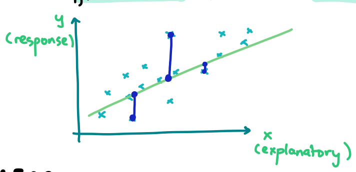
LEAST SQUARES ESTIMATES

How can we fit a straight line to bivariate data?



RESIDUALS (C6S978)

The "residuals" are the distances between the fitted line and the data.



LEAST SQUARES ESTIMATE (C6S979)

1. Usually, we find the fitted line $y = \alpha + \beta x$ that minimizes the sum of squares of the residuals.
2. Estimates of α & β, ie "$\hat{\alpha}$" & "$\hat{\beta}$", are called the "least squares estimate".
3. We want to find $\hat{\alpha}$ & $\hat{\beta}$ that minimizes

$$\boxed{g(\alpha,\beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2}$$

which are given by

$$\boxed{\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\beta} = \frac{S_{xy}}{S_{xx}}.}$$

Proof Sketch. We can get $\hat{\alpha}$ & $\hat{\beta}$ by solving the simultaneous eq^{n}s

$$\begin{cases} \dfrac{\partial g}{\partial \alpha} = \dfrac{\partial}{\partial \alpha}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2 = \sum_{i=1}^{n} 2(y_i-\alpha-\beta x_i)(-1) = 0 \\ \dfrac{\partial g}{\partial \beta} = \dfrac{\partial}{\partial \beta}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2 = \sum_{i=1}^{n} 2(y_i-\alpha-\beta x_i)(-x_i) = 0 \end{cases}$$

which resolves to $\alpha = \bar{y} - \beta\bar{x}$ & $\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)x_i = 0$.

Set $\alpha = \hat{\alpha}$, $\beta = \hat{\beta}$. Algebra gives us the desired result (see slides for full details.) ☒

4. In particular,

$$\boxed{\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \cdot \sqrt{\frac{S_{yy}}{S_{xx}}} = r\sqrt{\frac{S_{yy}}{S_{xx}}}.}$$

Thus, the sign of $\hat{\beta}$ = sign of r, and $\hat{\beta}$ and r are linearly related.

5. In R, we can do this using

```
> lm(y ~ x)
```

or we can create a "model object"

```
> mod <- lm(y~x)
> summary(mod)
```

SIMPLE LINEAR REGRESSION (C6S998)

1. **Idea:**
   ① σ — the sd of the x's (unknown)
   ② $\mu(x) = \alpha + \beta x$ — the mean y-value in the study population with x-value x.

* x = explanatory variate
y = response variate

2. In particular,
   - $\alpha = \mu(0)$ = mean y-value amongst data s.t $x=0$ (not really useful)
   - β represents the "increase" in the mean y-value in the study population for a one 'unit' increase in the x-value. (this is the same regardless of x.)
3. We assume $Y_i \sim G(\alpha+\beta x_i, \sigma)$ for $i=1,\dots,n$, and so σ represents the variability in the response variate Y in the study population for each value of the explanatory variate x.

LIKELIHOOD FUNCTION FOR α & β (C6S1004)

1. Our likelihood function for α & β is

$$\boxed{L(\alpha,\beta) = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2\right)}$$

(since we assume $Y_i \sim G(\alpha+\beta x_i, \sigma)$)

2. So, to maximize $L(\alpha,\beta)$, we minimize

$$\boxed{\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)}$$

but this is just the least squares problem!

3. Therefore, the MLEs of α & β are

$$\boxed{\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\beta} = \frac{S_{xy}}{S_{xx}}.}$$

REGRESSION PARAMETERS: $\hat{\alpha}$, $\hat{\beta}$

We call the values of $\hat{\alpha}$ and $\hat{\beta}$ above the "regression parameters".

DISTRIBUTION OF $\tilde{\beta}$ (C6S1011)

1. A corresponding estimator for β is

$$\boxed{\tilde{\beta} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i .}$$

2. If each $Y_i \sim G(\alpha + \beta x_i, \sigma)$ independently, then it follows that

$$\boxed{\tilde{\beta} \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right).}$$

Proof sketch. $\tilde{\beta}$ is a linear combination of Gaussian r.v. $\Rightarrow$ $\tilde{\beta}$ is also a Gaussian r.v. ☒

PIVOTAL QUANTITY OF β IF σ IS KNOWN (C6S1013)

1. Thus, if σ is known, a pivotal quantity for $\tilde{\beta}$ is

$$\boxed{\frac{\tilde{\beta} - \beta}{\sigma / \sqrt{S_{xx}}} \sim G(0,1).}$$

ESTIMATING σ^2 IN SIMPLE LINEAR REGRESSION & THE MEAN SQUARED ERROR: S_e^2 (C6S1014)

1. If σ^2 is unknown, we estimate it via

$$\boxed{S_e^2 = \frac{1}{n-2} \underline{\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2} = \frac{1}{n-2}(S_{yy} - \hat{\beta} S_{xy}) .}$$

the "mean squared error" the "sum of squared errors"

PIVOTAL QUANTITY OF β IF σ IS UNKNOWN (C6S1020)

1. We can show that

$$\boxed{\frac{(n-2) S_e^2}{\sigma^2} \sim \chi^2_{n-2} .}$$

*note $n-2$.

2. Hence, recall that

$$T = \frac{Z}{\sqrt{U/k}} \sim t_k, \quad U \sim \chi^2_k, \quad Z \sim G(0,1)$$

and so

$$\boxed{\frac{\tilde{\beta} - \beta}{S_e / \sqrt{S_{xx}}} \sim t_{n-2} .}$$

CI FOR β (C6S1021)

1. Thus, if $P(-a \leq T \leq a) = p$ for $T \sim t_{n-2}$, a $100p\%$ confidence interval for β is

$$\boxed{P\left(\tilde{\beta} - a\frac{S_e}{\sqrt{S_{xx}}} \leq \beta \leq \tilde{\beta} + a\frac{S_e}{\sqrt{S_{xx}}}\right).}$$

(or $P(T \leq a) = \frac{1+p}{2}$).

2. For testing $H_0 : \beta = \beta_0$, the p-value is

$$\boxed{p = 2\left[1 - P\left(T \leq \frac{|\hat{\beta} - \beta_0|}{S_e / \sqrt{S_{xx}}}\right)\right]}$$

where $T \sim t_{n-2}$.

HYPOTHESIS OF NO (LINEAR) RELATIONSHIP (C6S1023)

1. A discrepancy measure for testing $H_0 : \beta = \beta_0$ is

$$\boxed{\frac{|\hat{\beta} - \beta_0|}{S_e / S_{xx}} \sim t_{n-2}}$$

which is larger if the data are surprising if H_0 is true.

