Keven Qiu

Contents

1 Introduction to Statistical Sciences

1.1 Empirical studies and Statistical Sciences

1.2 Data Collection

Def 1: A variate is a characteristic of a unit.

Def 2: An attribute of a population or process is a function of the variates over the population or process.

1.3 Data Summaries

Measures of location:

- Sample mean $\overline{y} = \frac{1}{n}\sum\limits_{i=1}^{n} y_i$.
- Sample median $\hat{m}$ or the middle value when n is odd and average of two middle values when n is even.
- Sample mode or the value of y which appears with the highest frequency.

Measures of dispersion or variability:

- Sample variance:

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

- Range $= y_{(n)} - y_{(1)}$ where $y_{(n)} = \max(y_1, y_2, \ldots, y_n)$ and $y_{(1)} = \min(y_1, y_2, \ldots, y_n)$.
- Interquartile Range.

Measures of shape:

- Sample skewness is a measure of the (lack of) symmetry in the data.

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

- Sample kurtosis measures the heaviness of the tails and peakedness of the data relative to Normal data. If kurtosis is greater than 3 then this indicates heavier tails and a more peaked center than Normal data.

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

Def 3: Let $\{y_{(1)}, y_{(2)}, \ldots, y_{(2)}\}$ where $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ be the order statistic for the data set $\{y_1, y_2, \ldots, y_n\}$. For $0 \leq p \leq 1$, the pth (sample) quantile (also called the $100p$th (sample) percentile), is a value, call it $q(p)$, determined as follows:

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

Def 4: The quantiles $q(0.25)$, $q(0.5)$ and $q(0.75)$ are called the lower or first quartile, the median, and the upper or third quartile.

Def 5: The interquartile range is $IQR = q(0.75) - q(0.25)$.

Def 6: The five number summary of a data set consists of the smallest observation, the lower quartile, the median, the upper quartile and the largest value, that is the five values: $y_{(1)}, q(0.25), q(0.5), q(0.75), y_{(n)}$.

Def 7: The sample correlation, denoted by r, for data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is

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

where

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

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

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

Def 8: For categorical data in the form of a General two-way table, the relative risk of event A in group B as compared to group $\overline{B}$ is

$$\text{relative risk} = \frac{\frac{y_{11}}{(y_{11}+y_{12})}}{\frac{y_{21}}{(y_{21}+y_{22})}}$$

Graphical Summaries:

- Frequency histograms (Standard and Relative Frequency). In a relative frequency histogram, the bar's areas add up to 1.
- Bar graphs
- Empirical cumulative distribution function
- Boxplots. The minimum and maximum whisker of the boxplots is equal to $q(0.25) - 1.5(IQR)$ and $q(0.75) + 1.5(IQR)$ respectively. The outliers are represented with some symbol.

- Scatterplots
- Run charts

Def 9: For a data set $\{y_1, y_2, \ldots, y_n\}$, the empirical cumulative distribution function or e.c.d.f is defined by

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

1.4 Probability Distributions and Statistical Models

Data summaries and properties of probability models:

- The sample mean $\overline{y}$ corresponds to the population mean $E(Y) = \mu$.
- The sample standard deviation s corresponds to σ, the population SD of Y, where $\sigma^2 = E[(Y - \mu)^2]$.
- The sample median $\hat{y}$ corresponds to the population median m.
- The relative frequency histogram corresponds to the probability histogram of Y for discrete distributions and the probability density function of Y for continuous distributions.

1.5 Data Analysis and Statistical Inference

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

Statistical Inference: Using the data obtained in the study of a process or population to draw general conclusions about the process or population itself.

2 Statistical Models and Maximum Likelihood Estimation

2.1 Choosing a Statistical Model

Binomial Distribution: $Y \sim \text{Binomial}(n, \theta)$

$$P(Y = y; \theta) = f(y; \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \text{ for } y = 0, 1, \ldots, n$$

$$E(Y) = n\theta,\ Var(Y) = n\theta(1 - \theta)$$

Poisson Distribution: $Y \sim \text{Poisson}(\theta)$

$$f(y; \theta) = \frac{\theta^y e^{-\theta}}{y!}$$

$$E(Y) = \theta, Var(Y) = \theta$$

Exponential Distribution: $Y \sim \text{Exponential}(\theta)$

$$f(y;\theta) = \frac{1}{\theta}e^{-y/\theta}$$

$$E(Y) = \theta, Var(Y) = \theta^2$$

Gaussian/Normal Distribution: $Y \sim G(\mu, \sigma)$ or $Y \sim N(\mu, \sigma^2)$

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

$$E(Y) = \mu, Var(Y) = \sigma^2$$

Multinomial Distribution: $(Y_1, Y_2, \ldots, Y_k) \sim \text{Multinomial}(n; \boldsymbol{\theta})$

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k; \boldsymbol{\theta}) &= f(y_1, y_2, \ldots, y_k; \boldsymbol{\theta}) \\ &= \frac{n!}{y_1! y_2! \cdots y_k!}\theta_1^{y_1}\theta_2^{y_2}\cdots\theta_k^{y_k} \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$

2.2 Maximum Likelihood Estimation

Def 10: A point estimate of a parameter is the value of a function of the observed data $y_1, y_2, \ldots, y_n$ and other known quantities such as the sample size n. We use $\hat{\theta}$ to denote an estimate of the parameter θ.

Statistic: A function of the data which does not involve any unknown quantities such as unknown parameters.

Def 11: Let the discrete (vector) random variable $\mathbf{Y}$ represent potential data that will be. used to estimate θ, and let $\mathbf{y}$ represent the actual observed data that are obtained in a specific application. The likelihood function for θ is defined as

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

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

Def 12: The value of θ which maximizes $L(\theta)$ for given data $\mathbf{y}$ is called the maximum likelihood estimate (m.l. estimate) of θ. It is the value of θ which maximizes the probability of observing the data $\mathbf{y}$. This value is denoted $\hat{\theta}$.

Def 13: The relative likelihood function is defined as

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

Note that $0 \leq R(\theta) \leq 1$ for all $\theta \in \Omega$.

Def 14: The log likelihood function is defined as

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

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

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

Likelihood function for Binomial distribution:

$$\begin{aligned} L(\theta) &= P(y \text{ units have characteristic}; \theta) \\ &= \binom{n}{y} \theta^y (1-\theta)^{n-y} && \text{for } 0 \le \theta \le 1 \\ &= \theta^y (1-\theta)^{n-y} && \text{for } 0 \le \theta \le 1 \end{aligned}$$

(We can drop the constant $\binom{n}{y}$)
If $y \neq 0$ and $y \neq n$ then it can be shown that $L(\theta)$ attains its maximum value at $\theta = \hat{\theta} = \frac{y}{n}$ by solving $\frac{dL(\theta)}{d\theta} = 0$. The estimate $\hat{\theta} = y/n$ is called the sample proportion.

Likelihood function for Poisson distribution: Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from a Poisson(θ) distribution. The likelihood function is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(y_i;\theta) = \prod_{i=1}^{n} P(Y_i = y_i;\theta) && \text{for } \theta \in \Omega \\ &= \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} = \left(\prod_{i=1}^{n} \frac{1}{y_i!} \right) \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta} && \text{for } \theta \ge 0 \end{aligned}$$

or more simply

$$L(\theta) = \theta^{n\overline{y}} e^{-n\theta}$$

The log likelihood is

$$l(\theta) = n(\overline{y} \ln \theta - \theta)$$

with derivative $\frac{d}{d\theta} l(\theta) = n\left(\frac{\overline{y}}{\theta} - 1\right) = \frac{n}{\theta}(\overline{y} - \theta)$. The maximum likelihood estimate of θ is $\hat{\theta} = \overline{y}$.

Combining likelihoods based on independent experiments: If we have two data sets $\mathbf{y}_1$ and $\mathbf{y}_2$ from two independent studies for estimating θ, then since the corresponding random variables $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent we have

$$P(\mathbf{Y}_1 = \mathbf{y}_1, \mathbf{Y}_2 = \mathbf{y}_2;\theta) = P(\mathbf{Y}_1 = \mathbf{y}_1;\theta) P(\mathbf{Y}_2 = \mathbf{y}_2;\theta)$$

and we obtain the "combined" likelihood function $L(\theta)$ based on $\mathbf{y}_1$ and $\mathbf{y}_2$ together as

$$L(\theta) = L_1(\theta) \times L_2(\theta) \text{ for } \theta \in \Omega$$

where $L_j(\theta) = P(\mathbf{Y}_j = \mathbf{y}_j;\theta), j = 1, 2$

2.3 Likelihood Functions for Continuous Distributions

Def 15: If $y_1, y_2, \ldots, y_n$ are the observed values of a random sample from a distribution with probability density function $f(y;\theta)$, then the likelihood function is defined as

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

Likelihood function for Exponential distribution:

$$L(\theta) = \theta^{-n} e^{-\frac{n\overline{y}}{\theta}}$$

The log likelihood function is

$$l(\theta) = -n\left(\ln\theta + \frac{\overline{y}}{\theta}\right)$$

with derivative

$$\begin{aligned}\frac{d}{d\theta} l(\theta) &= -n\left(\frac{1}{\theta} - \frac{\overline{y}}{\theta^2}\right)\\ &= \frac{n}{\theta^2}(\overline{y} - \theta)\end{aligned}$$

Maximum likelihood estimate of θ is $\hat{\theta} = \overline{y}$.

