Chapter 1

Empirical Studies (Sections 1.1-1.2)

A **unit** is an individual person, place, or thing about which we can take some measurement(s)
A **population** is a collection of units
A **process** is also a collection of units, but those units are 'produced' over time
Population: All current UW undergraduate students.
Process: All UW undergraduate students for the next ten years.

Variates are characteristics of the units which are usually represented by letters such as x, y
Continuous variates are those that can be measured to an infinite degree of accuracy
Discrete variates are those that can only take a finite or countably infinite number of values
Categorical variates - units fall into a (non-numeric) category, such as hair colour
Ordinal variates - ordering is implied, but not necessarily through a numeric measure. Examples include 'strongly disagree, disagree, neutral, agree, strongly agree' in surveys.

An **attribute** of a population or process is a function of a variate which is defined for all units in the population or process. Examples: The modal number of completed assignments. The proportion of assignments submitted in the final 24 hours.

Sample survey - information is obtained about a finite population by selecting a representative sample from the population and determining the variates of interest for each unit in the sample
Observational study - information about a population or process is collected without any attempt to change one or more variates for the sampled units.
Experimental study - the experimenter intervenes and changes or sets the values of one or more variates for the units in the study.

Observational	Survey
Population of interest infinite or conceptual (persons at risk of a disease)	Finite, tangible (eligible voters in the next election)
Data collected routinely over time	Often only one point of contact with participants
More passive, learning about population's daily lives and/or habits	Specific questions about participants' lives and experiences

Numerical Summaries (Section 1.3)

Sample mean: for variable $y, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, which implies $\sum_{i=1}^{n} y_i = n\bar{y}$
Sample median: for continuous random variable $Y, P(Y \leq m) = \int_{-\infty}^{m} g(y)dy = 0.5$
for discrete odd number of observations, $\hat{m} = y_{(\frac{n+1}{2})}$, for discrete even, $\hat{m} = \frac{1}{2}[y_{(\frac{n}{2})} + y_{(\frac{n+1}{2})}]$
Mean: not robust (extreme cases affect value)
Median: robust resistance to outlying data

Ordered sample: denoted by $\{y_{(1)}, y_{(2)}, ..., y_{(n)}\}$ where $y_{(1)} \leq y_{(2)} \leq ... \leq y_{(n)}$
Sample mode: most useful for discrete or categorical data with small number of possible values. For frequency or grouped data, highest frequency is called **sample modal class**.

Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}(y_i^2) - n\bar{y}^2]$
Sample standard deviation: SQUARE ROOT of variance!

For Gaussian distribution $Y \sim G(\mu, \sigma)$,
$P(\mu - \sigma \leq Y \leq \mu + \sigma) \approx 0.68, P(\mu - 2\sigma \leq Y \leq \mu + 2\sigma) \approx 0.95$

Range: $y_{(n)} - y_{(1)}$ is a very crude measure of the spread of the data, very susceptible to outliers!

q(p) Quantile is the value such that a fraction p of the data fall at or below this value.
In R, this can be calculated by: `quantile(c(7,11,13,17,19), 0.25)`
where c denotes the dataset, calculates the 25th percentile and returns 11
$q(0.25)$ = lower or first quartile, $q(0.75)$ = upper or third quartile
IQR= $q(0.75) - q(0.25)$. Used as a measure of spread instead of the crude range. More robust.

Sample skewness =

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2]^{\frac{3}{2}}}$$

If data are symmetric (Gaussian, uniform) = sample skewness close to 0.
Long right tail = positive sample skewness (> 0)
Long left tail = negative sample skewness (< 0)
Since denominator always positive, $\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3 < 0$ implies negative skew, vise versa.

Sample kurtosis =

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

Data that look Gaussian have a sample kurtosis close to 3.
Data with large tails (more prone to extreme values) have a sample kurtosis larger than 3.
Data with shorter tails have a sample kurtosis less than 3.
Data that look uniform have a sample kurtosis close to 1.8.

To see if Gaussian model is appropriate for a particular sample:
> Is the sample mean close to the sample median?
> Are approximately 95% of points within two sample standard deviations of the sample mean?
> Is the sample skewness close to 0?
> Is the sample kurtosis close to 3?
Note: We never 'prove' an assumption, instead, we see if we can find evidence against an assumption. Never use definitive statements, use 'reasonably' and 'approximately'.

Five Number Summary
> The minimum $y_{(1)}$
> The lower quartile q(0.25)
> The median q(0.5)
> The upper quartile q(0.75)
> The maximum $y_{(n)}$

For numerical summaries for bivariate data (Sample correlation) and relative risk,
as well as graphical summary diagrams, refer to notes on next page.

Graphical Summaries (Section 1.3)

A **frequency histogram** has rectangles with height proportional to the number of observations (observed frequencies) for partitions divided in non-overlapping intervals.
Instead of observed frequencies, we can also plot the **relative frequencies** by scaling - bar of height 0.5 means half the data are contained in the interval.

In a **standard histogram**, the intervals are of equal width and the heights are equal to the frequencies. The sum of the areas of the bars will equal the sample size n.

In a **relative frequency** histogram, the area of the rectangle $= \frac{number\ of\ observations\ in\ interval}{total\ observations}$.
The sum of the areas of the rectangles equals 1.

Relative frequency histograms can superimpose a probability density function on top, since the Y axes match. **Empirical CDFs** are to CDFs as relative frequency histograms are to PDFs - a way to graphically compare CDFs to the actual shape of the data.

Definition of the empirical c.d.f: $\frac{number\ of\ values\ in\ \{y_1,...,y_n\}\ which\ are\ \leq y}{n}$

Box plots
Box height = IQR, Outliers beyond $q(0.25) - 1.5 \times IQR$ and $q(0.75) + 1.5 \times IQR$

```
  o   <- each outlier is plotted as a point
----- <- top whisker represents maximum value that isn't an outlier
  |
----- <- top of box represents third quartile
|   |
|---| <- line in the box represents median
|   |
----- <- bottom of box represents first quartile
  |
----- <- bottom whisker represents minimum value that isn't an outlier
```

We can use **scatter plots** for bivariate data. Simply plot the points $(x_i, y_i), i = 1, ..., n$
The variates are split into two types: **response variates** and **explanatory variates**.
The explanatory variate x is to partially explain or determine the distribution of y.

Data Analysis and Statistical Models (Sections 1.4-1.5, 2.1)

Descriptive statistics are portrayals of the data, or parts of the data, in numerical and graphical ways to show features of interest.
When the specific data obtained in the study of a population or process are used to draw general conclusions about the population or process itself using **inductive reasoning**, we call this process **statistical inference**.

Chapter 2

In a statistical model, a **random variable** (Like Y in $Y \sim G(\mu, \sigma)$) is used to represent a characteristic or variate of a randomly selected unit from the population or process.

Binomial(n, θ): for sequences of n independent success/fail trials each with θ probability of success, where Y is the number of successes.
Poisson(θ): for random occurences of an event over time averaging θ per unit time, where Y is the number of event occurences.
Exponential(θ): for random occurences of an event over time averaging θ per unit time, where Y is the time between event occurences.
Gaussian$(\theta), \theta = (\mu, \sigma)$ represent the distribution of continuous measurements, or when many statistical models are summed together, where Y is the value of the quantity.