2. Since $\mu(x) = \alpha + \beta x$, a test of $H_0 : \beta = 0$ is a test of the hypothesis that $\mu(x)$ does not depend on x.

CI FOR $\mu(x) = \alpha + \beta x$ (C6S1034)

1. A point estimate for $\mu(x)$ is

$$\boxed{\hat{\mu}(x) = \hat{\alpha} + \hat{\beta} x = \bar{y} + \hat{\beta}(x - \bar{x})}$$

* x is a value we "plug in" to μ.

and so the corresponding estimator is

$$\boxed{\tilde{\mu}(x) = \bar{Y} + \tilde{\beta}(x - \bar{x}).}$$

2. We can show that

$$\boxed{\tilde{\mu}(x) = \sum_{i=1}^{n} \left(\frac{1}{n} + (x - \bar{x})\frac{(x_i - \bar{x})}{S_{xx}}\right) Y_i}$$

where $Y_i \sim G(\alpha + \beta x_i, \sigma)$ for each i.

3. This has distribution

$$\boxed{\tilde{\mu}(x) \sim G\left(\mu(x), \sigma\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}\right)}$$

where $\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta} x$ & $\mu(x) = \alpha + \beta x$.

4. Equivalently

$$\frac{\tilde{\mu}(x) - \mu(x)}{\sigma\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}} \sim G(0,1).$$

Since σ is unknown, we use the pivotal quantity

$$\boxed{\frac{\tilde{\mu}(x) - \mu(x)}{S_e\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}}$$

5. A $100p\%$ CI for $\mu(x) = \alpha + \beta x$ is

$$\boxed{\hat{\alpha} + \hat{\beta} x \pm a S_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}}$$

where $P(T \leq a) = \frac{1+p}{2}$ & $T \sim t_{n-2}$.

CI FOR α (C6S1038)

1. Since $\mu(0) = \alpha + \beta(0)$, a $100p\%$ CI for α is given by

$$\boxed{\hat{\alpha} \pm a S_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$$

CI FOR AN INDIVIDUAL RESPONSE Y AT x

Question:

"What is the CI for Y such that x?"

PREDICTION INTERVAL [FOR A FUTURE RESPONSE Y] (C6S1049)

1. A "100p% prediction interval" for a future response Y is

$$\boxed{\hat{\alpha} + \hat{\beta}x \pm a s_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}}$$

* note the +1 (which is absent in CIs).

where $P(T \leq a) = \frac{1+p}{2}$, $T \sim t_{n-2}$.

How do we get this?

Let Y = potential observation for given value of x.

We then have

$Y = \mu(x) + R$, $R \sim G(0, \sigma)$

independent of $Y_1, \dots, Y_n$. We established

$Y \sim G(\alpha + \beta x, \sigma)$ & $\tilde{\mu}(x) \sim G(\alpha + \beta x, \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}})$.

Then

$$\begin{aligned} Y - \tilde{\mu}(x) &= Y - \mu(x) + \mu(x) - \tilde{\mu}(x) \\ &= R + [\mu(x) - \tilde{\mu}(x)]. \end{aligned}$$

Note R is independent of $\tilde{\mu}(x)$ since it is not connected to the existing sample.

Thus the equation is a linear combination of ind, normally dist rv, and so is also normally dist.

$$\begin{aligned} \therefore E(Y - \tilde{\mu}(x)) &= E(R + [\mu(x) - \tilde{\mu}(x)]) \\ &= E(R) + E(\mu(x)) - E(\tilde{\mu}(x)) \\ &= 0 + \mu(x) - \mu(x) = 0. \end{aligned}$$

$$\begin{aligned} \therefore Var(Y - \tilde{\mu}(x)) &= Var(Y) + Var(\tilde{\mu}(x)) \\ &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right] \\ &= \sigma^2\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]. \end{aligned}$$

Thus

$$Y - \tilde{\mu}(x) = G(0, \sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}.$$

Thus

$$\frac{Y - \tilde{\mu}(x)}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0,1)$$

and since σ is unknown we use

$$\frac{Y - \tilde{\mu}(x)}{S_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

The corresponding CI is the one above.

2. In R, we can use

```
> predict(data, data.frame(x=75), interval='prediction')
```

GAUSSIAN RESPONSE MODELS

1. The general form of a Gaussian response model is

$$\boxed{Y_i \sim G(\mu(x_i), \sigma), \quad i=1,\ldots,n}$$

independently where x_i are assumed to be known constants (possibly vectors).

2. We can also write this as

$$\boxed{Y_i = \mu(x_i) + R_i.}$$

where $R_i \sim G(0, \sigma)$ independently.

3. In particular,
   ① $\mu(x_i)$ is a "deterministic" component;
   ② R_i is a "random" component.

LINEAR REGRESSION MODELS (C6S1056)

1. In "linear regression models", the deterministic component takes the form

$$\boxed{E(Y_i) = \mu(x_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}$$

So that $E(Y_i)$ is a linear function of a vector of explanatory variates for unit i $(x_i = (x_{i1}, \ldots, x_{ik}))$ & unknown parameters $\beta_0, \ldots, \beta_k$.

2. In particular,
   ① The β_j are called the "regression coefficients"; &
   ② The x_j are called the "covariates".

MULTIPLE LINEAR REGRESSION (C6S1058)

1. If we wish to fit the model

$$\boxed{E(Y_i) = \mu(x_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}$$

then we seek parameters $\beta_0, \ldots, \beta_k$ that minimize

$$\boxed{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij}\right)^2}$$

2. In R, we use

```
> mod <- lm(y ~ x1 + x2)
> summary(mod)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.01375    5.01527  -0.202  0.84133
x1           0.73142    0.07664   9.544 3.83e-10 ***
x2           0.28225    0.09850   2.866  0.00797 **
```

Then $\hat{\beta}_0 = -1.01375$
$\hat{\beta}_1 = 0.73142$
$\hat{\beta}_2 = 0.28225$

INTERPRETING $\hat{\beta}_j$ (C6S1060)

1. $\hat{\beta}_j$ can be interpreted as the amount of increase in response y when x_j increases by one unit when the other predictors $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ are held fixed.

HYPOTHESIS TEST OF $H_0: \beta_j = 0$ (C6S1061)

1. To test $H_0: \beta_j = 0$, $H_1: \beta_j \neq 0$, we use the test statistic

$$\boxed{t_j = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}} = \frac{\text{estimate}}{\text{std. error}}}$$

2. If H_0 is true, then

$$\boxed{t_j \sim t_{n-k-1}}$$

where $k = \#$ of parameters.

SUM OF SQUARED ERRORS / SSE (C6S1063)

1. The "sum of squared errors" is

$$\boxed{SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

where $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}$.