Likelihood function for Gaussian distribution: The likelihood function for $\boldsymbol{\theta} = (\mu, \sigma)$ is

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

or more simply

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

The log likelihood is

$$l(\boldsymbol{\theta}) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \overline{y})^2 - \frac{n(\overline{y}-\mu)^2}{2\sigma^2}$$

Maximum likelihood estimate of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma})$, where

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y} \text{ and } \hat{\sigma} = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right]^{\frac{1}{2}}$$

2.4 Likelihood Functions for Multinomial Models

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

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

The log likelihood function is

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

Maximum likelihood estimates of $\theta_1, \theta_2, \ldots, \theta_k$ is

$$\hat{\theta}_i = \frac{y_i}{n}$$

Distribution	Observed Data	Maximum Likelihood Estimate	Maximum Likelihood Estimator	Relative Likelihood Function
Binomial(n, θ)	y	$\hat{\theta} = \frac{y}{n}$	$\tilde{\theta} = \frac{Y}{n}$	$R(\theta) = \left(\frac{\theta}{\hat{\theta}}\right)^y \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n-y}$ $0 < \theta < 1$
Poisson(θ)	$y_1, y_2, \ldots, y_n$	$\hat{\theta} = \bar{y}$	$\tilde{\theta} = \overline{Y}$	$R(\theta) = \left(\frac{\theta}{\hat{\theta}}\right)^{n\hat{\theta}} e^{n(\hat{\theta}-\theta)}$ $\theta > 0$
Geometric(θ)	$y_1, y_2, \ldots, y_n$	$\hat{\theta} = \frac{1}{1+\bar{y}}$	$\tilde{\theta} = \frac{1}{1+\overline{Y}}$	$R(\theta) = \left(\frac{\theta}{\hat{\theta}}\right)^n \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n\bar{y}}$ $0 < \theta < 1$
Negative Binomial(k, θ)	$y_1, y_2, \ldots, y_n$	$\hat{\theta} = \frac{k}{k+\bar{y}}$	$\tilde{\theta} = \frac{k}{k+\overline{Y}}$	$R(\theta) = \left(\frac{\theta}{\hat{\theta}}\right)^{nk} \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n\bar{y}}$ $0 < \theta < 1$
Exponential(θ)	$y_1, y_2, \ldots, y_n$	$\hat{\theta} = \bar{y}$	$\tilde{\theta} = \overline{Y}$	$R(\theta) = \left(\frac{\hat{\theta}}{\theta}\right)^n e^{n\left(\frac{1-\hat{\theta}}{\theta}\right)}$ $\theta > 0$

2.5 Invariance Property of Maximum Likelihood Estimate

Theorem 16: If $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ is the maximum likelihood estimate of $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ then $g(\hat{\boldsymbol{\theta}})$ is the maximum likelihood estimate of $g(\boldsymbol{\theta})$.

2.6 Checking the Model

Comparing Observed and Expected Frequencies: Compare the observed frequencies with the expected frequencies calculated using the assumed model. This method is useful for data from a discrete probability model.

Poisson model expected frequency:

$$e_j = n\frac{\hat{\theta}^j e^{-\hat{\theta}}}{j!}$$

Exponential model expected frequency in $[a_{j-1}, a_j]$:

$$e_j = n\int\limits_{a_{j-1}}^{a_j} \frac{1}{\hat{\theta}} e^{-y/\hat{\theta}}\, dy = n\left(e^{-\frac{a_{j-1}}{\hat{\theta}}} - e^{-\frac{a_j}{\hat{\theta}}}\right)$$

Graphical Checks of Models: Useful for continuous data.

Empirical cumulative distribution functions and cumulative distribution functions: A second graphical method is to plot the empirical cdf $\hat{F}(y)$ and then superimpose on this a plot of the cdf $P(Y \leq y; \theta) = F(y; \theta)$.

Qqplots for checking Gaussian model:
To check if $G(\mu, \sigma)$ matches the data set $\{y_1, \ldots, y_n\}$, order the data as $\{y_{(1)}, \ldots, y_{(n)}\}$. Let $Q(p)$ be the pth (theoretical) quantile for the $G(\mu, \sigma)$ distribution, that is $P(Y \leq Q(p)) = p$ where $Y \sim G(\mu, \sigma)$. Also $q(p)$ is the pth sample quantile. Then plot $\left(Q\left(\frac{i}{n+1}\right), q\left(\frac{i}{n+1}\right)\right)$ where the points should lie in a reasonably straight line.

3 Planning and Conducting Empirical Studies

3.1 Empirical Studies

Empirical study: A study which is carried out to learn about a population or process by collecting data. It is helpful to think about planning and conducting a study using a set of steps (PPDAC) as the following:

- Problem: a clear statement of the study's objectives, usually involving one or more questions.
- Plan: the procedures used to carry out the study including how the data will be collected.
- Data: the physical collection of the data, as described in the Plan.

- Analysis: the analysis of the data collected in light of the Problem and the Plan.
- Conclusion: the conclusions that are drawn about the Problem and their limitations.

3.2 The Steps of PPDAC

Problem: The problem step describes what the experimenters are trying to learn or what questions they want to answer. Often this can be done using questions starting with "What".

- What conclusions are the experimenters trying to draw?
- What group of things or people do the experimenters want the conclusions to apply?
- What variates can be defined?
- What is(are) the question(s) the experimenters are trying to answer?

There are three common types of statistical problems that are encountered.

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

Def 17: The target population or target process is the collection of units to which the experimenters conducting the empirical study wish the conclusions to apply.

We want the problem specified in terms of attributes of the target population/process. This includes the mean, proportion, variability, and also graphical attributes such as population histogram, population cdf, or a scatterplot.

Plan: The plan step depends on the questions posed in the problem step. The plan step includes a description of the population or process of units from which units will be selected, what variates will be collected, and how the variates will be measured.

Def 18: The study population or study process is the collection of units available to be included in the study.

Def 19: If the attributes in the study population/process differ from the attributes in the target population/process then the difference is called study error.

Def 20: The sampling protocol is the procedure used to select a sample of units from the study population/process. The number of units sampled is called the sample size.

Def 21: If the attributes in the sample differ from the attributes in the study population/process the difference is called sample error.

Def 22: If the measured value and the true value of a variate are not identical the difference is called measurement error.

Response bias: When those that do respond have a somewhat different characteristic than the population at large, the quality of the data is threatened, especially when the response rate is lower.

Data The goal of the data step is to collect the data according to the plan. Any deviations from the plan should be noted. The data must be stored in a way that facilitates the analysis.

Analysis The analysis step includes both simple and complex calculations to process the data into information. Numerical and graphical methods and other methods are used in this step to summarize the data. A key component of the analysis step is the selection of an appropriate model that describes the data and how the data were collected.

Conclusions The purpose of the conclusion step is to address the questions posed in the problem. An attempt should be made to quantify or discuss potential errors as described in the plan step. Limitations are discussed as well.

4 Estimation

4.1 Statistical Models and Estimation

In statistical estimation we use two models:

(1) A model which describes the variability in the variate(s) of interest in the population or process being studied.

(2) A model which takes in to account how the data were collected and which is constructed in conjunction with the model in (1).

4.2 Estimators and Sampling Distributions

Def 23: A (point) estimator $\tilde{\theta}$ is a random variable which is a function $\tilde{\theta} = g(Y_1, Y_2, \ldots, Y_n)$ of the random variables $Y_1, Y_2, \ldots, Y_n$. The distribution of $\tilde{\theta}$ is called the sampling distribution of the estimator.

For a sample of size n drawn without replacement from a **finite population** of size N,

$$sd(\overline{Y}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

4.3 Interval Estimation Using the Likelihood Function

Def 24: Relative Likelihood function: see Def 13

Def 25: A $100p\%$ likelihood interval for θ is the set $\{\theta : R(\theta) \geq p\}$.

Notes:

(1) Usually likelihood intervals cannot be found explicitly. They may be determined more accurately by solving the equation $R(\theta) - p = 0$ using the uniroot function R.

(2) A likelihood interval is an interval of the form $[L(\mathbf{y}), U(\mathbf{y})]$ where $L(\mathbf{y})$ and $U(\mathbf{y})$ are functions of the observed data $\mathbf{y}$. $L(\mathbf{y})$ and $U(\mathbf{y})$ are the two solutions of the equation $R(\theta) - p = 0$ with $L(\mathbf{y}) \le U(\mathbf{y})$. Since $R(\theta) = R(\theta; \mathbf{y})$ depends on the data $\mathbf{y}$, the solutions of L and U will also depend on $\mathbf{y}$.