PDF of random variable is $f(y;\theta) = \frac{d}{dy}P(Y \leq y)$ for $y \in range(Y)$
For discrete, $f(y;\theta) = P(Y = y)$, for continuous, $f(y;\theta) = P(a \leq Y \leq b) = \int_a^b f(y;\theta)dy$

Estimation (Sections 2.2, 2.3, 2.5)

A **point estimate** of a parameter θ is an estimate $\hat{\theta} = \hat{\theta}(y)$ that is the result of a function of observed data $(y_1, ..., y_n)$ and other known quantities such as the sample size n.
$G(\mu, \sigma) : \hat{\mu} = \bar{y}$, the sample mean. $Bin(n, \theta) : \hat{\theta} = \frac{y}{n}$, the sample proportion.

A **Likelihood function** for θ is $L(\theta) = L(\theta; y) = P(Y = y; \theta)$ for $\theta \in \Omega$
where Ω is the parameter space - the set of possible values of θ.
It is the probability of observing the data y as a function of the unknown constant θ, scaled by an arbitrary positive constant k.

The value of θ that maximizes $L(\theta)$ for given data y is called the **maximum likelihood estimate** (MLE) of θ and is denoted by $\hat{\theta}$. We do this by differentiating the likelihood function with respect to θ, then setting that derivative to 0 and solving for θ to find the global max.

The **relative likelihood function** is $R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}, \theta \in \Omega$ for $0 \leq R(\theta) \leq 1, R(\hat{\theta}) = 1$.
The y-axis of the relative likelihood function is the likelihood of the x-axis, not the probability of anything. Likewise, θ isn't a probability either, just an unknown constant. A value $R(\theta) = 0.6$ means θ is 0.6 times as likely as the maximum likelihood estimate.

Likelihoods are the frequency of certain hypotheses out of all hypotheses, while **probabilities** are the frequency of certain results out of all results. A **probability function** is when we have a parameter value and want to describe what samples will look like, while the **likelihood function** is when we have a sample and want to describe what the parameter value looked like.

Log likelihood function is defined as $\ell(\theta) = logL(\theta) = lnL(\theta), 0 < \theta < 1$. $\hat{\theta}$ maximises $\ell(\theta)$.

Suppose we have two independent data sets y, z for two independent random variables Y, Z that share a parameter θ. $L(\theta) = P(Y = y, Z = z; \theta) = P(Y = y; \theta)P(Z = z; \theta) = L(\theta; y)L(\theta; z)$.
If we observe data $Y = (Y_1, Y_2, ..., Y_n)$ that are independent and identically distributed each with probability function $P(Y_i = y_i; \theta)$ then $L(\theta) = \prod_{i=1}^n P(Y_i = y_i; \theta), \theta \in \Omega$.

The likelihood function for θ in a **binomial model** is $L(\theta; y) = \theta^y(1-\theta)^{n-y}$
The log likelihood function for binomial model is $\ell(\theta) = ylog\theta + (n-y)log(1-\theta), 0 < \theta < 1$.

For a **Poisson** function $L(\theta) = \theta^{n\bar{y}}e^{-n\theta}$, $\frac{d}{d\theta}L(\theta) = (\bar{y}-\theta)\theta^{n\bar{y}-1}ne^{-n\theta}$, $\ell(\theta) = n\bar{y}log(\theta) - n\theta, \hat{\theta} = \bar{y}$.

Suppose $Y = (Y_1, Y_2, ..., Y_n)$ is a random sample from a **continuous distribution**. We define the likelihood function for θ based on the observed data $y = (y_1, y_2, ..., y_n)$ as:
$L(\theta) = L(\theta; y) = \prod_{i=1}^{n} f(y_i; \theta), \theta \in \Omega$

For **Exponential**: PDF is $\frac{1}{\theta}e^{-y/\theta}$, $L(\theta) = \prod_{i=1}^{n}\frac{1}{\theta}e^{-y/\theta} = \frac{1}{\theta^n}e^{-\sum_{i=1}^{n} y_i/\theta} = \theta^{-n}e^{-n\bar{y}/\theta}$
$\ell(\theta) = log(\theta^{-n}e^{-n\bar{y}/\theta}) = -nlog(\theta) - \frac{n\bar{y}}{\theta} = -n(log(\theta) + \frac{\bar{y}}{\theta})$
Maximising: $\frac{d}{d\theta}\ell(\theta) = -n(\frac{1}{\theta} - \frac{\bar{y}}{\theta^2}) = \frac{n}{\theta^2}(\bar{y} - \theta)$, so $\bar{y} = \hat{\theta}$ as MLE.

For **Gaussian**: $\hat{\mu} = \bar{y}$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, like the sample variance but with n rather than n-1 for the denominator. For Gaussian calculation details and summary table, see next page(s).

Invariance property: if $\hat{\theta}$ is the MLE of θ then $g(\hat{\theta})$ is the MLE of $g(\theta)$. To find the maximum likelihood of a function, we just need to find the maximum likelihood of its parameter.
EX: the variance of the binomial distribution is $\sigma^2 = n\theta(1-\theta)$, so MLE of variance is $n\hat{\theta}(1-\hat{\theta})$

Check the Fit of a Model (Section 2.6)

To check the fit of model we can compare the observed frequencies based on the data with the expected frequencies calculated using probabilities based on the model we use.

For example, for a Poisson model we would first fit it to the dataset using the sample mean as θ. Then, we can calculate the theoretical frequency for an interval $P(y_1 \leq Y \leq Y_2)$ predicted by the Poisson model, and then compare them with the actual number of times we observe values between y_1 and y_2. We usually choose around 10-15 intervals such that each one contains at least one sample.

PPDAC (Chapter 3)

Problem: A clear statement of the study's objectives.
+ Units and the target population or target process must be defined. The **target population** is the collection of units to which the experimenters conducting the empirical study wish the conclusions to apply.
+ Also consider **variates** and **attributes** of interest. To determine the variates, look at what is measured or recorded on each unit.
+ Problems fall into three types: **descriptive** (what is the value of some attribute?), **causative** (does A cause B?), and **predictive** (what would be the effect of X?). Usually we cannot answer causative problems from observational studies or sample surveys.

The **study population** or study process is the collection of units available to be included (could be included) in the study. Often the study population is a strict subset of the target population, but not always (rats in medical trials).

! **Study error**: The difference of **attributes** in the **study population** from the attributes in the **target population**. A difference between the study and target population is not itself an example of study error. There must be a difference in attributes. It's not differences between the target population and the sample, or the study population and the sample.

Plan: The procedures used to carry out the study including how the data will be collected. For an empirical study the Plan should indicate the **target and study populations**, the **sampling protocol**, the **variates** which are to be measured, and the **quality** of the measurement systems that are intended for us.
+ The **sampling protocol** is the procedure used to select sample of units from study pop.
+ The number of units sampled is called the **sample size**.

The **sample** is only a subset of the units in the study population.

! **Sample error**: attributes in **sample** differ from the attributes in the **study population**. Different sampling protocols may lead to different sample errors. The values of the study population attributes are unknown, so the sample error is unknown. Statistical models are used to quantify the size of this error.
! **Measurement error**: when the measured value and the true value of a variate are not identical. Response bias is when study respondents systematically tend to give incorrect answers. For example, people tend to exaggerate their income on financial surveys. **Human error** like wrongly entering data can also be source of measurement error.

Departures from the Plan may arise **over time** (e.g. persons may drop out of a long term medical study) These departures will also impact the Analysis and Conclusion.

Data: The physical collection of the data, as described in the Plan. Variates must be clearly defined and satisfactory methods of measuring them must be used.