2. The smaller this is, the less 'error' in our model fit.

R^2 STATISTIC (C6S1064)

1. The "R^2 statistic" is

$$\boxed{R^2 = 1 - \frac{SSE}{S_{yy}}.}$$

2. In particular,

$$\boxed{R^2 = \frac{\text{variation explained by regression}}{\text{total variation}}}$$

ADJUSTED R^2-STATISTIC (C6S1065)

1. The "adjusted R^2" is

$$\boxed{\text{adj. } R^2 = 1 - \frac{SSE/(n-k-1)}{S_{yy}/(n-1)}}$$

where $k = \#$ of explanatory variates.

2. This tries to "compensate" that adding more & more variables will (potentially artificially) increase R^2.

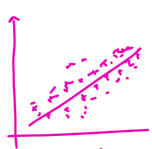
ASSUMPTIONS (C6S1067)

1. Assumptions we make for Gaussian linear response models:
   ① Y_i (given covariates x_i) has a Gaussian distribution with std dev σ which does not depend on the covariates; &
   ② $E(Y_i) = \mu(x_i)$ is a linear combination of known covariates $x_i = (x_{i1}, \ldots, x_{ik})$ and the unknown regression coefficients $\beta_0, \ldots, \beta_k$.

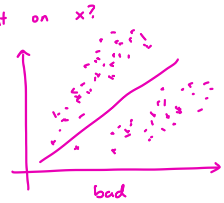
2. We must check these are suitable!

GRAPHICAL METHOD TO CHECK MODEL (C6S1069)

1. We can use graphical methods to do this: in particular,
   ① Make a scatterplot of y against x;
   ② Do the points seem to fit reasonably along a straight line?
      - ie is it linear
   ③ Are the points generally "spread out" to the same extent regardless of x?
      - ie is σ dependent on x?



good



bad
(gets more spread out as $x\uparrow$)

USING RESIDUALS TO CHECK MODEL (C6S1070)

1. We let our "fitted response" to be

$$\boxed{\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}}$$

- for simple LR, $\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x$.

2. The "residuals" are

$$\boxed{\hat{r}_i = y_i - \hat{\mu}_i \,;}$$

this represents what has been "left over" after the model has been fitted to the data.

3. We assume $Y_i \sim G(\mu(x_i), \sigma)$, and in particular,

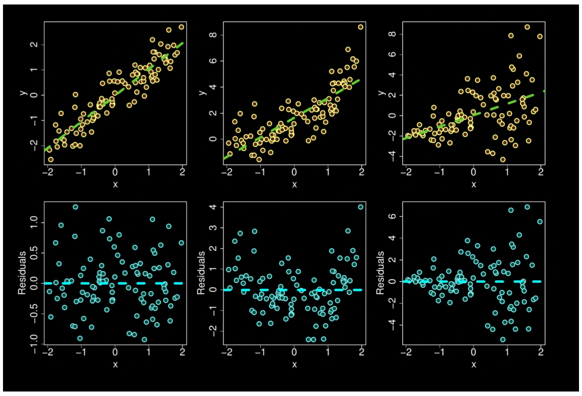
$$\boxed{Y_i = \mu_i + R_i,}$$

where $R_i \sim G(0, \sigma)$.

RESIDUAL PLOTS (C6S1073)

1. A "residual plot" is a plot of the points $(x_i, \hat{r}_i)$.

2. If the model assumptions hold, the points
   ① should lie more or less within a horizontal band around the line $\hat{r}_i = 0$
   ② with no obvious pattern.



STANDARDIZED RESIDUALS : $\hat{r}_i^*$ (C6S1077)

1. We define the "standardized residuals" to be

$$\boxed{\hat{r}_i^* = \frac{\hat{r}_i}{s_e} = \frac{y_i - \hat{\mu}_i}{s_e}.}$$

2. If we plot $(x_i, \hat{r}_i^*)$ instead of $(x_i, \hat{r}_i)$:
   ① The plot looks the same, but be "scaled";
   &
   ② The $\hat{r}_i^*$ values lie in the range $(-3, 3)$ since $\hat{r}_i^* \approx G(0,1)$.

RESIDUAL PLOT TYPE 2 (C6S1080)

1. If we have a more general linear model

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

we can plot $(\hat{\mu}_i, \hat{r}_i^*)$, where

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}.$$

2. We can use this to check the assumption about the form of $\mu(x_i)$;
   we check if the points appear randomly scattered around a horizontal line at 0.

QQ-PLOT OF RESIDUALS (C6S1082)

1. Since $\hat{r}_i^* \approx G(0,1)$, a QQ-plot should give approximately a straight line if the model assumptions hold.

MULTICOLLINEARITY (C6S1094)

1. "Multicollinearity" describes a situation when two (or more) of our explanatory variates are highly correlated.

2. This can occur when we have collected data on several variates on the same subject.

3. This can make us deduce incorrect conclusions.

PREDICTING BEYOND THE RANGE OF COVARIATES (C6S1097)

1. We may be tempted to predict an outcome for a covariate value outside the range of those in our dataset.

2. However,
   ① Our model assumptions may no longer hold, and we have no way of checking them.
   ② Our predictions might also not make sense.

BINARY OUTCOMES (C6S1099)

ODDS RATIO [OF AN EVENT]: odds(E) (C6S1105)

1. The "odds" of an event E is

$$\boxed{\text{odds}(E) = \frac{P(E)}{1-P(E)}.}$$

2. If $\text{odds}(E) = \frac{a}{b}$ ("the odds of E are a to b"), then

$$\boxed{P(E) = \frac{a}{a+b}.}$$

GENERALIZED LINEAR MODELS / GLMs (C6S1108)

1. "GLMs" have the following properties:
   ① A probability distribution for the outcome variable;
   ② A linear model $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$; &
   ③ A "link function" relating the linear model to the parameters of the outcome distribution.

LOGISTIC REGRESSION (C6S1109)

1. "Logistic regression" is a GLM for binary outcome data.
2. Assumptions:
   ① Outcome can be modelled by a binomial rv;
   &
   ② We want to model p, the probability of success.
3. A common link function is "logit":

$$\boxed{g(p) = \text{logit}(p) = \log\left(\frac{p}{1-p}\right), \quad p \in [0,1]}$$

which maps from $[0,1] \to [-\infty, +\infty]$.

4. This is the log odds of success.

USING logit IN LOGISTIC REGRESSION (C6S1110)

1. Assume we had a single explanatory variate x_i, and let p_i be the probability unit i experiences the outcome.
2. We can use logit to relate this probability to a linear model of our data:

$$\boxed{g(p_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i}$$

3. This can be rewritten as

$$\boxed{p_i = \frac{1}{1+\exp(-(\beta_0+\beta_1 x_i))}}$$

4. In R, we use

```
< mod1 <- glm( offer ~ grade , family = 'binomial' )
< summary (mod1)
```

to get the estimates for the linear model.