Guidelines for Interpreting Likelihood Intervals:

- Values of θ inside a 50% likelihood interval are very plausible in light of the observed data.
- Values of θ inside a 10% likelihood interval are plausible in light of the observed data.
- Values of θ outside a 10% likelihood interval are implausible in light of the observed data.
- Values of θ outside a 1% likelihood interval are very implausible in light of the observed data.

Def 26: The log relative likelihood function is

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

for $\theta \in \Omega$ where $l(\theta) = \log L(\theta)$ is the log likelihood function.
$r(\theta)$ can also be used to obtain a $100p\%$ likelihood interval since $R(\theta) \ge p$ if and only if $r(\theta) \ge \log p$.

4.4 Confidence Intervals and Pivotal Quantities

Def 27: Suppose the interval estimator $[L(\mathbf{Y}), U(\mathbf{Y})]$ has the property that

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

Suppose the interval estimate $[L(\mathbf{y}), U(\mathbf{y})]$ is constructed for the parameter θ based on observed data $\mathbf{y}$. The interval estimate $[L(\mathbf{y}), U(\mathbf{y})]$ is called a $100p\%$ confidence interval for θ and p is called the confidence coefficient.
Note: $P[L(\mathbf{Y}) \le \theta \le U(\mathbf{Y})]$ is called the coverage probability of the interval estimator $[L(\mathbf{Y}), U(\mathbf{Y})]$.

Def 28: A pivotal quantity $Q = Q(\mathbf{Y}; \theta)$ is a function of the data $\mathbf{Y}$ and the unknown parameter θ such that the distribution of the random variable Q is fully known. That is, probability statements such as $P(Q \le b)$ and $P(Q \ge a)$ depend on a and b but not on θ or any other unknown information.

Confidence Interval for mean μ of a Gaussian distribution with known sd σ:

$$Q = Q(\mathbf{Y}; \mu) = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim G(0, 1)$$

Since the above and $G(0, 1)$ is a completely known distribution, Q is a pivotal quantity. A two-sided confidence interval takes the form:

$$\text{point estimate } \pm a \times \text{ standard deviation of the estimator}$$

$$\text{Estimate} \pm \text{margin of error}$$

$$\text{Estimate} \pm z^* SE$$

where a is a quantile from the $G(0, 1)$ distribution.

1. $\overline{y} \pm 1.96\frac{\sigma}{\sqrt{n}}$ is a 95% confidence interval for μ.
2. $\overline{y} \pm 1.6449\frac{\sigma}{\sqrt{n}}$ is a 90% confidence interval for μ.
3. $\overline{y} \pm 2.5758\frac{\sigma}{\sqrt{n}}$ is a 99% confidence interval for μ.

Asymptotic Gaussian Pivotal Quantities: Suppose $\tilde{G}$ is a point estimator of the unknown parameter θ. Suppose also that the Central Limit Theorem can be used to obtain the result that

$$\frac{\tilde{\theta} - \theta}{g(\tilde{\theta})/\sqrt{n}}$$

has approx. a $G(0, 1)$ distribution for large n where $E(\tilde{\theta}) = \theta$ and $sd(\tilde{\theta}) = g(\theta)/\sqrt{n}$ for some real valued function $g(\theta)$. If we replace θ by $\tilde{\theta}$ in the denominator then it can be shown that

$$Q_n(\tilde{\theta}; \theta) = \frac{\tilde{\theta} - \theta}{g(\tilde{\theta})/\sqrt{n}}$$

also has approx. a $G(0, 1)$ distribution for large n.

Approximate confidence interval for Binomial model Suppose $Y \sim \text{Binomial}(n, \theta)$. The maximum likelihood estimator of θ is $\tilde{\theta} = Y/n$ with

$$E(\tilde{\theta}) = E\left(\frac{Y}{n}\right) = \theta$$

and

$$sd(\tilde{\theta}) = sd\left(\frac{Y}{n}\right) = \sqrt{\frac{\theta(1-\theta)}{n}}$$

By CLT the random variable

$$\frac{\tilde{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}}$$

has approx. a $G(0, 1)$ distribution for large n. $g(\theta) = \sqrt{\theta(1-\theta)}$. Therefore, replacing the denominator with $\tilde{\theta} = Y/n$, we have the random variable

$$Q_n = Q_n(Y; \theta) = \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}}$$

Thus,

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

is an approximately 95% confidence interval for θ where $\hat{\theta} = y/n$ and y is observed data.

4.5 The Chi-squared and t Distributions

The χ^2 Distribution

To define the Chi-squared distribution we need the Gamma function and its properties:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy \text{ for } \alpha > 0$$

Properties of the Gamma function:

1. $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
2. $\Gamma(\alpha) = (\alpha - 1)!$ for $\alpha = 1, 2, \ldots$
3. $\Gamma(1/2) = \sqrt{\pi}$

The $\chi^2(k)$ distribution, $W = Z_1^2 + Z_2^2 + \cdots + Z_k^2$ where $Z_i \sim G(0,1)$, is a continuous family of distributions on $(0, \infty)$ with probability density function:

$$f(x;k) = \boxed{\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

where $k \in \{1, 2, \ldots\}$ is a parameter of the distribution. We write $X \sim \chi^2(k)$. The parameter k is referred to as the "degrees of freedom" (d.f.) parameter.

For $k = 2$, the p.d.f. is the Exponential(2) p.d.f.
For $k > 2$, the p.d.f. is unimodal with maximum value at $x = k - 2$.
For values $k \geq 30$, the p.d.f. resembles that of a $N(k, 2k)$ p.d.f.

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

$$E(X) = k \text{ and } Var(X) = 2k$$

$$E(X^j) = 2^j \frac{\Gamma\left(\frac{k}{2} + j\right)}{\Gamma\left(\frac{k}{2}\right)} \text{ for } j = 1, 2, \ldots$$

Theorem 29: Let $W_1, W_2, \ldots, W_n$ be independent random variables with $W_i \sim \chi^2(k_i)$. Then

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

Theorem 30: If $Z \sim G(0,1)$, then the distribution of $W = Z^2$ is $\chi^2(1)$.

Corollary 31: If $Z_1, Z_2, \ldots, Z_n$ are mutually independent $G(0,1)$ random variables and $S = \sum\limits_{i=1}^n Z_i^2$, then $S \sim \chi^2(n)$.

Useful Results:

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

Student's t distribution

Student's t distribution or t distribution has p.d.f.

$$f(t;k) = c_k \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}} \quad \text{for } t \in \mathbb{R} \text{ and } k = 1, 2, \ldots$$

where the constant c_k is

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

The parameter k is called the degrees of freedom. We write $T \sim t(k)$.

Theorem 32: Suppose $Z \sim G(0,1)$ and $U \sim \chi^2(k)$ independently. Let

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

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

4.6 Likelihood-Based Confidence Intervals

Likelihood Ratio Statistic:

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

Theorem 33: If $L(\theta)$ is based on $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, a random sample of size n, and if θ is the true value of the scalar parameter, then (under mild mathematical conditions) the distribution of $\Lambda(\theta)$ converges to a $\chi^2(1)$ distribution as $n \to \infty$. This means that $\Lambda(\theta)$ can be used as a pivotal quantity for sufficiently large n in order to obtain approximate confidence intervals for θ.

Theorem 34: A $100p\%$ likelihood interval is an approximate $100q\%$ confidence interval where $q = P(W \leq -2\log p) = 2P(Z \leq \sqrt{-2\log p}) - 1$ and $W \sim \chi^2(1)$, $Z \sim N(0,1)$.
A $100p\%$ likelihood interval is defined by $\{\theta; R(\theta) \geq p\}$ which can be rewritten as

$$\{\theta; R(\theta) \geq p\} = \left\{\theta : -2\log\left[\frac{L(\theta)}{L(\hat{\theta})}\right] \leq -2\log p\right\}$$

Theorem 35: If a is a value such that $p = 2P(Z \leq a) - 1$ where $Z \sim N(0,1)$, then the likelihood interval $\left\{\theta : R(\theta) \geq e^{-a^2/2}\right\}$ is an approximate $100p\%$ confidence interval.

Approximate confidence intervals for Binomial model:

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

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

Note: $E(S^2) = \sigma^2$.

Theorem 36: Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from the $G(\mu, \sigma)$ distribution with sample mean $\overline{Y}$ and sample variance S^2. Then

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

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

Confidence intervals for μ: The confidence interval for μ when σ is **unknown** is:

$$\overline{y} \pm a\frac{s}{\sqrt{n}} = [\overline{y} - as/\sqrt{n}, \overline{y} + as/\sqrt{n}]$$

Behaviours of confidence interval as $n \to \infty$: As sample size n increases, $E(S) \approx \sigma$, the sample sd s gets closer to the true sd σ. Secondly as the degrees of freedom $k = n-1$ increase, the quantiles of the t distribution approach the quantiles of the $G(0,1)$ distribution.
In general for large n, the width of the confidence interval gets narrower as n increases at the rate of $1/\sqrt{n}$.