Analysis: The analysis of the data, accounting for considerations in the Problem and the Plan. Analysis should include numerical and graphical summaries of the data.
+ A key part is the selection of an appropriate model that describes the data and how the data were collected.
+ Checking the fit of the model must be included (like using QQ plots)

Conclusion: The conclusions that are drawn about the Problem, and limitations of the study.
+ Potential study, sample or measurement errors, as described in the Plan step, should be discussed and quantified if possible.
+ Departures from the Plan that affect the Analysis must be addressed.

Estimation (Sections 4.1-4.2)

When we observe data y from a random variable Y, we can say y is a realization of Y where $Y \sim G(\mu, \sigma)$, so $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is a realisation of $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim G(\mu, \frac{\sigma}{n})$.

A **point estimate** of θ is a function $\hat{\theta} = g(y_1, ..., y_n)$ of the observed data used to estimate the unknown parameter θ. For example, a point estimate for θ in a Poisson distribution is $\hat{\theta} = \bar{y}$, for binomial it's $\frac{\sum y_i}{n}$.

Let $Y_1, ..., Y_n$ be potential observations in a random sample. We can associate with the point estimate with a random variable $\tilde{\theta} = g(Y_1, ..., Y_n)$. So a **point estimator** is a random variable which is a function $\tilde{\theta} = g(Y_1, ..., Y_n)$ of the random variables $Y_1, ..., Y_n$.
An **estimator** is a rule that tells us how to process the data to obtain an **estimate** of an unknown parameter θ. The numerical value $\hat{\theta} = g(y_1, ..., y_n)$ is the value obtained using this rule for a particular observed dataset.
TLDR: $\hat{\theta}$ is an estimate that's an numerical value, $\tilde{\theta}$ is an estimator that's a random variable.

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

For $Y_i \sim G(\mu, \sigma)$, probability we draw a sample yielding an estimate $\hat{\mu}$ that is close to μ:
- Increases as n increases.
- Decreases as σ increases.
- Does not change with different population mean μ

Quick CLT: When we take a sample $Y_1, \ldots, Y_n$ from our population, each entry in the sample can be thought of as **independent and identically distributed**. Using this observation, we can approximate the mean of different distributions with:

- If $Y \sim \text{Binomial}(n, \theta)$ (Y of n samples are successes) and n is large, then

$$Y \sim \text{Gaussian}\left(n\theta, \sqrt{n\theta(1-\theta)}\right), \frac{Y}{n} \sim G\left(\theta, \sqrt{\frac{\theta(1-\theta)}{n}}\right)$$

- If $Y_i \sim \text{Poisson}(\theta)$, $1 \leq i \leq n$ (n samples, yielding Y_i events for each) and n is large, then

$$\overline{Y} \sim \text{Gaussian}\left(\theta, \sqrt{\frac{\theta}{n}}\right).$$

- If $Y_i \sim \text{Exponential}(\theta)$, $1 \leq i \leq n$ (n samples, taking Y_i time before an event for each) and n is large, then

$$\overline{Y} \sim \text{Gaussian}\left(\theta, \sqrt{\frac{\theta^2}{n}}\right).$$

Quick Reminder: If $Y \sim G(\mu, \sigma)$, **then** $\frac{Y-\mu}{\sigma} \sim G(0, 1)$.

If $Y \sim G\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ approximately, then it is $\text{sd}(Y)$ that affects the probability that our sample will result in an estimate close to μ.
- **A larger sample size n will decrease sd(Y)**, and more of our sample estimates will be close to the true value μ.
- **A smaller population standard deviation σ will decrease sd(Y)**, and more of our sample estimates will be close to the true value μ.
- The shape of our population distribution will affect how many of our sample estimates will be close to the true value μ, but predicting how is trickier.
- **The true mean μ does not directly affect how many of our sample estimates are close to the true μ, but it may do so indirectly through affecting sd(Y)**
(Like for poisson data)

Likelihood Intervals (Section 4.3)

Recall RLF, if for some number θ_0 we have $R(\theta_0) = 0.1$, we can say:
"The data are 10 times more likely if $\theta = \hat{\theta}$ than if $\theta = \theta_0$."
If $R(\theta_0) = 0.5$, then the data are half as likely if $\theta = \theta_0$ than if $\theta = \hat{\theta}$.

To indicate uncertainty in an estimate, we define an **interval estimate** of θ given observed data y to be $[L(y), U(y)]$, where $L(y)$ and $U(y)$ represent the lower and upper bounds.
For example, the interval estimate $\left[\overline{y} - \frac{2\sigma}{\sqrt{n}}, \overline{y} + \frac{2\sigma}{\sqrt{n}}\right]$ summarizes the uncertainty in the point estimate $\hat{\mu}$. As the sample size n increases, the confidence interval shrinks, converging on $\overline{y}$.

The $100p\%$ **likelihood interval** is defined as $\{\theta : R(\theta; y) \geq p\}$ where θ is the unknown parameter, $R(\theta)$ is the relative likelihood function, and $0 \leq p \leq 1$.

It's important to note that the likelihood interval is always a function of the observed data. The likelihood interval is an example of an interval estimate.
The wider the likelihood interval, the more uncertainty about our estimate, and the less we know about the value of the unknown parameter.
- Values in the 10% likelihood interval are generally considered plausible.
- Values inside the 50% likelihood interval are generally considered very plausible.
- Values outside the 10% likelihood interval are generally implausible, 1% very implausible.

The log relative likelihood function often looks pretty quadratic, while the relative likelihood function often resembles a bell curve. We sometimes write the $100p\%$ likelihood interval as $\{\theta : r(\theta; y) \geq \ln p\}$, where $r(\theta; y)$ is the log relative likelihood function, and the max is 0.

Confidence Intervals (Section 4.4)

To determine how good the **interval estimator** $[L(Y), U(Y)]$ is, we want to compute $P(L(Y) \leq \theta \leq U(Y))$.

The $100p\%$ confidence interval is the interval estimate $[L(y), U(y)]$ such that $P(L(y) \leq \theta \leq U(y)) = p$ where θ is the unknown parameter and $0 \leq p \leq 1$.
The center of the confidence interval is usually the point estimate for the mean, so the confidence interval is often just the point estimate plus or minus the **margin of error**. Also, p is called the **confidence coefficient**.

We can only say we are **100p% confident** that the interval contains the true but **unknown** value of θ. **We cannot say** $P(\theta \in [L(y), U(y)]) = p$**.** since θ is just a constant, so doesn't have a distribution or probability of being in there.
All we know is that if we take a very large number of samples, and calculated a confidence interval for each sample, then about 100p% of those confidence intervals would contain θ.

In general, we calculate the sample mean $\bar{y}$ and sample standard deviation σ, and then choose a value. The confidence interval is

$$\left[\bar{y} - z^* \cdot \frac{\sigma}{\sqrt{n}},\ \bar{y} + z^* \cdot \frac{\sigma}{\sqrt{n}}\right],$$

where z^* is the critical value corresponding to the desired confidence level.

- **Sample size** increases means width of the confidence interval decreases. Larger sample sizes provide more information about the population, reducing the error of the estimate.
- As the **confidence level** increases, the width of the confidence interval increases. We want to be more certain that the interval contains the true population parameter.
- As the **population variance** increases, the width of the confidence interval also increases. A higher variability in the data leads to less precise estimates of the population parameter.

A **pivotal quantity** $Q = Q(Y;\theta)$ is a function of the data Y and the unknown parameter θ such that the distribution of the random variable Q is completely known. This makes $P(a \le Q(Y;\theta) \le b)$ depend only on a and b but not on θ or any other unknown information.