5. We use

```
< predict( mod1, newdata = data.frame(grade=80), type='response')
```

to get the odds directly.

ODDS RATIO & LOG ODDS RATIO (C6S1121)

1. Let O_1 be the odds of E_1, & O_2 be the odds of E_2.
2. The "odds ratio" of E_1 relative to E_2 is

$$\boxed{\text{odds ratio} = \frac{O_1}{O_2}.}$$

3. The "log odds ratio" of E_1 relative to E_2 is

$$\boxed{\text{log odds ratio} = \log\left(\frac{O_1}{O_2}\right).}$$

4. If $\eta_i = \beta_0 + \beta_1 x_i$, then $\hat{\beta}_1$ is our estimate of the log odds ratio of a one unit increase in x.

ASSUMPTIONS FOR LOGISTIC REGRESSION (C6S1128)

1. Assumptions for logistic regression:
   ① Events are independent; &
   ② A linear relationship exists between predictors and the log odds.
2. One option: split the data into tertiles / quantiles etc.

COMPARING MEANS OF TWO POPULATIONS

TWO-SAMPLE GAUSSIAN PROBLEM (C6S1139)

1. A "two-sample Gaussian problem" involves

$$\boxed{\begin{aligned} &Y_{1i} \sim G(\mu_1, \sigma), \quad i=1,\dots,n_1 \quad \text{independently; } \& \\ &Y_{2i} \sim G(\mu_2, \sigma), \quad i=1,\dots,n_2 \quad \text{independently.}\end{aligned}}$$

2. This is a special case of the Gaussian response model.

HYPOTHESIS TEST THAT TWO MEANS ARE THE SAME (C6S1140)

1. To check if $\mu_1 = \mu_2$, we use

$H_0: \mu_1 = \mu_2$

or equivalently

$$\boxed{H_0: \mu_1 - \mu_2 = 0.}$$

POINT ESTIMATORS FOR μ_1 & μ_2 (C6S1141)

1. First, note the ML estimators for μ_1 & μ_2 are

$$\boxed{\begin{aligned} \tilde{\mu}_1 &= \bar{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n} Y_{1i} \\ \tilde{\mu}_2 &= \bar{Y}_2 = \frac{1}{n_2}\sum_{i=1}^{n} Y_{2i}\end{aligned}}$$

and so a point estimator for $\mu_1 - \mu_2$ is

$$\boxed{\tilde{\mu}_1 - \tilde{\mu}_2 = \bar{Y}_1 - \bar{Y}_2.}$$

2. Since $\tilde{\mu}_1 = \bar{Y}_1 \sim G(\mu_1, \frac{\sigma}{\sqrt{n_1}})$, $\tilde{\mu}_2 = \bar{Y}_2 \sim G(\mu_2, \frac{\sigma}{\sqrt{n_2}})$ independently, it follows that

$$\boxed{\tilde{\mu}_1 - \tilde{\mu}_2 = \bar{Y}_1 - \bar{Y}_2 \sim G\left(\mu_1 - \mu_2, \ \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)}$$

and so

$$\boxed{\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim G(0,1).}$$

POINT ESTIMATOR FOR σ; THE POOLED ESTIMATOR FOR VARIANCE: S_p^2 (C6S1143)

1. First, define

$$\boxed{\begin{aligned} S_1^2 &= \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(Y_{1i} - \bar{Y}_1)^2 \\ S_2^2 &= \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_{2i} - \bar{Y}_2)^2\end{aligned}}$$

which are the point estimators of σ^2 based on only the Y_{1i}, & only the Y_{2i}, respectively.

2. Our point estimator of σ^2 is then

$$\boxed{\begin{aligned} S_p^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \\ &= \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}(Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2}(Y_{2i} - \bar{Y}_2)^2\right]\end{aligned}}$$

which is the "pooled estimator of variance".

3. Note $E(S_p^2) = \sigma^2$, so the estimator is unbiased. (It is NOT the ML estimator.)

4. Our pivotal quantity for S_p^2 is

$$\boxed{\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2_{n_1 + n_2 - 2}.}$$

PIVOTAL QUANTITY FOR $\mu_1 - \mu_2$ (C6S1146)

From the previous results, thus

$$\boxed{\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}}$$

(since $Z \sim G(0,1)$, $U \sim \chi^2_k$ independently $\Rightarrow T = \frac{Z}{\sqrt{U/k}} \sim t_k$)

CI FOR $\mu_1 - \mu_2$ (C6S1147)

A $100p\%$ CI for $\mu_1 - \mu_2$ is thus

$$\boxed{CI = \bar{y}_1 - \bar{y}_2 \pm a s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $P(T \le a) = \frac{1+p}{2}$ & $T \sim t_{n_1 + n_2 - 2}$.

TEST STATISTIC FOR $H_0: \mu_1 - \mu_2 = 0$ (C6S1149)

1. The test statistic for $H_0: \mu_1 - \mu_2 = 0$ is

$$\boxed{D = \frac{|\bar{Y}_1 - \bar{Y}_2 - 0|}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$$

with observed value

$$\boxed{d = \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.}$$

2. In particular, large values of d would be surprising if H_0 is true.

p-VALUE FOR $H_0: \mu_1 - \mu_2 = 0$ (C6S1150)

The p-value is thus

$$\boxed{p = 2[1 - P(T \le d)]}$$

where $T \sim t_{n_1 + n_2 - 2}$.

COMPARISON OF 2 MEANS WITH UNEQUAL VARIANCES (C6S1156)

APPROXIMATE PIVOTAL QUANTITY FOR $\mu_1 - \mu_2$ (C6S1157)

1. Suppose instead that

$$\boxed{\begin{aligned} &Y_{1i} \sim G(\mu_1, \sigma_1), \quad i=1,\dots,n_1 \quad \text{independently} \\ &Y_{2i} \sim G(\mu_2, \sigma_2), \quad i=1,\dots,n_2 \quad \text{independently}\end{aligned}}$$

where we don't assume $\sigma_1 = \sigma_2$.

2. If n_1, n_2 are large (≥ 30), we can use the approximate pivotal quantity

$$\boxed{\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx G(0,1).}$$

3. Thus, an approximate CI for $\mu_1 - \mu_2$ is

$$\boxed{CI = \mu_1 - \mu_2 \pm a\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $P(Z \le a) = \frac{1+p}{2}$, $Z \sim G(0,1)$.

PAIRED DATA

1. "Paired data" considers scenarios where the Y_{1i}'s are related to the Y_{2i}'s.

eg Y_{1i} = movies, Y_{2i} = their sequels

2. Suppose once again that

$$\boxed{\begin{array}{ll} Y_{1i} \sim G(\mu_1, \sigma_1), & i=1,\dots,n_1 \text{ independently} \\ Y_{2i} \sim G(\mu_2, \sigma_2), & i=1,\dots,n_2 \text{ independently} \end{array}}$$

but the set of Y_{1i}, Y_{2i} is not independent with each other.