Theorem 37: Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from the $G(\mu, \sigma)$ distribution with sample variance S^2.

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

Confidence intervals for σ^2 **and** σ: A $100p\%$ confidence interval $P(a \leq U \leq b) = p$ for σ^2 is

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

and $100p\%$ confidence interval for σ is

$$\left[s\sqrt{\frac{(n-1)}{b}}, s\sqrt{\frac{(n-1)}{a}}\right]$$

Note the swapping of a and b.
For convenience, a and b are chosen such that

$$P(U \leq a) = P(U > b) = \frac{1-p}{2}$$

This is because the $\chi^2(n-1)$ distribution is not symmetric.
Note that using the table to find a and b such that

$$P(U \leq a) = \frac{1-p}{2} \text{ and } P(U \leq b) = p + \frac{1-p}{2} = \frac{1+p}{2}$$

Note that unlike confidence intervals for μ, the confidence interval for σ^2 is not symmetric about s^2.

In some cases we are interested in an upper bound on σ. In this case we take $b = \infty$ and find a such that $P(a \leq U) = p$ or $P(U \leq a) = 1 - p$ so that a one-sided $100p\%$ confidence interval for σ is

$$\left[0, s\sqrt{\frac{n-1}{a}}\right]$$

Prediction Interval for a Future Observation: Since $Y - \overline{Y}$ is a linear combination of independent Gaussian random variables then $Y - \overline{Y}$ also has a Gaussian distribution with mean

$$E(Y - \overline{Y}) = \mu - \mu = 0$$

and variance

$$Var(Y - \overline{Y}) = Var(Y) + Var(\overline{Y}) = \sigma^2 + \frac{\sigma^2}{n}$$

Since

$$\frac{Y - \overline{Y}}{\sigma\sqrt{1 + \frac{1}{n}}} \sim G(0, 1)$$

independently of

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

then by Theorem 32

$$\frac{\frac{Y-\overline{Y}}{\sigma\sqrt{1+\frac{1}{n}}}}{\sqrt{S^2/\sigma^2}} = \frac{Y - \overline{Y}}{S\sqrt{1 + \frac{1}{n}}} \sim t(n-1)$$

is a pivotal quantity which can be used to obtain an interval of values for Y. Let a be the value such that

$$P(-a \leq T \leq a) = p \text{ or } P(T \leq a) = \frac{1+p}{2} \text{ where } T \sim t(n-1)$$

Therefore,

$$\left[\overline{y} - as\sqrt{1 + \frac{1}{n}}, \overline{y} + as\sqrt{1 + \frac{1}{n}}\right]$$

is an interval of values for the future observation Y with confidence coefficient p.
The interval is called a $100p\%$ prediction interval instead of a confidence interval since Y is not a parameter but a random variable.

4.8 Chapter 4 Summary

Table 4.3
Approximate Confidence Intervals for Named Distributions
based on Asymptotic Gaussian Pivotal Quantities

Named Distribution	Observed Data	Point Estimate $\hat{\theta}$	Point Estimator $\tilde{\theta}$	Asymptotic Gaussian Pivotal Quantity	Approximate $100p\%$ Confidence Interval
Binomial(n, θ)	y	$\frac{y}{n}$	$\frac{Y}{n}$	$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}}$	$\hat{\theta} \pm a\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$
Poisson(θ)	$y_1, y_2, \ldots, y_n$	$\bar{y}$	$\overline{Y}$	$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}}{n}}}$	$\hat{\theta} \pm a\sqrt{\frac{\hat{\theta}}{n}}$
Exponential(θ)	$y_1, y_2, \ldots, y_n$	$\bar{y}$	$\overline{Y}$	$\frac{\tilde{\theta}-\theta}{\frac{\tilde{\theta}}{\sqrt{n}}}$	$\hat{\theta} \pm a\frac{\hat{\theta}}{\sqrt{n}}$

Note: The value a is given by $P(Z \leq a) = \frac{1+p}{2}$ where $Z \sim G(0,1)$. In R, $a = \texttt{qnorm}\left(\frac{1+p}{2}\right)$

Table 4.4
Confidence/Prediction Intervals for Gaussian
and Exponential Models

Model	Unknown Quantity	Pivotal Quantity	$100p\%$ Confidence/Prediction Interval
$G(\mu,\sigma)$ σ known	μ	$\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim G(0,1)$	$\bar{y} \pm a\sigma/\sqrt{n}$
$G(\mu,\sigma)$ σ unknown	μ	$\frac{\overline{Y}-\mu}{S/\sqrt{n}} \sim t(n-1)$	$\bar{y} \pm bs/\sqrt{n}$
$G(\mu,\sigma)$ μ unknown σ unknown	Y	$\frac{Y-\overline{Y}}{S\sqrt{1+\frac{1}{n}}} \sim t(n-1)$	$100p\%$ Prediction Interval $\bar{y} \pm bs\sqrt{1+\frac{1}{n}}$
$G(\mu,\sigma)$ μ unknown	σ^2	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left[\frac{(n-1)s^2}{d}, \frac{(n-1)s^2}{c}\right]$
$G(\mu,\sigma)$ μ unknown	σ	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left[\sqrt{\frac{(n-1)s^2}{d}}, \sqrt{\frac{(n-1)s^2}{c}}\right]$
Exponential(θ)	θ	$\frac{2n\overline{Y}}{\theta} \sim \chi^2(2n)$	$\left[\frac{2n\bar{y}}{d_1}, \frac{2n\bar{y}}{c_1}\right]$

Notes: (1) The value a is given by $P(Z \le a) = \frac{1+p}{2}$ where $Z \sim G(0,1)$.
In R, $a = \texttt{qnorm}\left(\frac{1+p}{2}\right)$

(2) The value b is given by $P(T \le b) = \frac{1+p}{2}$ where $T \sim t(n-1)$. In R, $b = \texttt{qt}\left(\frac{1+p}{2}, n-1\right)$

(3) The values c and d are given by $P(W \le c) = \frac{1-p}{2} = P(W > d)$ where $W \sim \chi^2(n-1)$.
In R, $c = \texttt{qchisq}\left(\frac{1-p}{2}, n-1\right)$ and $d = \texttt{qchisq}\left(\frac{1+p}{2}, n-1\right)$

(4) The values c_1 and d_1 are given by $P(W \le c_1) = \frac{1-p}{2} = P(W > d_1)$ where $W \sim \chi^2(2n)$.
In R, $c_1 = \texttt{qchisq}\left(\frac{1-p}{2}, 2n\right)$ and $d_1 = \texttt{qchisq}\left(\frac{1+p}{2}, 2n\right)$

5 Hypothesis Testing

5.1 Introduction

Def 38: A test statistic or discrepancy measure D is a function of the data $\mathbf{Y}$ that is constructed to measure the degree of "agreement" between the data $\mathbf{Y}$ and the null hypothesis H_0.
Usually we define D so that $D = 0$ represents the best possible agreement between the data and H_0, and values of D not close to 0 indicate poor agreement.

We want to determine $P(D \geq d; H_0)$ where the notation "$; H_0$" means assuming H_0 is true.

Two types of hypotheses:

(1) The hypothesis $H_0 : \theta = \theta_0$ where it is assumed that the data $\mathbf{Y}$ have arisen from a family of distributions with probably (density) function $f(\mathbf{y}; \theta)$ with parameter θ.

(2) The hypothesis $H_0 : Y \sim f_0(y)$ where it is assumed that the data $\mathbf{Y}$ have a specified probability (density) function $f_0(y)$.

Two types of errors:

- Type I Error: Reject the null hypothesis when it was true.
- Type II Error: Fail to reject the null hypothesis when it was false.

Statistical test of hypothesis: First, assume that the hypothesis H_0 will be tested using some random data $\mathbf{Y}$. We then adopt a test statistic or discrepancy measure $D(\mathbf{Y})$ for which, normally, large values of D are less consistent with H_0. Let $d = D(\mathbf{y})$ be the corresponding observed value of D. We then calculate the $p - value$ or observed significance level of the test. There are one-sided and two-sided hypothesis tests.

Def 39: Suppose we use the test statistic $D = D(\mathbf{Y})$ to test the hypothesis H_0. Suppose also that $d = D(\mathbf{y})$ is the observed value of D. The $p - value$ or observed significance level of the test of hypothesis H_0 using test statistic D is

$$p - value = P(D \geq d; H_0)$$

$p - value$	Interpretation
$p - value > 0.1$	No evidence against H_0 based on the observed data.
$0.05 < p - value \leq 0.1$	Weak evidence against H_0 based on the observed data.
$0.01 < p - value \leq 0.05$	Evidence against H_0 based on the observed data.
$0.001 < p - value \leq 0.01$	Strong evidence against H_0 based on the observed data.
$p - value \leq 0.001$	Very strong evidence against H_0 based on the observed data.