In general, we can use a **pivotal quantity to construct a confidence interval** as follows:

1. Determine numbers a and b such that

$$P[a \le Q(Y;\theta) \le b] = p.$$

2. Re-express the inequality $a \le Q(Y;\theta) \le b$ in the form

$$L(Y) \le \theta \le U(Y),$$

then

$$p = P[a \le Q(Y;\theta) \le b] = P[L(Y) \le \theta \le U(Y)],$$

so the coverage probability equals p.

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

For $Y \sim \text{Gaussian}(\mu, \sigma)$ where σ is known, $P(-a \le Z \le a) = p \implies P(Z \le a) = \frac{1+p}{2}$.
(In R, $qnorm(\frac{1+p}{2})$) Then, the $100p\%$ confidence interval for samples is

$$\left[\bar{y} - a\frac{\sigma}{\sqrt{n}},\ \bar{y} + a\frac{\sigma}{\sqrt{n}}\right], \text{The width of our confidence interval is } 2a\frac{\sigma}{\sqrt{n}}$$

So our $100p\%$ CI for μ is of the form: point estimate $\pm$ (distribution quantile) $\times$ sd(estimator).
This result also holds for other distributions if the sample size is large enough, due to CLT.
Common confidence intervals we often use are 90%, 95%, and 99%, with corresponding values of a equal to 1.645, 1.960, and 2.576 respectively.

We can often find random variables $Q_n = Q_n(Y;\theta)$ such that as $n \to \infty$, the distribution of Q_n ceases to depend on θ or other unknown information. We call Q_n an **asymptotic** or **approximate pivotal quantity**.

1. Binomial Data:
For data from Binomial(n,θ) distribution, an approximate $100p\%$ confidence interval for θ is:

$$\hat{\theta} \pm a\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}},$$

where $\hat{\theta} = \frac{y}{n}$, and $P(Z \le a) = \frac{1+p}{2}$ with $Z \sim$ Gaussian$(0,1)$.
To ensure the confidence interval has a width $\le 2\ell$, choose the sample size n such that:

$$n \ge \left[\frac{a \cdot (0.5)}{\ell}\right]^2$$

2. Poisson Data:
For data from a Poisson(θ) distribution, an approximate $100p\%$ confidence interval for θ is:

$$y \pm a\frac{\sqrt{y}}{\sqrt{n}},$$

where $P(Z \le a) = \frac{1+p}{2}$, $Z \sim$ Gaussian$(0,1)$.
3. Exponential Data:
For data from Exponential(θ) distribution, an approximate $100p\%$ confidence interval for θ is:

$$y \pm a\frac{y}{\sqrt{n}},$$

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

Chi-squared and t Distributions (Section 4.5)

The **chi-squared distribution with k degrees of freedom**, denoted χ^2_k or $\chi^2(k)$, has the probability density function (p.d.f.) given by:

$$f(y;k) = \frac{1}{2^{k/2}\Gamma(k/2)} y^{k/2-1} e^{-y/2}, \quad y > 0.$$

The Gamma function, $\Gamma(z)$, is defined as: $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$.
If $W \sim \chi^2_k$, then the **expected value and variance** are given by: $E[W] = k, \mathrm{Var}(W) = 2k$.

The **R command** `pchisq(w, df)` will return the probability $P(W \le w)$ where $W \sim \chi^2_{\mathrm{df}}$, a chi-squared distribution with `df` degrees of freedom.
The command `qchisq(q, df)` will return a value w such that $P(W \le w) = q$

- If $W_1, W_2, \ldots, W_n$ are independent random variables with $W_i \sim \chi^2_{k_i}$, then

$$S = \sum_{i=1}^{n} W_i \sim \chi^2_{\sum k_i}.$$

- If $Z \sim G(0,1)$, then $Z^2 = W \sim \chi^2_1$, and so if $W \sim \chi^2_1$, then

$$P(W \le w) = 2P(Z \le \sqrt{w}) - 1,$$

where $Z \sim G(0,1)$.

- If $Z_1, Z_2, \ldots, Z_n \sim G(0,1)$ independently, then

$$S = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2.$$

- If $W \sim \chi_2^2$, then $W \sim$ Exponential(2), and thus

$$P(W \geq w) = e^{-w/2}.$$

- If $Z \sim G(0,1)$ and $U \sim \chi_k^2$ independently, then

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

- If T_k follows a t distribution with k degrees of freedom, as k increases, the probability density function (p.d.f.) of T_k approaches that of a $G(0,1)$ random variable.

Likelihood-Based Confidence Intervals (Section 4.6)

The **likelihood ratio statistic** $\Lambda(\theta)$ (a random variable that depends on data Y) is:

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

where $\tilde{\theta} = \tilde{\theta}(Y)$ is the maximum likelihood estimator of θ.
For large n, we can show that: $\Lambda(\theta) \sim \chi_1^2$ Therefore, $\Lambda(\theta)$ is an approximate pivotal quantity that can be used to obtain approximate confidence intervals for θ. Given **likelihood level p** the **confidence interval level q** is found by: $q = P(\Lambda(\theta) \leq -2\log p)$ where $\Lambda(\theta) \sim \chi_1^2$

Convert a 100q% Confidence Interval to a 100p% Likelihood Interval:
An approximate 100q% confidence interval is: $\{\theta : R(\theta) \geq e^{-c/2}\}$ where $P(W \leq c) = q, W \sim \chi_1^2$

1. Find c such that:

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

   where W follows a chi-squared distribution with 1 degree of freedom and Z follows a standard normal distribution. To find c, solve for the value of c that satisfies:

$$P(Z \leq \sqrt{c}) = \frac{1+q}{2}$$

2. Then, p is given by:

$$p = e^{-c/2}$$

Convert a 100p% Likelihood Interval to a 100q% Confidence Interval:
A 100p% likelihood interval) is:$\{\theta : R(\theta) \geq p\}$
Start with p, and calculate q as: $q = 2P(Z \leq \sqrt{-2\log p}) - 1$

Note that as p increases, likelihood interval gets narrower, while confidence interval gets wider.

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

1. 100p% confidence interval for the mean μ of a Gaussian distribution with **unknown standard deviation** σ:

$$\bar{y} \pm a \cdot \frac{s}{\sqrt{n}} \quad \text{where} \quad P(T \leq a) = \frac{1+p}{2}, \quad T \sim t_{n-1}, \quad s \text{ is the sample standard deviation.}$$

2. 100p% confidence interval for the **variance** σ^2 of a Gaussian distribution with **unknown mean** μ:

$$\left[\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right] \quad \text{where} \quad P(W \leq a) = \frac{1-p}{2}, \quad P(W \leq b) = \frac{1+p}{2}, \quad W \sim \chi^2_{n-1}.$$

3. For Gaussian experiment with known standard deviation σ, choose sample size n such that:

$$n \geq \left(\frac{a\sigma}{\ell}\right)^2 \quad \text{where} \quad P(Z \leq a) = \frac{1+p}{2}, \quad Z \sim N(0,1),$$

to ensure a 100p% confidence interval has width $\leq 2\ell$.

4. Differences between a confidence interval for μ where σ is and isn't known: the interval will be wider when σ is unknown.

Hypothesis Testing (Section 5.1)

Statistical tests work starting supposing the **null hypothesis** H_0 is true and then try to quantify the strength of evidence provided by our observed data. Design an experiment to see if our alternative hypothesis H_A is more plausible than the null hypothesis.