3. Then

$$\boxed{Var(\bar{Y}_1 - \bar{Y}_2) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2Cov(Y_{1i}, Y_{2i}).}$$

which is smaller than for an unpaired experiment.

4. To make inferences about $\mu_1 - \mu_2$, we analyze the within-pair differences

$$\boxed{Y_i = Y_{1i} - Y_{2i} \quad \forall i = 1,\dots,n}$$

by assuming

$$\boxed{Y_i = Y_{1i} - Y_{2i} \sim G(\mu_1 - \mu_2, \sigma) \quad \forall i = 1,\dots,n}$$

independently.

5. To test $H_0: \mu_1 - \mu_2 = 0$, we use the test statistic

$$\boxed{D = \frac{|\bar{Y} - 0|}{s/\sqrt{n}} \sim t_{n-1} \text{ (if } H_0 \text{ is true).}}$$

and our p-value is

$$\boxed{p = 2(1 - P(T \leq d))}$$

where $T \sim t_{n-1}$.

Chapter 7: Multinomial Models & Goodness of Fit Tests

MULTINOMIAL LIKELIHOOD FUNCTION (C7S1184)

1. We consider

$$\boxed{(Y_1, \dots, Y_k) \sim \text{Multinomial}(n, \theta_1, \dots, \theta_k).}$$

2. The likelihood function based on $y_1, \dots, y_k$ is

$$\boxed{\begin{aligned} L(\theta_1, \dots, \theta_k) &= P(Y_1 = \theta_1, \dots, Y_k = \theta_k \,;\, \theta_1, \dots, \theta_k) \\ &= \frac{n!}{y_1! \cdots y_k!} \theta_1^{y_1} \cdots \theta_k^{y_k}. \end{aligned}}$$

Ditching the constant gets us that

$$\boxed{L(\theta_1, \dots, \theta_k) = \theta_1^{y_1} \cdots \theta_k^{y_k}.}$$

TESTING $H_0: \theta_1 = \dots = \theta_k$ (C7S1184)

1. Suppose we wish to test

$$\boxed{H_0: \theta_1 = \dots = \theta_k \iff H_0: \theta = \theta_0 = \left(\tfrac{1}{k}, \dots, \tfrac{1}{k}\right).}$$

2. We note that the MLE of θ_j is

$$\boxed{\hat{\theta}_j = \frac{y_j}{n}}.$$

LIKELIHOOD RATIO TEST STATISTIC FOR $H_0: \theta = \theta_0$ (C7S1187)

1. The likelihood ratio test statistic for H_0 is then

$$\boxed{\Lambda(\theta_0) = -2\log\left(\frac{L(\theta_0)}{L(\tilde{\theta})}\right)}$$

where

$$\boxed{\begin{aligned} \tilde{\theta} &= \left(\frac{Y_1}{n}, \dots, \frac{Y_k}{n}\right) \\ \theta_0 &= \left(\frac{1}{k}, \dots, \frac{1}{k}\right) \\ L(\theta_1, \dots, \theta_k) &= \theta_1^{y_1} \cdots \theta_k^{y_k} = \prod_{j=1}^{k} \theta_j^{y_j}. \end{aligned}}$$

2. This evaluates to

$$\boxed{\Lambda(\theta_0) = 2\sum_{j=1}^{k} Y_j \log\left(\frac{\overbrace{Y_j}^{\text{observed}}}{\underbrace{E_j}_{\text{expected}}}\right), \quad E_j = \frac{n}{k}}$$

where E_j is the "expected frequency" of Y_j.

p-VALUE FOR $H_0: \theta = \theta_0$ (C7S1193)

1. The observed value of $\Lambda(\theta_0)$ is

$$\boxed{\lambda(\theta_0) = 2\sum_{j=1}^{k} y_j \log\left(\frac{y_j}{e_j}\right)}$$

where $e_j = \frac{n}{k}$.

2. For a sufficiently large sample, we can use the p-value

$$\boxed{p\text{-value} = P(W \geq \lambda(\theta_0)), \quad W \sim \chi^2_{k-1-p}}$$

where p = # of parameters estimated in forming H_0.

* degrees of freedom = $k-1-p$ because these are the parameters that are "free to move".
(When we estimate parameters, they are "locked").

3. Guideline: "sufficiently large" roughly implies $e_j \geq 5$ for all j.
- if this is not satisfied, we can "collapse" two or more adjacent categories with the smallest expected probabilities.

PEARSON'S χ^2 GOODNESS OF FIT STATISTIC (C7S1201)

1. The "Pearson's χ^2 test statistic" is

$$\boxed{D = \sum_{j=1}^{k} \frac{(Y_j - E_j)^2}{E_j} \sim \chi^2_{k-1-p}}$$

2. For large n, D & Λ are asymptotically equivalent & have the same asymptotic χ^2 distribution.

3. Note we need to account for any parameters we estimated.

eg If we assume $Poi(\theta)$, we need to account for

- sample size, n
- MLE of θ, $\hat{\theta}$

$\Big\}$ so degrees of freedom $v = k-1-1 = k-2$.

4. Our p-value is thus

$$\boxed{p = P(W \geq d), \quad d = \sum_{j=1}^{k} \frac{(y_j - e_j)^2}{e_j}, \quad W \sim \chi^2_{k-1-p}.}$$

TWO-WAY TABLES & INDEPENDENCE TESTS (C7S1218)

1. "2-way tables" have the following form:

	B	$\bar{B}$	total
A	y_{11}	y_{12}	$r_1 = y_{11} + y_{12}$
$\bar{A}$	y_{21}	y_{22}	$n - r_1$
total	$c_1 = y_{11} + y_{21}$	$n - c_1$	n

2. We are concerned whether there is a relationship between A & B, and in particular, whether they are independent.

MODEL FOR TEST OF INDEPENDENCE (C7S1224)

1. We define the random variables

$$\boxed{\begin{aligned} Y_{11} &= \#\text{ of } A \cap B \text{ outcomes} \\ Y_{12} &= \#\text{ of } A \cap \bar{B} \text{ outcomes} \\ Y_{21} &= \#\text{ of } \bar{A} \cap B \text{ outcomes} \\ Y_{22} &= \#\text{ of } \bar{A} \cap \bar{B} \text{ outcomes.} \end{aligned}}$$

2. Then our model is

$$\boxed{(Y_{11}, Y_{12}, Y_{21}, Y_{22}) \sim \text{Multinomial}(n, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})}$$

where

$$\boxed{\begin{aligned} \theta_{11} &= P(A \cap B) & \theta_{12} &= P(A \cap \bar{B}) \\ \theta_{21} &= P(\bar{A} \cap B) & \theta_{22} &= P(\bar{A} \cap \bar{B}) \end{aligned}}$$