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

Test of Hypothesis for μ

Suppose we wish to test the hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_A : \mu \neq \mu_0$. The test statistic is

$$D = \frac{\left|\overline{Y} - \mu_0\right|}{S/\sqrt{n}} \sim t(n-1)$$

Let

$$d = \frac{|\overline{y} - \mu_0|}{s/\sqrt{n}}$$

be the observed value of D in a sample with mean $\overline{y}$ and s.d. s, then

$$\begin{aligned} p - value &= P(D \geq d; H_0 \text{ is true}) \\ &= P(|T| \geq d) \qquad \text{where } T \sim t(n-1) \\ &= 2[1 - P(T \leq d)] \end{aligned}$$

One-sided test of hypothesis for μ

The null hypothesis $H_0 : \mu = \mu_0$ and the alternative hypothesis $H_A : \mu > \mu_0$. The test statistic is

$$D = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}}$$

Let the observed value of D be

$$d = \frac{\overline{y} - \mu_0}{s/\sqrt{n}}$$

Then

$$\begin{aligned} p - value &= P(D \geq d; H_0 \text{ is true}) \\ &= P(T \geq d) \\ &= 1 - P(T \leq d) \qquad \text{where } T \sim t(n-1) \end{aligned}$$

Relationship between Hypothesis Testing and Interval Estimation:
Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from the $G(\mu, \sigma)$ distribution. Suppose we test $H_0 : \mu = \mu_0$.

$$p - value \geq 0.05$$

$$\text{if and only if } P\left(\frac{\left|\overline{Y} - \mu_0\right|}{S/\sqrt{n}} \geq \frac{|\overline{y} - \mu_0|}{s/\sqrt{n}}; H_0 : \mu = \mu_0 \text{ is true}\right) \geq 0.05$$

$$\text{if and only if } P\left(|T| \geq \frac{|\overline{y} - \mu_0|}{s/\sqrt{n}}\right) \geq 0.05 \text{ where } T \sim t(n-1)$$

$$\text{if and only if } P\left(|T| \leq \frac{|\overline{y} - \mu_0|}{s/\sqrt{n}}\right) \leq 0.95$$

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

$$\text{if and only if } \mu_0 \in [\overline{y} - as/\sqrt{n}, \overline{y} + as/\sqrt{n}]$$

The $p-value$ for testing $H_0 : \mu = \mu_0$ is greater than or equal to 0.05 if and only if the value $\mu = \mu_0$ is an element of a 95% confidence interval for μ. Note both endpoints of the interval correspond to a $p - value$ equal to 0.05 while values inside the interval will have $p - values$ greater than 0.05.

More generally, suppose we have data $\mathbf{y}$ and a model $f(\mathbf{y}; \theta)$. Suppose we use the same pivotal quantity to construct the confidence interval for θ and to test the hypothesis $H_0 : \theta = \theta_0$. Then the parameter value $\theta = \theta_0$ is an element of the $100q\%$ confidence interval for θ if and only if the $p - value$ for testing $H_0 : \theta = \theta_0$ is greater than or equal to $1 - q$.

Test of Hypothesis for σ

The null hypothesis $H_0 : \sigma = \sigma_0$ or equivalently $H_0 : \sigma^2 = \sigma_0^2$. We use the test statistic

$$U = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

Large and small values of U provide evidence against H_0. U has a Chi-squared distribution when H_0 is true. The following approximates the $p - value$ because χ^2 is asymmetric:

1. Let $u = (n-1)s^2/\sigma_0^2$ denote the observed value of U from the data.
2. If u is large (that is, if $P(U \leq u) > \frac{1}{2}$) compute the $p - value$ as

   $$p - value = 2P(U \geq u)$$

   where $U \sim \chi^2(n-1)$.
3. If u is small (that is, if $P(U \leq u) < \frac{1}{2}$) compute the $p - value$ as

   $$p - value = 2P(U \leq u)$$

   where $U \sim \chi^2(n-1)$.

Note: only one of the two values $2P(U \geq u)$ and $2P(U \leq u)$ will be less than 1 and this is the desired $p - value$.

5.3 Likelihood Ratio Test of Hypothesis - One Parameter

When a pivotal quantity does not exist then a general method for finding a test statistic with good properties can be based on the likelihood function.

Assume that the null hypothesis is θ_0 and alternative is θ_1. Then using the ratio of the likelihood values:

$$\frac{L(\theta_0)}{L(\theta_1)}$$

If the value of the ratio is much greater than 1 then the data support the value θ_0 more than θ_1.

If there is no alternative hypothesis, then it is natural to replace θ_1 by the most plausible value, i.e., the maximum likelihood estimate $\hat{\theta}$. The ratio is just the relative likelihood function at θ_0:

$$R(\theta_0) = \frac{L(\theta_0)}{L(\hat{\theta})}$$

If $R(\theta_0)$ is close to one, then θ_0 is plausible in light of the observed data, but if $R(\theta_0)$ is very small and close to zero, then θ_0 is not plausible in light of the observed data and suggests evidence against H_0. Therefore, the random variable $L(\theta_0)/L(\tilde{\theta})$ is a natural statistic for testing $H_0 : \theta = \theta_0$.
It is actually easier to use the likelihood ratio statistic:

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

which is a one-to-one function of $L(\theta_0)/L(\tilde{\theta})$. We choose this because if H_0 is true, then $\Lambda(\theta_0)$ has approximately $\chi^2(1)$ distribution. Large observed values of $\Lambda(\theta_0)$ indicate evidence against H_0.

To determine the $p - value$ we first calculate the observed value of $\Lambda(\theta_0)$:

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

The approximate $p - value$ is then

$$\begin{aligned} p - value &\approx P[W \geq \lambda(\theta_0)] && \text{where } W \sim \chi^2(1) \\ &= P\left(|Z| \geq \sqrt{\lambda(\theta_0)} \right) && \text{where } Z \sim G(0,1) \\ &= 2 \left[1 - P\left(Z \leq \sqrt{\lambda(\theta_0)} \right) \right] \end{aligned}$$

Summary: We can test $H_0 : \theta = \theta_0$ using our test statistic the likelihood ratio test statistic Λ. Large observed values of $\Lambda(\theta_0)$ correspond to evidence rejecting the null hypothesis H_0. If H_0 is true, $\Lambda(\theta_0)$ has approximately a $\chi^2(1)$ distribution.
If $\hat{\theta}$ is close in value to θ_0 then $R(\theta_0)$ will be close in value to 1 and $\lambda(\theta_0)$ will be close in value to 0.

Likelihood ratio test statistic for Binomial model: The relative likelihood function for the Binomial model is

$$R(\theta) = \left(\frac{\theta}{\hat{\theta}} \right)^y \left(\frac{1-\theta}{1-\hat{\theta}} \right)^{n-y}$$

for $0 \leq \theta \leq 1$. The likelihood ratio test statistic for testing H_0 is

$$\Lambda(\theta_0) = -2 \log \left[\left(\frac{\theta_0}{\tilde{\theta}} \right)^y \left(\frac{1-\theta_0}{1-\tilde{\theta}} \right)^{n-y} \right]$$

where $\tilde{\theta} = Y/n$ is the maximum likelihood estimator of θ. The observed value of $\Lambda(\theta_0)$ is

$$\lambda(\theta_0) = -2 \log R(\theta_0) = -2 \log \left[\left(\frac{\theta_0}{\hat{\theta}} \right)^y \left(\frac{1-\theta_0}{1-\hat{\theta}} \right)^{n-y} \right]$$

where $\hat{\theta} = y/n$.

Likelihood ratio test statistic for Exponential model: Suppose $y_1, \ldots, y_n$ is an observed random sample from $Exponential(\theta)$ distribution.

$$L(\theta) = \theta^{-n} e^{-n\overline{y}/\theta}$$

Since MLE is $\hat{\theta} = \overline{y}$,

$$R(\theta) = \left(\frac{\hat{\theta}}{\theta}\right)^n e^{n(1-\hat{\theta}/\theta)}$$

The likelihood ratio test statistic is

$$\Lambda(\theta_0) = -2\log\left[\left(\frac{\tilde{\theta}}{\theta_0}\right)^n e^{n(1-\tilde{\theta}/\theta_0)}\right]$$

where $\tilde{\theta} = \overline{Y}$ and the observed value of $\Lambda(\theta_0)$ is

$$\lambda(\theta_0) = -2\log\left[\left(\frac{\hat{\theta}}{\theta_0}\right)^n e^{n(1-\hat{\theta}/\theta_0)}\right]$$

Likelihood ratio test of hypothesis for μ for $G(\mu, \sigma)$, known σ: To test the hypothesis $H_0 : \mu = \mu_0$ we use the likelihood ratio statistic

$$\Lambda(\mu_0) = \left(\frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}}\right)^2$$

We see that $\Lambda(\mu_0)$ has exactly a $\chi^2(1)$ distribution for all values of n since $\frac{\overline{Y}-\mu_0}{\sigma/\sqrt{n}} \sim G(0, 1)$.