A **test statistic** or **discrepancy measure** is a function of the data defined as: $D = g(Y)$ constructed to measure the degree of 'agreement' between the observed data Y and the null hypothesis H_0. D is a function of Y, so it is a random variable. Once we observe $Y = y$, the **observed value** of D is denoted as: $d = g(y)$.

Typically, we define D such that:
- $d = 0$ represents the best possible agreement between the data and H_0.
- Larger values of d indicate poorer agreement between the data and H_0, meaning stronger evidence against H_0.
For **binomial**, discrepancy measure is $D = |Y - n\theta_0|$

The **p-value** of the test of hypothesis H_0 using test statistic D is $P(D \geq d; H_0)$. In other words, 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.

1. **Specify the Null Hypothesis**: Define H_0 to be tested using data Y.
2. **Choose a Test Statistic**: Select a test statistic or discrepancy measure $D(Y)$, where large values of D indicate results less consistent with H_0. Let $d = D(y)$ be the observed value of the test statistic from the data.
3. **Calculate the p-value**: p-value $= P(D \geq d \mid H_0 \text{ is true})$
   This represents the probability of obtaining a test statistic at least as extreme as d under the assumption that the null hypothesis H_0 is true.

4. **Draw a Conclusion**:

If p-value is small, either the null hypothesis is true and we only got what we did by chance, or the null hypothesis is false. A **small p-value** gives strong evidence against the null hypothesis, while a **large p-value** says that there's little evidence that the null hypothesis is false.

p-value	**Interpretation**
$p > 0.10$	No evidence against H_0 based on the observed data.
$0.05 < p \leq 0.10$	Weak evidence against H_0 based on the observed data.
$0.01 < p \leq 0.05$	Evidence against H_0 based on the observed data.
$0.001 < p \leq 0.01$	Strong evidence against H_0 based on the observed data.
$p \leq 0.001$	Very strong evidence against H_0 based on observed data.

Hypothesis Testing for Parameters in the $G(\mu, \sigma)$ Model (Section 5.2)

For testing the null hypothesis $H_0 : \mu = \mu_0$, we use the discrepancy measure: $D = |Y - \mu_0|$
The p-value for this test is given by: $P(|Y - \mu_0| \geq |y - \mu_0|)$
where $Y \sim G(\mu_0, \sigma/\sqrt{n})$.

Hypothesis Test for $H_0 : \mu = \mu_0$ (Unknown Standard Deviation σ)
Let $d = \frac{|y-\mu_0|}{s/\sqrt{n}}$ be the observed value of D for an experiment which has been conducted.

$$\text{p-value} = P\left(\frac{|Y - \mu_0|}{S/\sqrt{n}} \geq \frac{|y - \mu_0|}{s/\sqrt{n}}\right) = P(|T| \geq d) \text{ where } T \sim t_{n-1}$$

$$\text{p-value} = 2[1 - P(T \leq d)]$$

where $T \sim t_{n-1}$ and s is the sample standard deviation.

Hypothesis Test for $H_0 : \sigma^2 = \sigma_0^2$
Define

$$u = \frac{(n-1)s^2}{\sigma_0^2}$$

Let $U \sim \chi^2_{n-1}$.
If u is large (i.e., $P(U \leq u) > 0.5$), then: p-value $= 2P(U \geq u)$
If u is small (i.e., $P(U \leq u) < 0.5$), then: p-value $= 2P(U \leq u)$

Relationship Between Confidence Intervals and Hypothesis Tests

Suppose we test $H_0 : \mu = \mu_0$ for $G(\mu, \sigma)$ data, then

$$\text{p-value} \geq 0.05 \quad \Leftrightarrow \quad P\left(\frac{|Y - \mu_0|}{S/\sqrt{n}} \geq \frac{|y - \mu_0|}{s/\sqrt{n}}\right) \geq 0.05$$

$$\Leftrightarrow \quad P\left(|T| \geq \frac{|y - \mu_0|}{s/\sqrt{n}}\right) \geq 0.05 \quad \text{where} \quad T \sim t_{n-1}$$

$$\Leftrightarrow \quad P\left(|T| \leq \frac{|y - \mu_0|}{s/\sqrt{n}}\right) \leq 0.95$$

$$\Leftrightarrow \quad \frac{|y - \mu_0|}{s/\sqrt{n}} \leq a \quad \text{where} \quad P(|T| \leq a) = 0.95$$

This is equivalent to

$$-a \leq \frac{y-\mu_0}{s/\sqrt{n}} \leq a \quad \text{where} \quad P(|T| \leq a) = 0.95$$

which can be rearranged to

$$y - a\frac{s}{\sqrt{n}} \leq \mu_0 \leq y + a\frac{s}{\sqrt{n}}$$

Conclusion: The parameter value $\theta = \theta_0$ is inside a $100q\%$ confidence interval for θ if and only if the p-value for testing $H_0 : \theta = \theta_0$ is greater than $1 - q$.
Ex. If testing $H_0 : \theta = \theta_0$ results in a p-value ≤ 0.05, then a 95% confidence interval for θ will **not** contain θ_0, and vice-versa.

Note: This result only approximately holds if we use different pivotal quantities, such as a Gaussian approx. to calculate the p-value and a chi-squared approximation to calculate the CI.

Likelihood Ratio Tests (Section 5.3)

If H_0 is **true**, we know the (approximate) distribution of the discrepancy measure. For a general test of $H_0 : \theta = \theta_0$, we might consider

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

because if H_0 is true, then $\Lambda(\theta_0) \sim \chi^2_1$.

Process for using the likelihood ratio statistic for a test of $H_0 : \theta = \theta_0$
1. Propose a model for your data and form the likelihood function $L(\theta)$. Use this to derive an expression for $\hat{\theta}$.
2. Gather data and calculate $\hat{\theta}$ for your observed data.
3. Compute the observed value of the test statistic

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

4. The p-value is

$$\text{p-value} \approx P(W \geq \lambda(\theta_0)), \quad W \sim \chi^2_1$$

$$P(W \geq \lambda(\theta_0)) = 2\left[1 - P\left(Z \leq \sqrt{\lambda(\theta_0)}\right)\right], \quad W \sim \chi^2_1 \text{ and } Z \sim G(0,1)$$

Gaussian Response Models (Sections 6.1-6.3)

Univariate data - measuring one variate on each unit
Bivariate data - two variates are measured per unit
Residuals - vertical distances between a fitted line and the data

Estimates of α and β (denoted $\hat{\alpha}$ and $\hat{\beta}$) are called the **least squares estimates**
For a line $y = \alpha + \beta x$, we can define the residual for each pair of points (x_i, y_i) as $r_i = y_i - (\alpha + \beta x_i)$.
To find the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$, we minimize

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

by differentiating and finding $\hat{\alpha}, \hat{\beta}$ that solve these equations for 0. Recall that $\sum_{i=1}^{n} y_i = n\bar{y}$, and similar for x. We get:

$$\alpha = \bar{y} - \hat{\beta}\bar{x}, \quad \beta = \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

For sample correlation r $= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$:

$$r = \hat{\beta} \sqrt{\frac{S_{xx}}{S_{yy}}}, \quad \text{and} \quad \hat{\beta} = r \sqrt{\frac{S_{yy}}{S_{xx}}}.$$

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

- $\hat{\beta}$ and r have the same sign.
- If $S_{yy} \gg S_{xx}$, then $\hat{\beta}$ is (much) larger than r, and vice versa.