HYPOTHESIS OF INDEPENDENCE (C7S1227)

1. To test whether A & B are independent, we use the null hypothesis

$$\boxed{H_0: P(A \cap B) = P(A)P(B)}$$

2. This is equivalent to

$$\boxed{H_0: \theta_{11} = \alpha\beta, \quad \alpha = P(A), \quad \beta = P(B)}$$

LIKELIHOOD FUNCTION FOR $H_0: \theta_{11} = \alpha\beta$ (C7S1229)

1. The likelihood function (ignoring constants) is

$$\boxed{L(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}) = \theta_{11}^{y_{11}} \theta_{12}^{y_{12}} \theta_{21}^{y_{21}} \theta_{22}^{y_{22}}}$$

2. The ML estimates are

$$\boxed{\hat{\theta}_{ij} = \frac{y_{ij}}{n}, \qquad i = 1,2, \; j = 1,2}$$

with corresponding estimators

$$\boxed{\tilde{\theta}_{ij} = \frac{Y_{ij}}{n}, \qquad i = 1,2, \; j = 1,2}$$

PARAMETER ESTIMATION UNDER $H_0: \theta_{11} = \alpha\beta$ (C7S1230)

1. If $H_0: \theta_{11} = \alpha\beta$ is true, then the likelihood function is

$$\boxed{\begin{aligned} L(\theta) &= \theta_{11}^{y_{11}} \theta_{12}^{y_{12}} \theta_{21}^{y_{21}} \theta_{22}^{y_{22}} \\ \Rightarrow L(\alpha, \beta) &= (\alpha\beta)^{y_{11}} [\alpha(1-\beta)]^{y_{12}} [(1-\alpha)\beta]^{y_{21}} [(1-\alpha)(1-\beta)]^{y_{22}} \\ &= \alpha^{y_{11}+y_{12}} (1-\alpha)^{y_{21}+y_{22}} \beta^{y_{11}+y_{21}} (1-\beta)^{y_{12}+y_{22}}. \end{aligned}}$$

2. The ML estimates under H_0 are

$$\boxed{\hat{\alpha} = \frac{y_{11} + y_{12}}{n}, \qquad \hat{\beta} = \frac{y_{11} + y_{21}}{n}}$$

with corresponding ML estimators

$$\boxed{\tilde{\alpha} = \frac{Y_{11} + Y_{12}}{n}, \qquad \tilde{\beta} = \frac{Y_{11} + Y_{21}}{n}.}$$

LIKELIHOOD RATIO TEST STATISTIC FOR $H_0: \theta_{11} = \alpha\beta$ (C7S1233)

1. The likelihood ratio statistic is

$$\boxed{\begin{aligned}\Lambda(\theta) &= -2\log\left(\frac{L(\tilde\alpha,\tilde\beta)}{L(\tilde\theta_{11},\tilde\theta_{12},\tilde\theta_{21},\tilde\theta_{22})}\right)\\ &= -2\log\left(\frac{\tilde\alpha^{Y_{11}+Y_{12}}(1-\tilde\alpha)^{Y_{21}+Y_{22}}\,\tilde\beta^{Y_{11}+Y_{21}}(1-\tilde\beta)^{Y_{12}+Y_{22}}}{\tilde\theta_{11}^{Y_{11}}\,\tilde\theta_{12}^{Y_{12}}\,\tilde\theta_{21}^{Y_{21}}\,\tilde\theta_{22}^{Y_{22}}}\right).\\ &\approx \chi^2_1\end{aligned}}$$

where

degrees of freedom $= \underbrace{4}_{k} - \underbrace{1}_{\text{sample size, } n} - \underbrace{2}_{\text{since we are estimating } \alpha \,\&\, \beta} = 1.$

2. This is also equivalent to

$$\boxed{\Lambda(\theta) = 2\left[Y_{11}\log\left(\frac{Y_{11}}{E_{11}}\right) + Y_{12}\log\left(\frac{Y_{12}}{E_{12}}\right) + Y_{21}\log\left(\frac{Y_{21}}{E_{21}}\right) + Y_{22}\log\left(\frac{Y_{22}}{E_{22}}\right)\right]}$$

where

$$\boxed{\begin{aligned}&E_{11} = n\hat\alpha\hat\beta, && E_{12} = n\hat\alpha(1-\hat\beta),\\ &E_{21} = n(1-\hat\alpha)\hat\beta, && E_{22} = n(1-\hat\alpha)(1-\hat\beta).\end{aligned}}$$

* using $\Lambda(\theta) = 2\sum_{j=1}^{k} Y_j \log\left(\frac{Y_j}{E_j}\right)$.

3. Our p-value is thus approximately

$$\boxed{p = P(W \geq \lambda) = 2[1 - P(Z \leq \sqrt{\lambda})], \quad W \sim \chi^2_1, \; Z \sim G(0,1)}$$

where λ is our observed value of Λ.

LARGER TWO-WAY TABLES (C7S1254)

1. Let Y_{ij} = number of individuals in category A_i & category B_j in a sample size of n.

2. Let

$$\boxed{\theta_{ij} = P(A_i B_j).}$$

Then our model is

$$\boxed{(Y_{11}, Y_{12}, \ldots, Y_{ab}) \sim \text{Multinomial}(n; \theta_{11}, \theta_{12}, \ldots, \theta_{ab}).}$$

HYPOTHESIS FOR INDEPENDENCE (C7S1256)

1. Let

$$\boxed{\alpha_i = P(A_i), \quad \beta_j = P(B_j).}$$

2. To test if A & B are independent variates, we test

$$\boxed{H_0: \theta_{ij} = \alpha_i\beta_j \quad \forall i = 1, \ldots, a, \; j = 1, \ldots, b.}$$

3. Under H_0, we can show the expected frequencies e_{ij} are

$$\boxed{e_{ij} = \frac{r_i c_j}{n}, \quad i = 1, \ldots, a, \quad j = 1, \ldots, b.}$$

where

$$\boxed{\begin{aligned}r_i &= \#\text{ of outcomes under } A_i\\ c_j &= \#\text{ of outcomes under } B_j.\end{aligned}}$$

total of row i

total of column j.