5.4 Likelihood Ratio Test of Hypothesis - Multiparameter

5.5 Chapter 5 Summary

Table 5.2
Hypothesis Tests for Named Distributions
based on Asymptotic Gaussian Pivotal Quantities

Named Distribution	Point Estimate $\hat{\theta}$	Point Estimator $\tilde{\theta}$	Test Statistic for $H_0 : \theta = \theta_0$	Approximate $p - value$ based on Gaussian approximation
Binomial(n, θ)	$\frac{y}{n}$	$\frac{Y}{n}$	$\frac{\left\lvert\tilde{\theta}-\theta_0\right\rvert}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}$	$2P\left(Z \geq \frac{\left\lvert\hat{\theta}-\theta_0\right\rvert}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}\right)$ $Z \sim G(0,1)$
Poisson(θ)	$\bar{y}$	$\overline{Y}$	$\frac{\left\lvert\tilde{\theta}-\theta_0\right\rvert}{\sqrt{\frac{\theta_0}{n}}}$	$2P\left(Z \geq \frac{\left\lvert\hat{\theta}-\theta_0\right\rvert}{\sqrt{\frac{\theta_0}{n}}}\right)$ $Z \sim G(0,1)$
Exponential(θ)	$\bar{y}$	$\overline{Y}$	$\frac{\left\lvert\tilde{\theta}-\theta_0\right\rvert}{\frac{\theta_0}{\sqrt{n}}}$	$2P\left(Z \geq \frac{\left\lvert\hat{\theta}-\theta_0\right\rvert}{\frac{\theta_0}{\sqrt{n}}}\right)$ $Z \sim G(0,1)$

Note: To find $2P(Z \geq d)$ where $Z \sim G(0,1)$ in R, use $2*(1-$ `pnorm`$(d))$

Table 5.3
Hypothesis Tests for Gaussian
and Exponential Models

Model	Hypothesis	Test Statistic	Exact $p-value$
$G(\mu,\sigma)$ σ known	$H_0 : \mu = \mu_0$	$\frac{\lvert\overline{Y}-\mu_0\rvert}{\sigma/\sqrt{n}}$	$2P\left(Z \geq \frac{\lvert\bar{y}-\mu_0\rvert}{\sigma/\sqrt{n}}\right)$ $Z \sim G(0,1)$
$G(\mu,\sigma)$ σ unknown	$H_0 : \mu = \mu_0$	$\frac{\lvert\overline{Y}-\mu_0\rvert}{S/\sqrt{n}}$	$2P\left(T \geq \frac{\lvert\bar{y}-\mu_0\rvert}{s/\sqrt{n}}\right)$ $T \sim t(n-1)$
$G(\mu,\sigma)$ μ unknown	$H_0 : \sigma = \sigma_0$	$\frac{(n-1)S^2}{\sigma_0^2}$	$\min(2P\left(W \leq \frac{(n-1)s^2}{\sigma_0^2}\right),$ $2P\left(W \geq \frac{(n-1)s^2}{\sigma_0^2}\right))$ $W \sim \chi^2(n-1)$
Exponential(θ)	$H_0 : \theta = \theta_0$	$\frac{2n\bar{Y}}{\theta_0}$	$\min(2P\left(W \leq \frac{2n\bar{y}}{\theta_0}\right),$ $2P\left(W \geq \frac{2n\bar{y}}{\theta_0}\right))$ $W \sim \chi^2(2n)$

Notes:

(1) To find $P(Z \geq d)$ where $Z \sim G(0,1)$ in R, use $1-$ `pnorm`(d)

(2) To find $P(T \geq d)$ where $T \sim t(k)$ in R, use $1-$ `pt`(d,k)

(3) To find $P(W \leq d)$ where $W \sim \chi^2(k)$ in R, use `pchisq`(d,k)

6 Gaussian Response Models

6.1 Introduction

Def 40: A Gaussian response model is one for which the distribution of the response variate Y, given the associated vector of covariates $\vec{x} = (x_1, x_2, \ldots, x_k)$ for an individual unit, is of the form

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

If observations are made on n randomly selected units we write the model as

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

for $i = 1, \ldots, n$ independently.

In most examples we assume $\sigma(\vec{x}_i) = \sigma$ is constant. The choice of $\mu(\vec{x})$ is guided by past information and on current data from the population/process. We often assume $\mu(\vec{x}_i)$ is a linear function of the covariates. These models are called Gaussian linear models or linear regression models and can be written as

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

for $i = 1, \ldots, n$ independently with

$$\mu(\vec{x}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

where $\vec{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ is the vector of known covariates associated with unit i and $\beta_0, \ldots, \beta_k$ are unknown parameters. β_j's are called the regression coefficients.
Remark: Sometimes the model is written as

$$Y_i = \mu(\mathbf{x}_i) + R_i$$

where $R_i \sim G(0, \sigma)$. In this form we see that Y_i is the sum of a deterministic componenet, $\mu(\mathbf{x}_i)$ (a constant), and a stochastic component, R_i (a random variable).

$G(\mu, \sigma)$ **Model**: Suppose $Y \sim G(\mu, \sigma)$ models a response variate y in some population/process. A random sample $Y_1, Y_2, \ldots, Y_n$ is selected, and we want to estimate the model parameters and possibly to test hypotheses about them. The model is in the form

$$Y_i = \mu + R_i$$

where $R_i \sim G(0, \sigma)$ so this is a special case of the Gaussian response model in which the mean function is constant. The estimator of the parameter μ that we used is the maximum likelihood estimator $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$. This estimator is also a "least squares estimator". $\overline{Y}$ has the property that it is closer to the data than any other constant, or

$$\min_{\mu} \sum_{i=1}^{n} (Y_i - \mu)^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

6.2 Simple Linear Regression

Consider the model with independent Y_i's such that

$$Y_i \sim G(\mu(x_i), \sigma) \text{ where } \mu(x_i) = \alpha + \beta x_i$$

This is of the form $\mu(\vec{x}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$ with (β_0, β_1) replaced by (α, β). The x_i's assumed to be known constants. The unknown parameters are α, β, σ. The likelihood function for (α, β, σ) is

$$\begin{aligned} L(\alpha, \beta, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right] \\ &= \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right] \end{aligned}$$

The log likelihood function is

$$l(\alpha, \beta, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

where both L and l are for $\alpha \in \mathbb{R}, \beta \in \mathbb{R}, \sigma > 0$.

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \\ \frac{\partial l}{\partial \beta} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i = 0 \\ \frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = 0 \end{aligned}$$

The maximum likelihood estimates are:

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n} x_i(x_i - \overline{x})} = \frac{S_{xy}}{S_{xx}} \\ \hat{\alpha} &= \overline{y} - \hat{\beta}\overline{x} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n}\left(S_{yy} - \hat{\beta}S_{xy}\right) \end{aligned}$$

where

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

Least squares estimation

If we are given data $(x_i, y_i), i = 1, 2, \ldots, n$ then one criterion which could be used to obtain a line of best fit to these data is to fit the line which minimizes the sum of the squares of the distances between the observed points, $(x_i, y_i), i = 1, 2, \ldots, n$, and the fitted line $y = \alpha + \beta x$. Mathematically we want to find the values of α and β which minimize the function

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

Such estimates are called least squares estimates. To find the least squares estimates we solve the two equations:

$$\frac{\partial g}{\partial \alpha} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0$$
$$\frac{\partial g}{\partial \beta} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0$$

These are also the maximum likelihood equations from above.
The least squares estimates and the MLEs obtained are the same estimates. Note the line $y = \hat{\alpha} + \hat{\beta} x$ is often called the fitted regression line for y on x or simply the fitted line.

Interpretation of β: β is the change in the mean response variate in the study population for every one explanatory variate increase.

Distribution of the estimator $\tilde{\beta}$

The maximum likelihood estimator corresponding to $\tilde{\beta}$ is

$$\tilde{\beta} = \frac{1}{S_{xx}} \sum_{i=1}^{n} x_i (Y_i - \overline{Y}) = \sum_{i=1}^{n} a_i Y_i$$

where $a_i = \frac{(x_i - \overline{x})}{S_{xx}}$ since $\sum_{i=1}^{n} x_i (Y_i - \overline{Y}) = \sum_{i=1}^{n} (x_i - \overline{x}) Y_i$.
$\tilde{\beta}$ is a linear combination of the Gaussian random variables Y_i and therefore has a Gaussian distribution.