A Gaussian response model is one for which the distribution of the response variate Y, given the associated vector of covariates $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ for an individual unit, is of the form:

$$Y \sim G(\mu(\mathbf{x}), \sigma(\mathbf{x})).$$

Parameters in the $Y_i \sim G(\alpha + \beta x_i, \sigma)$ model: The line of best fit with the Gaussian model is known as the **simple linear regression model**. We're assuming that the variance is the same for each x_i, so the entire model has only the unknown parameters α, β, and σ.

- α: Mean value of the response variate in the study population of individuals for whom the explanatory variate is zero.
- β: Increase in the mean value of the response variate in the study population for a one unit increase in the value of the explanatory variate.
- σ: The variability in the response variate Y in the study population, and does not vary with x.

Since our model is $Y_i \sim G(\alpha + \beta x_i, \sigma)$ for $i = 1, 2, \ldots, n$ independently,
where the x_i $(i = 1, 2, \ldots, n)$ are known constants, then the **likelihood function** is:

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

To maximize $L(\alpha, \beta) = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right)$, we would minimize $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$, same as minimizing the sum of squared residual.

If $Y_i \sim G(\alpha + \beta x_i, \sigma)$ for $i = 1, 2, \ldots, n$, independently, where the x_i $(i = 1, 2, \ldots, n)$ are known constants, then the **least squares estimator of** β,

$$\tilde{\beta} = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})Y_i, \text{ has } \textbf{distribution } \tilde{\beta} \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right).$$

Pivotal quantity if σ is known:

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

If we know if $Y \sim G(\mu, \sigma)$, with σ unknown, we could use the pivotal quantity: $\frac{Y-\mu}{S/\sqrt{n}} \sim t_{n-1}$.
Since σ^2 is usually unknown, we estimate it using

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

The quantity $\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ is called the s**um of squared errors**, and s_e^2 is called the **mean squared error**. s_e^2 is not the maximum likelihood estimate of σ^2, but we use it to estimate σ^2 because if we define

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \tilde{\alpha} - \tilde{\beta}x_i)^2, \quad \tilde{\beta} = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})Y_i \quad \text{and} \quad \tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}, \text{ then } E[S_e^2] = \sigma^2.$$

In our linear model, if S_e^2 is the estimator of the mean squared error, then

$$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2_{n-2} \quad \text{and we can show that } \textbf{piviotal quantity} \text{ is } \frac{\tilde{\beta} - \beta}{S_e/\sqrt{S_{xx}}} \sim t_{n-2}.$$

$$\implies \text{for constructing confidence interval, } P\left(\tilde{\beta} - a\frac{S_e}{\sqrt{S_{xx}}} \leq \beta \leq \tilde{\beta} + a\frac{S_e}{\sqrt{S_{xx}}}\right),$$

then for observed data, a $100p\%$ **confidence interval** for β is

$$\left[\hat{\beta} - a\frac{S_e}{\sqrt{S_{xx}}}, \hat{\beta} + a\frac{S_e}{\sqrt{S_{xx}}}\right],$$

where $P(T \leq a) = \frac{1+p}{2}$ and $T \sim t_{n-2}$. This interval has width $2a\frac{S_e}{\sqrt{S_{xx}}}$.

The **width of the interval increases with** S_e, which is our estimate of σ from our $Y_i \sim G(\alpha + \beta x_i, \sigma)$ model. Greater variability in our response variates leads to greater uncertainty in our estimates.

The **width of the interval decreases with** $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, which is proportional to the sample variance of x. If we have larger variability in x, then our estimate of β is based on a more variable set of values of x. This reduces uncertainty in our estimate of β as it is based on a broader range of x values.

Since $\mu_x = \alpha + \beta x$, a test of $H_0 : \beta = 0$
is a test of the hypothesis that μ_x does not (linearly) depend on x.
The p-value for testing $H_0 : \beta = \beta_0$ is

$$\text{p-value} = 2\left[1 - P\left(T \leq \frac{|\hat{\beta} - \beta_0|}{S_e/\sqrt{S_{xx}}}\right)\right], \text{where } T \sim t_{n-2}$$

The point estimate for μ_x is $\hat{\mu}_x = \hat{\alpha} + \hat{\beta}x = \bar{y} + \hat{\beta}(x - \bar{x})$,
with corresponding estimator $\tilde{\mu}_x = \tilde{\alpha} + \tilde{\beta}x = \bar{Y} + \tilde{\beta}(x - \bar{x})$

Because $\tilde{\beta} = \frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x})Y_i$, we have:

$$\tilde{\mu}_x = \bar{Y} + \tilde{\beta}(x - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n} Y_i + (x - \bar{x}) \cdot \frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x})Y_i. = \sum_{i=1}^{n}\left(\frac{1}{n} + \frac{(x - \bar{x})(x_i - \bar{x})}{S_{xx}}\right) Y_i,$$

which is a linear combination of Gaussian random variables, $Y_i \sim G(\alpha + \beta x_i, \sigma)$. **Distribution**:

$$\tilde{\mu}_x \sim G\left(\mu_x, \sigma\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}\right), \text{ where } \tilde{\mu}_x = \tilde{\alpha} + \tilde{\beta}x \text{ and } \mu_x = \alpha + \beta x$$

Don't know σ, use the following **pivotal quantity**:

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

The **confidence interval** for μ_x is:

$$\mu_x \pm aS_e\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}, \text{ where } P(T \leq a) = \frac{1 + p}{2}, \text{ and } T \sim t_{n-2}.$$

The width of the confidence interval decreases as n increases because $\frac{1}{n}$ becomes smaller.
The width increases as $(x - \bar{x})^2$ increases because this term directly contributes to the variance.
The width decreases as S_{xx} increases because the variance is inversely proportional to S_{xx}.

Since $\mu(0) = \alpha + \beta(0) = \alpha$, a $100p\%$ confidence interval for α is given by:

$$\hat{\alpha} \pm aS_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}, \text{ where } P(T \leq a) = \frac{1 + p}{2}, \text{ and } T \sim t_{n-2}.$$

If $\bar{x}$ **is large in magnitude** (i.e., the x_i are typically far from 0), the term $\frac{\bar{x}^2}{S_{xx}}$ becomes large, which **widens the confidence interval** for α. This indicates higher uncertainty in the estimate of α when the explanatory variables are centered far from zero.

Confidence interval for an individual response
We can write the error in the point estimator of Y as:

$$Y - \tilde{\mu}_x = Y - \mu_x + \mu_x - \tilde{\mu}_x = R + [\mu_x - \tilde{\mu}_x].$$

R is independent of $\tilde{\mu}_x$ because it is not connected to the existing sample.

$$E[Y - \tilde{\mu}_x] = E(R + [\mu_x - \tilde{\mu}_x]) = E(R) + E[\mu_x] - E[\tilde{\mu}_x] = 0 + \mu_x - \mu_x = 0.$$

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

Distribution:

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

Pivotal quantity:

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

Interval estimate:

$$\hat{\alpha} + \hat{\beta}x \pm aS_e\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}, \text{ where } P(T \leq a) = \frac{1 + p}{2} \text{ and } T \sim t_{n-2}.$$

This interval is called a $100p\%$ **prediction interval** instead of a confidence interval, because here Y is not a parameter but a 'future' observation.

Checking model assumptions for Gaussian response models:

- Y_i (given covariates x_i) has a Gaussian distribution.
- That distribution has standard deviation σ, which does not depend on the covariates.
- $E[Y_i] = \mu(x_i)$ is a linear combination of known covariates $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ and the unknown regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$.