LIKELIHOOD RATIO TEST STATISTIC (C7S1259)

1. The likelihood ratio test statistic is then

$$\boxed{\Lambda = 2\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij}\log\left(\frac{Y_{ij}}{E_{ij}}\right)}$$

with observed value

$$\boxed{\lambda = 2\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}\log\left(\frac{y_{ij}}{e_{ij}}\right).}$$

2. In particular, if H_0 is true and $e_{ij} \geq 5 \;\; \forall i,j$, then

$$\boxed{\Lambda \approx \chi^2_{(a-1)(b-1)}.}$$

Proof. degrees of freedom,

$$\begin{aligned}\nu &= k - 1 - p\\ &= ab - 1 - \underbrace{(a-1)}_{\alpha_1, \ldots, \alpha_{a-1}} - \underbrace{(b-1)}_{\beta_1, \ldots, \beta_{b-1}}\\ &= (a-1)(b-1). \quad \#\end{aligned}$$

* Since $\alpha_a = 1 - \alpha_1 - \cdots - \alpha_{a-1}$, we don't estimate α_a! Similar with β_b.

3. Our p-value is thus

$$\boxed{p = P(W \geq \lambda),}$$

where $W \sim \chi^2_{(a-1)(b-1)}$.

Chapter 8: Causality or Relationship?

CAUSAL EFFECT (C8S1278)

1. We say x has a "causal effect" on Y if, when all other factors that affect Y are held constant, a change in x induces a change in a property of the distribution of Y.

   * this is impractical since we cannot hold all factors that affect Y to be constant!

2. We should design studies so that alternative explanations of what causes changes in the distribution of Y can be ruled out, leaving x as the causal agent.

REASONS 2 VARIATES CAN BE RELATED

EXPLANATORY VARIATE IS THE DIRECT CAUSE OF THE RESPONSE VARIATE (C8S1281)

1. Reason 1: A change in the explanatory variate directly causes a change in the response variate.

   eg drinking tea & thirst

2. Note that even in this case, we may not see a strong association.

   eg playing the lottery & winning the lottery

RESPONSE VARIATE IS THE DIRECT CAUSE OF THE EXPLANATORY VARIATE (C8S1283)

Reason 2: Similar to Reason 1, but now the causal relationship is "flipped"; the response variate directly causes the explanatory variate.

THE EXPLANATORY VARIATE IS A CONTRIBUTING, BUT NOT ONLY, CAUSE OF THE RESPONSE VARIATE (C8S1285)

1. Reason 3: The explanatory variate is a contributing cause, but not the sole cause, of the response variate.

   eg diet & type of cancer

2. In particular, we may think we have found a sole cause, when in actuality we have found a necessary contributor to the outcome.

   eg HIV & AIDS
   - we need HIV to get AIDS (so it is a necessary contributor)
   - but HIV is not necessarily the sole cause of AIDS (there might be other factors).

BOTH VARIATES ARE CHANGING OVER TIME (C8S1287)

Reason 4: Non-sensical associations can result from correlating two variates that are both changing over time.

eg global avg temp. & # of pirates
- they both decrease as time increases
- but are not related in any way

THE ASSOCIATION MAY BE NOTHING MORE THAN COINCIDENCE (C8S1289)

Reason 5: The association may be nothing more than coincidence.

BOTH VARIATES MAY RESULT FROM A COMMON CAUSE — CONFOUNDING / LURKING VARIATES (C8S1292)

1. Reason 6: An association between 2 variates may be observed because both variates are responding to changes in some unobserved variate or variates.

$$X \leftarrow Z \rightarrow Y$$

2. These variates are called "confounding" variates.

SIMPSON'S PARADOX (C8S1294)

"Simpson's paradox" describes the phenomenon where the association between 2 categorical variables is different than the association after controlling for one or more variables.

eg

Age	Coke	Pepsi
< 30	93% (81/87)	87% (234/270)
≥ 30	73% (192/263)	69% (55/80)
Total	78% (273/350)	83% (289/350)

- for each individual row, Coke is bigger
- but for the total, Pepsi is bigger!

→ age is a confounding variate.

THE IMPORTANCE OF RANDOMIZATION (C8S1301)

Randomization is important since it ensures the other variates will be approximately equally distributed across the categories.

ESTABLISHING CAUSATION IN OBSERVATIONAL STUDIES (C8S1304)

To establish causation in observational studies, we need at least the following 4 features:

① The association between the 2 variates must be observed in many studies of different types among different groups.
   - this reduces the chance an observed association is due to a defect in one type of study
   - or from a peculiarity in one group of subjects

② The association must continue to hold when the effects of plausible confounding variates are taken into account.

③ There must be a plausible scientific explanation for the direct influence of one variate on the other variate.
   - so a causal link does not depend on the observed association alone.

④ There must be a consistent response; ie one variate increases (or decreases) as the other variate increases.

CAUSAL INFERENCE (C8S1308)

1. "Causal inference" is concerned with identifying causal, merely than just associative, relationships.
2. We are interested in causal effects, and quantifying the effect of a change in some variate on some outcome.

COUNTERFACTUALS (C8S1311)

1. We write $Y(1)$ for the result of a response variate given one value of the explanatory variate, and similarly $Y(0)$ for another value of the explanatory variate.
2. We can then write the causal effect as

$$\boxed{Y(1) - Y(0).}$$

3. If we let $Y_i(1)$ & $Y_i(0)$ be the corresponding random variables for unit i, then it follows that the "average causal effect" is roughly

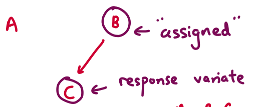
$$\boxed{\tau = E(Y(1) - Y(0)).}$$

4. Note for each unit we can only observe one of $Y(0)$ or $Y(1)$; the other is called a "counterfactual" or "potential outcome".

COVARIATE DEPENDENT TREATMENT (C8S1315)

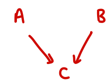
1. Consider the following 3 scenarios:

#1



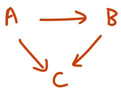
No relationship between A & C
We measure how B changes with C

#2



Relationship between A & C
We measure how B changes with C

#3



B is chosen based on A
Relationship exists between A & C
We measure how B changes with C

2. For scenarios #1 & #2, we can perform a standard analysis:
   1. Let Y_0, Y_1 be the r.v. denoting the value of C given B or $\neg$B respectively;
   2. Suppose $Y_0 \sim G(\mu_0, \sigma_0)$ & $Y_1 \sim G(\mu_1, \sigma_1)$.
   3. Carry out a test of $H_0: \mu_0 = \mu_1$ using methods from earlier in this course.
3. However, this may not necessarily work in scenario #3.
   → since A influences B, B's effects are exaggerated.

INVERSE PROBABILITY WEIGHTING (C8S1326)

PROPENSITY SCORE: $\pi(x)$ (C8S1327)

1. Let $\mu_0 = E(Y(0))$ & $\mu_1 = E(Y(1))$.
   The "propensity score" is

$$\boxed{\pi(x) = P(A=1 \mid X=x),}$$

   which is the probability A is true given covariate value x.
2. We write $\hat{\pi}(x)$ for our estimate of $\pi(x)$ for a given x.