The identities

$$\sum_{i=1}^{n} a_i = 0, \sum_{i=1}^{n} a_i x_i = 1, \sum_{i=1}^{n} a_i^2 = \frac{1}{S_{xx}}$$

give us

$$E(\tilde{\beta}) = \beta$$
$$Var(\tilde{\beta}) = \frac{\sigma^2}{S_{xx}}$$

Therefore,

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

Confidence Intervals for β and test of hypothesis of no relationship

Although the MLE of σ^2 is

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

we will estimate σ^2 using

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

since $E(S_e^2) = \sigma^2$ where

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

Confidence intervals for β are important because the parameter β represents the increase in the mean value of the response Y. If $\beta = 0$, then x has no effect on the mean of Y. Since

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

holds independently of

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

Then by Theorem 32,

$$\frac{\tilde{\beta} - \beta}{S_e/\sqrt{S_{xx}}} \sim t(n-2)$$

which is a pivotal quantity to obtain confidence intervals and construct hypothesis tests for β.
The confidence interval is:

$$\hat{\beta} \pm t^* \frac{s_e}{\sqrt{S_{xx}}}$$

To test the hypothesis of no relationship, $H_0 : \beta = 0$ we use test statistic

$$\frac{\left|\tilde{\beta} - 0\right|}{S_e/\sqrt{S_{xx}}}$$

with observed value

$$\frac{\left|\hat{\beta} - 0\right|}{s_e/\sqrt{S_{xx}}}$$

and $p - value$ given by

$$\begin{aligned} p - value &= P\left(|T| \geq \frac{\left|\hat{\beta} - 0\right|}{s_e/\sqrt{S_{xx}}}\right) \\ &= 2\left[1 - P\left(T \leq \frac{\hat{\beta} - 0}{s_e/\sqrt{S_{xx}}}\right)\right] \qquad T \sim t(n-2) \end{aligned}$$

A $100p\%$ confidence interval $P(a \le U \le b) = p$ for σ^2 is

$$\left[\frac{(n-2)s_e^2}{b}, \frac{(n-2)s_e^2}{a}\right]$$

and $100p\%$ confidence interval for σ is

$$\left[s_e\sqrt{\frac{(n-2)}{b}}, s_e\sqrt{\frac{(n-2)}{a}}\right]$$

where $P(U \le a) = P(U > b) = \frac{1-p}{2}$ and $U \sim \chi^2(n-2)$. σ corresponds to the variability in the response Y for each value of the covariate x.

Confidence intervals for the mean response $\mu(x) = \alpha + \beta x$

The maximum likelihood estimator of $\mu(x)$ obtains by by replacing the unknown parameters by their maximum likelihood estimators,

$$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x = \overline{Y} + \tilde{\beta}(x - \overline{x})$$

since $\tilde{\alpha} = \overline{Y} - \tilde{\beta}\overline{x}$. Since

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

we can rewrite it as

$$\tilde{\mu}(x) = \overline{Y} + \tilde{\beta}(x - \overline{x}) = \sum_{i=1}^{n} b_i Y_i \text{ where } b_i = \frac{1}{n} + (x - \overline{x})\frac{(x_i - \overline{x})}{S_{xx}}$$

Since $\tilde{\mu}(x)$ is a linear combination of Gaussian random variables it has a Gaussian distribution. We use the following identities

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

to give us

$$E[\tilde{\mu}(x)] = \mu(x)$$

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

Therefore, Therefore,

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

Since

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

holds independently of $\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2)$ then by Theorem 32 we get the pivotal quantity

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

The $100p\%$ confidence interval for $\mu(x)$ is

$$\hat{\mu}(x) \pm t^* s_e\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$$

where $\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x$.
Remark: Note that $\alpha = \mu(0)$, then a 95% confidence interval for α is given by the above with $x = 0$:

$$\hat{\alpha} \pm t^* s_e\sqrt{\frac{1}{n} + \frac{(\overline{x})^2}{S_{xx}}}$$

Prediction Interval for Future Response

Note that $Y \sim G(\mu(x), \sigma)$ or alternatively $Y = \mu(x) + R$ where $R \sim G(0, \sigma)$. For a point estimator of Y it is natural to use the maximum likelihood estimator $\tilde{\mu}(x)$ of $\mu(x)$. We are interested in the random variable $Y - \tilde{\mu}(x)$.

$$Y - \tilde{\mu}(x) = Y - \mu(x) + \mu(x) - \tilde{\mu}(x) = R + [\mu(x) - \tilde{\mu}(x)]$$

Then

$$E[Y - \tilde{\mu}(x)] = 0$$

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

Therefore,

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

or

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

Using Theorem 32,

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

The $100p\%$ prediction interval is

$$\hat{\mu}(x) \pm t^* s_e\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$$

Summary

Table 6.1
Confidence/Prediction Intervals for
Simple Linear Regression Model

Unknown Quantity	Estimate	Estimator	Pivotal Quantity	$100p\%$ Confidence/ Prediction Interval
β	$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$	$\tilde{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{S_{xx}}$	$\frac{\tilde{\beta}-\beta}{S_e/\sqrt{S_{xx}}} \sim t\left(n-2\right)$	$\hat{\beta} \pm a s_e/\sqrt{S_{xx}}$
α	$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$	$\tilde{\alpha} = \overline{Y} - \tilde{\beta}\bar{x}$	$\frac{\tilde{\alpha}-\alpha}{S_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}} \sim t\left(n-2\right)$	$\hat{\alpha} \pm a s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}$
$\mu\left(x\right) = \alpha + \beta x$	$\hat{\mu}\left(x\right) = \hat{\alpha} + \hat{\beta}x$	$\tilde{\mu}\left(x\right) = \tilde{\alpha} + \tilde{\beta}x$	$\frac{\tilde{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t\left(n-2\right)$	$\hat{\mu}\left(x\right) \pm a s_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$
σ^2	$s_e^2 = \frac{S_{yy}-\hat{\beta}S_{xy}}{n-2}$	$S_e^2 = \frac{\sum_{i=1}^{n}\left(Y_i-\tilde{\alpha}-\tilde{\beta}x_i\right)^2}{n-2}$	$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2\left(n-2\right)$	$\left[\frac{(n-2)s_e^2}{c}, \frac{(n-2)s_e^2}{b}\right]$
Y			$\frac{Y-\tilde{\mu}(x)}{S_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t\left(n-2\right)$	Prediction Interval $\hat{\mu}\left(x\right) \pm a s_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$

Notes: The value a is given by $P\left(T \leq a\right) = \frac{1+p}{2}$ where $T \sim t\left(n-2\right)$.
The values b and c are given by $P\left(W \leq b\right) = \frac{1-p}{2} = P\left(W > c\right)$ where $W \sim \chi^2\left(n-2\right)$.

Table 6.2
Hypothesis Tests for
Simple Linear Regression Model

Hypothesis	Test Statistic	$p - value$
$H_0 : \beta = \beta_0$	$\frac{\left\lvert\tilde{\beta}-\beta_0\right\rvert}{S_e/\sqrt{S_{xx}}}$	$2P\left(T \geq \frac{\left\lvert\hat{\beta}-\beta_0\right\rvert}{s_e/\sqrt{S_{xx}}}\right)$ where $T \sim t\,(n-2)$
$H_0 : \alpha = \alpha_0$	$\frac{\lvert\tilde{\alpha}-\alpha_0\rvert}{S_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}$	$2P\left(T \geq \frac{\lvert\hat{\alpha}-\alpha_0\rvert}{s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}\right)$ where $T \sim t\,(n-2)$
$H_0 : \sigma = \sigma_0$	$\frac{(n-2)S_e^2}{\sigma_0^2}$	$\min\left(2P\left(W \leq \frac{(n-2)s_e^2}{\sigma_0^2}\right), 2P\left(W \geq \frac{(n-2)s_e^2}{\sigma_0^2}\right)\right)$ $W \sim \chi^2\,(n-2)$

6.3 Checking the Model

There are two main components in Gaussian linear response models:

- The assumption that $E(Y_i) = \mu(x_i)$ is a linear combination of observed covariates with unknown coefficients.
- The assumption that the random variables Y_i (given any covariates x_i) has a Gaussian distribution with constant standard deviation σ.

Scatterplot with Fitted Line: If there is only one x covariate, a scatterplot of the data with the fitted line superimposed can be used. Such a plot is checking whether the response variate can be modeled by a random variable whose mean is a linear function of the explanatory variate and whose standard deviation is constant over the range of values of the explanatory variate.