For scatterplots, consider the following:

- Do the points seem to fit reasonably along a straight line? (i.e., is $E[Y_i] = \mu(x_i)$ a linear function of x?)
- Are the points generally 'spread out' to the same extent regardless of x? (i.e., does σ depend on x?)

The residual $\hat{r}_i$ can be thought of as observed values of R_i in the model:$Y_i = \mu_i + R_i$, where $R_i \sim G(0, \sigma)$, $i = 1, 2, \ldots, n$ independently.

- Residual plots can make visualization easier: assessing whether the points lie along a horizontal line rather than an angled line.
- Residual plots are more general: they can be used when we have more than one covariate.
- Residual (and standardized residual) plots should look like points randomly scattered about a horizontal line at 0.

A **Q-Q plot** of the standardized residuals $\hat{r}_i^* = \frac{\hat{r}_i}{s_e} = \frac{y_i - \hat{\mu}_i}{s_e}$ should give approximately a **straight line** if the model assumptions hold.

Our observed model may not **extrapolate**. There are two key problems:

- Our model assumptions may no longer hold, and we have no way to check them.
- Our predictions may not make sense.

The R2 statistic:

If we write $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$, we can define the sum of squared errors: $\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
The smaller this is, the less 'error' in our model fit. This is also the sum of the squared residuals from fitting our least squares regression line.

$$R^2 = 1 - \frac{\text{SSE}}{S_{yy}} = \frac{S_{yy} - \text{SSE}}{S_{yy}}, \text{ where } S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

We can interpret the R^2 as

$$R^2 = \frac{\text{Variation explained by the regression model}}{\text{Total variation}}$$

R^2 takes values between 0 (the regression explains none of the variation in our response) and 1 (the regression perfectly explains all variation in our response).

Adding new explanatory variables will always appear to 'improve' our model by increasing R^2, even if the new variable is unrelated to our response. This can lead to a problem known as **overfitting**. To address this issue, we often consider the adjusted R^2:

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

Comparing the Means of Two Populations (Section 6.4)

For sample sizes n_1 and n_2, $Y_{1i} \sim G(\mu_1, \sigma)$ for $i = 1, 2, \ldots, n_1$ independently, $Y_{2i} \sim G(\mu_2, \sigma)$ for $i = 1, 2, \ldots, n_2$ independently. We call this a **two-sample** Gaussian problem.
Null hypothesis: $H_0 : \mu_1 = \mu_2$. Easier to think of this as: $H_0 : \mu_1 - \mu_2 = 0$.

$$\tilde{\mu}_1 = \bar{Y}_1 \sim G\left(\mu_1, \frac{\sigma}{\sqrt{n_1}}\right) \quad \text{and} \quad \tilde{\mu}_2 = \bar{Y}_2 \sim G\left(\mu_2, \frac{\sigma}{\sqrt{n_2}}\right)$$

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

Therefore, the **standardized statistic** is:

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

A point estimator of σ^2, called the **pooled estimator** of variance is:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2\right),$$

$\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$ independently, so **pivotal quantity** is $\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$.

A 100p% **CI** assuming equal variances is:

$$\bar{y}_1 - \bar{y}_2 \pm a s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, \text{where } P(T \le a) = \frac{1+p}{2}, \quad T \sim t_{n_1+n_2-2}.$$

A 100p% **CI** NOT assuming equal variances is:

$$\bar{y}_1 - \bar{y}_2 \pm a s_p \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}, \text{where } P(T \le a) = \frac{1+p}{2}, \quad T \sim t_{n_1+n_2-2}.$$

Testing **Unpaired** $H_0 : \mu_1 - \mu_2 = 0$, we use the test statistic with observed value: $d = \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$.

The p-value is then given by: $2\left[1 - P(T \le d)\right], \quad \text{where } T \sim t_{n_1+n_2-2}.$

Testing **Paired:** Define $Y_i = Y_{1i} - Y_{2i} \sim G(\mu, \sigma)$ and test $H_0 : \mu = 0$.

$$d = \frac{|\bar{y} - 0|}{s/\sqrt{n}}, \quad p\text{-value} = 2\left[1 - P(T \le d)\right], \quad T \sim t_{n-1}.$$

General Gaussian Response Models (Section 6.5)

The general Gaussian response model (also referred to as multiple linear regression) relates explanatory variates $x_{1i}, x_{2i}, \ldots, x_{ki}$ to a response Y_i via the model

$$Y_i \sim G(\mu(\mathbf{x}_i), \sigma)$$

with

$$E[Y_i] = \mu(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \quad \text{for } i = 1, 2, \ldots, n, \text{ independently.}$$

where $E[Y_i]$ is a linear function of a vector of k explanatory variates for unit i and the unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$. Then we seek parameters $\beta_0, \beta_1, \ldots, \beta_k$ that minimize:

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

We can test $H_0 : \beta_j = 0$ for each parameter using the test statistic:

$$t_j = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}} = \frac{\text{ESTIMATE}}{\text{STANDARD ERROR}}.$$

To test the null hypothesis $H_0 : \beta_j = 0$, then refer to the corresponding estimator: $T_j \sim t_{n-k-1}$, which is a t-distribution with $n-k-1$ degrees of freedom, as we are estimating $k+1$ parameters overall.

Multinomial Models (Chapter 7)

Recall: joint distribution of $Y_1, Y_2, \ldots, Y_k$ is multinomial:

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k) = \frac{n!}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k}$$

$$0 < \theta_j < 1, \quad \sum_{j=1}^{k} \theta_j = 1, \quad y_j = 0, 1, \ldots, \quad \sum_{j=1}^{k} y_j = n.$$

Likelihood Ratio Test:
In general, for multinomial data $y_1, y_2, \ldots, y_k$, the likelihood function is given by:

$$L(\theta_1, \theta_2, \ldots, \theta_k) = \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k}, \quad 0 < \theta_j < 1, \quad \sum_{j=1}^{k} \theta_j = 1.$$

The maximum likelihood estimate of θ_j is: $\hat{\theta}_j = \frac{y_j}{n}, \quad j = 1, 2, \ldots, k,$
while the maximum likelihood estimator of θ_j is: $\tilde{\theta}_j = \frac{Y_j}{n}, \quad j = 1, 2, \ldots, k.$

The likelihood ratio test statistic for testing $H_0 : \theta = \theta_0 = \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$ is:

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

and the likelihood function is:

$$L(\theta_1, \theta_2, \ldots, \theta_k) = \prod_{j=1}^{k} (\theta_j)^{y_j}.$$

Finally, the ratio is given by:

$$\frac{L(\theta_0)}{L(\tilde{\theta})} = \frac{\prod_{j=1}^{k} \left(\frac{1}{k}\right)^{Y_j}}{\prod_{j=1}^{k} \left(\frac{Y_j}{n}\right)^{Y_j}} = \prod_{j=1}^{k} \left(\frac{n/k}{Y_j} \right)^{Y_j}. = \prod_{j=1}^{k} \left(\frac{E_j}{Y_j} \right)^{Y_j}$$

LR statistic for dataset with observed value:

$$\Lambda(\theta_0) = -2 \log \left[\prod_{j=1}^{k} \left(\frac{E_j}{Y_j} \right)^{Y_j} \right] = 2 \sum_{j=1}^{k} Y_j \log \left(\frac{Y_j}{E_j} \right) \sim \chi^2_{k-1-p}, \quad \lambda(\theta_0) = 2 \sum_{j=1}^{k} y_j \log \left(\frac{y_j}{e_j} \right)$$

If $y_j = e_j$, then category j will not affect the statistic.
If $y_j > e_j$, then category j will increase the statistic.
If $y_j < e_j$, then category j will decrease the statistic.
Our **categories are not independent**: if we have $y_j > e_j$ in one category, we must have $y_j < e_j$ in another category!