ESTIMATING μ_0 & μ_1 (C8S1333)

1. We can estimate $\hat{\mu}_0$ & $\hat{\mu}_1$ by

$$\boxed{\begin{aligned}\hat{\mu}_1 &= \frac{1}{n}\sum_{i=1}^{n}\frac{y_i \cdot 1_{\{A_i=1\}}}{\hat{\pi}(x_i)} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i a_i}{\hat{\pi}(x_i)} \\ \hat{\mu}_0 &= \frac{1}{n}\sum_{i=1}^{n}\frac{y_i \cdot 1_{\{A_i=0\}}}{1-\hat{\pi}(x_i)} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i(1-a_i)}{1-\hat{\pi}(x_i)}\end{aligned}}$$

   where $1_{\{P\}}$ is the indicator function for the statement P.
2. Intuition: we give more weight to more unusual data in $\hat{\mu}_1$, and the other way around in $\hat{\mu}_0$.
3. Note the above $\hat{\mu}_0$ & $\hat{\mu}_1$ are unbiased.

ESTIMATING THE CAUSAL EFFECT (C8S1336)

1. We can estimate the causal effect

$$\boxed{\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0.}$$

2. This is also unbiased.
   → since $\hat{\mu}_1, \hat{\mu}_0$ are unbiased.

ASSUMPTIONS FOR IPW (C8S1342)

1. Assumptions needed to perform IPW:
2. Consistency — a counterfactual outcome is equal to that which would have been observed had "B" been different.
   - we set

$$\boxed{Y_i = Y_i(0)(1-A_i) + Y_i(1)A_i.}$$

3. Stable Unit Treatment Value Assumption (SUTVA) — there is no interference in the value of the B's.
   - often reasonable, but not always!
   - eg vaccines
4. No Unmeasured Confounders (NUC) — the assignment of the B's is independent of the potential outcomes, given covariates.
   - we write

$$\boxed{Y(0), Y(1) \perp\!\!\!\perp A \mid X}$$

5. Positivity — there is a non-zero probability of assignment to a category of B for all subjects: ie

$$\boxed{0 < P(A_i = 1 \mid X=x) < 1 \quad \forall x}$$

DIRECTED ACYCLIC GRAPHS (C8S1350)

We can use directed acyclic graphs to model (hypothesized) causal relationships.

eg

screen time ⟶ obesity

↳ this shows we hypothesize screen time causes obesity.

MEDIATOR (C8S1351)

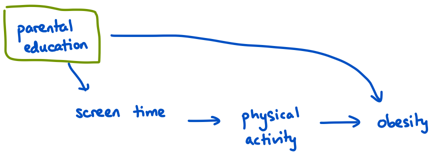
Consider

screen time → $\boxed{\text{physical activity}}$ → obesity

"physical activity" is a mediator.

CONFOUNDER (C8S1352)

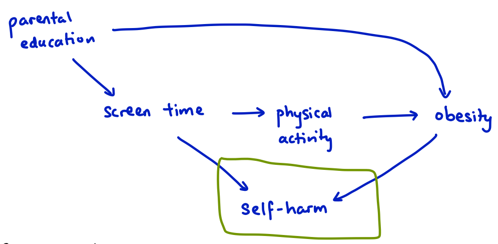
Consider



"Parental education" is a confounder.

COLLIDER (C8S1353)

Consider



"Self-harm" is a collider between screen time & obesity.

DIRECTED PATHS (C8S1356)

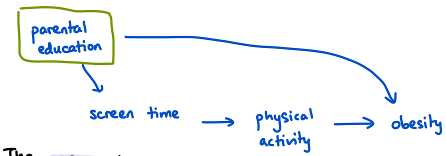
1. In a "directed path", all arrows point in the same direction.

eg screen time → physical activity → obesity

2. Any associations represent a causal relationship.

BACKDOOR PATHS (C8S1357)

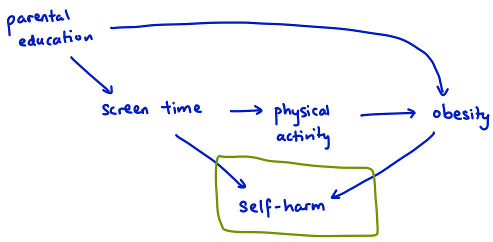
1. In a "backdoor path", two variables share the same cause.



2. The association by this path is non-causal.

CLOSED / BLOCKED PATHS (C6S1358)

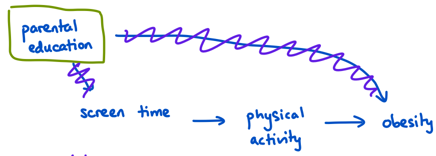
In a "closed path", two variables affect another variable.



CHANGING PATHS (C8S1359)

1. We can change the status of a path by controlling for or conditioning on a variable for our analysis.

eg

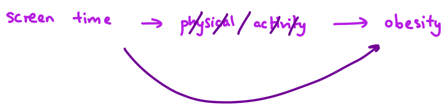


- parental education is a confounder for screen time & obesity
- "including" parental education in our analysis "closes" the backdoor path.

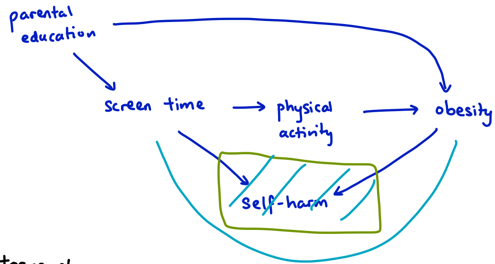
2. Note:

① Controlling for confounders is good;
- removes the confounder from analysis
- so estimate is more accurate.

② Controlling for mediators is bad.
- causes incorrect estimates of the overall association between the variables



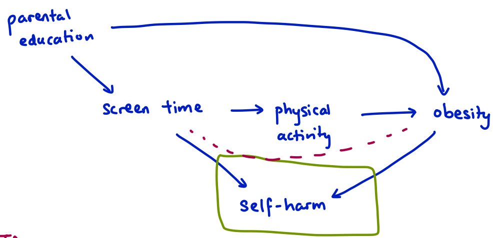
3. Conditioning on a collider changes the path from closed to open.



This transmits a non-causal association.

SELECTION BIAS (C8S1367)

1. Selection bias can be visualized in DAGs.
2. This often takes the form of colliders.



- If selected adolescents for the study because of a prior history of self-harm, we can view this as conditioning/controlling that collider.

LIMITATIONS (C8S1369)

Limitations of DAGs:

① They are qualitative; they do not indicate the form, size or direction of a relationship.

② They are limited by the information available to form them; it must be complete to give a proper causal interpretation.

③ They cannot depict random error; confounding can occur even when treatments are randomized, which is not shown in DAGs.