Larger values of $\lambda(\theta_0)$ provide stronger evidence against H_0, so for test of H_0, **p-value:**

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

Pearson Test:
For large n,

$$D = \sum_{j=1}^{k} \frac{(Y_j - E_j)^2}{E_j} \sim \chi^2_{k-1-p}, \quad d = \sum_{j=1}^{k} \frac{(y_j - e_j)^2}{e_j}$$

and so the **p-value** would be

$$p\text{-value} = P(D \geq d), \quad D \sim \chi^2_{k-1-p}.$$

Where y_j is the observed count in category j, and e_j is the expected count in category j assuming the null hypothesis is true.

Two-way tables (Section 7.3)
(Also see addition notes)
We have a model:

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

And a **null hypothesis**:

$$H_0 : \theta_{11} = \alpha\beta$$

Likelihood functions:

$$L(\tilde{\theta}) = \tilde{\theta}_{11}^{Y_{11}} \tilde{\theta}_{12}^{Y_{12}} \tilde{\theta}_{21}^{Y_{21}} \tilde{\theta}_{22}^{Y_{22}} = \left(\frac{Y_{11}}{n}\right)^{Y_{11}} \left(\frac{Y_{12}}{n}\right)^{Y_{12}} \left(\frac{Y_{21}}{n}\right)^{Y_{21}} \left(\frac{Y_{22}}{n}\right)^{Y_{22}}$$

$$L(\alpha, \beta) = (\alpha\beta)^{Y_{11}} [\alpha(1-\beta)]^{Y_{12}} [(1-\alpha)\beta]^{Y_{21}} [(1-\alpha)(1-\beta)]^{Y_{22}} = \alpha^{Y_{11}+Y_{12}} (1-\alpha)^{Y_{21}+Y_{22}} \beta^{Y_{11}+Y_{21}} (1-\beta)^{Y_{12}+Y_{22}}$$

$$0 < \alpha < 1 \quad \text{and} \quad 0 < \beta < 1$$

Likelihood ratio statistic:

$$\Lambda = -2 \log \left(\frac{L(\tilde{\alpha}, \tilde{\beta})}{L(\tilde{\theta})} \right) = -2 \log \left(\frac{\tilde{\alpha}^{Y_{11}+Y_{12}} (1-\tilde{\alpha})^{Y_{21}+Y_{22}} \tilde{\beta}^{Y_{11}+Y_{21}} (1-\tilde{\beta})^{Y_{12}+Y_{22}}}{\tilde{\theta}_{11}^{Y_{11}} \tilde{\theta}_{12}^{Y_{12}} \tilde{\theta}_{21}^{Y_{21}} \tilde{\theta}_{22}^{Y_{22}}} \right)$$

$$\text{where} \quad \tilde{\alpha} = \frac{Y_{11} + Y_{12}}{n}, \quad \tilde{\beta} = \frac{Y_{11} + Y_{21}}{n}, \quad \tilde{\theta}_{ij} = \frac{Y_{ij}}{n}$$

are the corresponding maximum likelihood estimators.

Expected values: $E_{11} = n\tilde{\alpha}\tilde{\beta}$ with observed value $e_{11} = n\hat{\alpha}\hat{\beta}$, **in general** $e_{ij} = \frac{r_i c_j}{n}$

Once we have our expected counts, we can compute our likelihood ratio test statistic

$$\lambda = 2 \left[y_{11} \log \left(\frac{y_{11}}{e_{11}} \right) + y_{12} \log \left(\frac{y_{12}}{e_{12}} \right) + y_{21} \log \left(\frac{y_{21}}{e_{21}} \right) + y_{22} \log \left(\frac{y_{22}}{e_{22}} \right) \right]$$

where to get our p-value we know we should consult a chi-squared distribution with $k - 1 - p$ degrees of freedom, where

- k is the number of categories.
- p is the number of parameters estimated in forming the null hypothesis.

For **generalised larger tables**, we have $(Y_{11}, Y_{12}, \ldots, Y_{ab}) \sim \text{Multinomial}(n; \theta_{11}, \theta_{12}, \ldots, \theta_{ab})$ where θ_{ij} is the probability that a randomly selected unit is in category A_i and B_j. That is,

$$\theta_{ij} = P(A_i \cap B_j), \quad i = 1, \ldots, a;\ j = 1, \ldots, b.$$

Our **null hypothesis** is that A_i and B_j are independent, so we test

$$H_0 : \theta_{ij} = P(A_i \cap B_j) = P(A_i)P(B_j).$$

Let

$$\alpha_i = P(\text{unit is type } A_i) \quad \text{and} \quad \beta_j = P(\text{unit is type } B_j).$$

To test whether A and B are independent variates, **we test**

$$H_0 : \theta_{ij} = \alpha_i \beta_j, \quad i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b.$$

$$\hat{\alpha}_i = \frac{r_i}{n}, \quad \hat{\beta}_j = \frac{c_j}{n}$$

and the **expected frequencies** e_{ij} are:

$$e_{ij} = \frac{r_i c_j}{n}, \quad i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b.$$

Finally to get p-value, we use:

$$\Lambda = 2 \sum_{i=1}^{a} \sum_{j=1}^{b} Y_{ij} \log \left(\frac{Y_{ij}}{E_{ij}} \right), \quad k = a \times b, \quad p = (a-1) + (b-1),$$

since we only actually estimate $a - 1$ parameters for A and $b - 1$ parameters for B. So under the hypothesis of independence, for sufficiently large n,

$$\Lambda \sim \chi^2_{(a-1)(b-1)}, \quad \text{approximately.}$$

Additional Notes:

- We usually try to keep all expected counts ≥ 5
- Always check your expected counts sum to n

Cause and Effect (Chapter 8)

x has a **causal effect** on Y if, when all other factors that affect Y are held constant (ideal, impractical since we cannot hold all other factors that affect y constant), a change in x induces a change in a **property of the distribution** of Y.

Defining causality: We look for a c**hange in the distribution** of a response variate, not simply a change in its observed value.

Reasons why two variates may be associated, but not necessarily causally related:

1. The **explanatory variate is the direct cause of the response variate** - even if one variate is the direct cause of another, we may not see a strong association
2. The **response variate is causing a change in the explanatory variate** - sometimes the causal connection is the opposite of what we might expect, 'reverse causality'
3. The explanatory variate is a **contributing but not sole cause** of the response variate - easy to be misled into thinking you have found a sole cause for a particular outcome, when you have actually found a necessary contributor to the outcome.
4. Both variates are **changing with time** - Nonsensical associations often result from correlating two variates that are both changing over time
5. The association may be due to **coincidence** or random chance.
6. **Both variates may result from a common cause** - An association between two variates may be observed because both variates are responding to changes in some unobserved variate or variates.
   These variates are sometimes referred to as **confounding** or lurking variates.

Given the number of possible explanations for association between two variates, how do we **establish a causal connection**?
Find groups of people who are - on average - the same, give members of each group a different treatment, then compare the response measures in each group.
Use **Randomization** - help account for the problem of confounding variates.

In **observational studies** controlling variates and using **randomization is not possible**.
We need:

1. The association between the two variates must be observed in many studies of different types among different groups.
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 that a causal link does not depend on the observed association alone.
4. There must be a consistent response, that is, one variate always increases (decreases) as the other variate increases.