Residual Plots: Consider the simple linear regression model for which $Y_i \sim G(\mu_i, \sigma)$ where $\mu_i = \alpha + \beta x_i$ and $R_i = Y_i - \mu_i \sim G(0, \sigma), i = 1, 2, \ldots, n$ independently. Residuals are defined as the difference between the observed response y_i and the fitted response $\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i$, that is $\hat{r}_i = y_i - \hat{\mu}_i, i = 1, 2, \ldots, n$.
Often we prefer to use standardized residuals

$$\hat{r}_i^* = \frac{\hat{r}_i}{s_e} = \frac{y_i - \hat{\mu}_i}{s_e} = \frac{y_i - \hat{\alpha} - \hat{\beta} x_i}{s_e} \text{ for } i = 1, 2, \ldots, n$$

Since $\hat{r}_i$'s behave roughly like a random sample from the $G(0, \sigma)$ distribution, the $\hat{r}_i^*$'s should behave like a random sample from the $G(0, 1)$ distribution.
Since $P(-3 \leq Z \leq 3) = 0.9973$ where $Z \sim G(0, 1)$, then roughly 99.73% of the observations should lie in the interval $[-3, 3]$.

Here are 3 residual plots which can be used to check model assumptions.

1. Plot points $(x_i, \hat{r}_i^*), i = 1, 2, \ldots, n$.
2. Plot points $(\hat{\mu}_i, \hat{r}_i^*), i = 1, 2, \ldots, n$.
3. Plot a Gaussian qqplot of the residuals $\hat{r}_i^*$.

If the model is satisfactory then the points in plots 1 and 2 should lie roughly within a horizontal band of constant width between -3 and 3. Approximately half the points should lie on either side of the line $\hat{r}_i^* = 0$.
If the model is satisfactory then the points in the qqplot 3 should lie roughly along a straight line with more variability in the points at both ends of the line.

Systematic departures from the expected pattern suggest the model assumptions do not hold. For example, if the points form a U-shaped pattern, then $\mu_i = \mu(x_i)$ is not correctly specified. A quadratic form $\mu(x_i) = \alpha + \beta x_i + \gamma x_i^2$ might be a better fit.
Sometimes the spread of the points about the line $\hat{r}_i^* = 0$ increases/decreases as x increases/decreases. We can transform the response variate to solve the non-constant variance, i.e., heteroscedasticity. $\log y$ and $\sqrt{y}$ are frequently used.

6.4 Comparison of Two Population Means

Two Gaussian Populations with Common Variance

Suppose $Y_{11},\ldots,Y_{1n_1}$ is a random sample from $G(\mu_1,\sigma)$ distribution and independently $Y_{21},\ldots,Y_{2n_2}$ is a random sample from a $G(\mu_2,\sigma)$ distribution. We can conform the notation by stacking these two sets of observations in a vector of $n = n_1 + n_2$ observations:

$$(Y_{11},\ldots,Y_{1n_1},Y_{21},\ldots,Y_{2n_2})^T$$

The likelihood function for μ_1,μ_2,σ is

$$L(\mu_1,\mu_2,\sigma) = \prod_{j=1}^{2}\prod_{i=1}^{n_j}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(y_{ji}-\mu_j)^2\right]$$

Maximum likelihood estimates are:

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i} = \overline{y}_1\\ \hat{\mu}_2 &= \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i} = \overline{y}_2\\ \hat{\sigma}^2 &= \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(y_{1i}-\overline{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i}-\overline{y}_2)^2\right]\end{aligned}$$

An estimate of the variance σ^2 called the pooled estimate of variance is

$$\begin{aligned}s_p^2 &= \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(y_{1i}-\overline{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i}-\overline{y}_2)^2\right]\\ &= \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\\ &= \frac{n_1+n_2}{n_1+n_2-2}\hat{\sigma}^2\end{aligned}$$

where

$$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_{1i}-\overline{y}_1)^2,\ \frac{1}{n_2-1}\sum_{i=1}^{n_2}(y_{2i}-\overline{y}_2)^2$$

are the sample variances obtained from the individual samples. The estimate s_p^2 can be written as

$$s_p^2 = \frac{w_1 s_1^2 + w_2 s_2^2}{w_1+w_2}$$

to show that s_p^2 is a weighted average of the sample variances s_j^2 with weights equal to $w_j = n_j - 1$.

Confidence intervals for $\mu_1 - \mu_2$

$$E(\overline{Y}_1 - \overline{Y}_2) = \mu_1 - \mu_2$$

$$Var(\overline{Y}_1 - \overline{Y}_2) = Var(\overline{Y}_1) + Var(\overline{Y}_2) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

An estimator for the variance from the pooled data is

$$S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

This has $n_1 + n_2 - 2$ degrees of freedom.

Theorem 41: If $Y_{11}, \ldots, Y_{1n_1}$ is a random sample from a $G(\mu_1, \sigma)$ distribution and independently $Y_{21}, \ldots, Y_{2n_2}$ is a random sample from a $G(\mu_2, \sigma)$ distribution then

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

and

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{j=1}^{2} \sum_{i=1}^{n_j} (Yji - \overline{Y}_j)^2 \sim \chi^2(n_1 + n_2 - 2)$$

Confidence Interval for $\mu_1 - \mu_2$:

$$\overline{y}_1 - \overline{y}_2 \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $P(T \le t^*) = (1 + p)/2$ and $T \sim t(n_1 + n_2 - 2)$.

Hypothesis Testing for $\mu_1 - \mu_2$: To test $H_0 : \mu_1 - \mu_2 = 0$ we use the test statistic

$$D = \frac{\left| \overline{Y}_1 - \overline{Y}_2 - 0 \right|}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with

$$p - value = P \left(|T| \ge \frac{|\overline{y}_1 - \overline{y}_2 - 0|}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right) = 2 \left[1 - P \left(T \le \frac{|\overline{y}_1 - \overline{y}_2 - 0|}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right) \right]$$

where $T \sim t(n_1 + n_2 - 2)$.

Confidence interval for σ:

$$\left[\sqrt{\frac{(n_1 + n_2 - 2)s_p^2}{b}}, \sqrt{\frac{(n_1 + n_2 - 2)s_p^2}{a}} \right]$$

where $P(U \le a) = (1 - p)/2$, $P(U \le b) = (1 + p)/2$, $U \sim \chi^2(n_1 + n_2 - 2)$.

Two Gaussian Populations with Unequal Variances

Assume that $Y_{11}, \ldots, Y_{1n_1}$ is a random sample from a $G(\mu_1, \sigma_1)$ distribution and independently $Y_{21}, \ldots, Y_{2n_2}$ is a random sample from a $G(\mu_2, \sigma_2)$ but $\sigma_1 \neq \sigma_2$.
If σ_1 and σ_2 are known then we can use the pivotal quantity

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim G(0, 1)$$

Confidence interval for $\mu_1 - \mu_2$:

$$\overline{y}_1 - \overline{y}_2 \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $P(Z \leq z^*) = (1+p)/2$ and $Z \sim G(0,1)$.

Hypothesis testing for $\mu_1 - \mu_2$: To test $H_0 : \mu_1 - \mu_2 = 0$ we use test statistic

$$D = \frac{\left|\overline{Y}_1 - \overline{Y}_2 - 0\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

with

$$p - value = P\left(|Z| \geq \frac{|\overline{y}_1 - \overline{y}_2 - 0|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) = 2\left[1 - P\left(Z \leq \frac{|\overline{y}_1 - \overline{y}_2 - 0|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)\right]$$

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

In the case that σ_1 and σ_2 are unknown, we can replace them with their estimators and if n_1 and n_2 are both large, then

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

and the confidence interval is

$$\overline{y}_1 - \overline{y}_2 \pm z^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Table 6.3
Confidence Intervals for
Two Sample Gaussian Model

Model	Parameter	Pivotal Quantity	$100p\%$ Confidence Interval
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ σ_1, σ_2 known	$\mu_1 - \mu_2$	$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ $\sim G(0,1)$	$\bar{y}_1 - \bar{y}_2 \pm a\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ $\sigma_1 = \sigma_2 = \sigma$ σ unknown	$\mu_1 - \mu_2$	$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $\sim t(n_1 + n_2 - 2)$	$\bar{y}_1 - \bar{y}_2 \pm b s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
$G(\mu_1, \sigma)$ $G(\mu_2, \sigma)$ μ_1, μ_2 unknown	σ^2	$\frac{(n_1+n_2-2)S_p^2}{\sigma^2}$ $\sim \chi^2(n_1 + n_2 - 2)$	$\left[\frac{(n_1+n_2-2)s_p^2}{d}, \frac{(n_1+n_2-2)s_p^2}{c}\right]$
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ $\sigma_1 \neq \sigma_2$ σ_1, σ_2 unknown	$\mu_1 - \mu_2$	asymptotic Gaussian pivotal quantity $\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ for large n_1, n_2	approximate $100p\%$ confidence interval $\bar{y}_1 - \bar{y}_2 \pm a\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Notes:
The value a is given by $P(Z \leq a) = \frac{1+p}{2}$ where $Z \sim G(0,1)$.
The value b is given by $P(T \leq b) = \frac{1+p}{2}$ where $T \sim t(n_1 + n_2 - 2)$.
The values c and d are given by $P(W \leq c) = \frac{1-p}{2} = P(W > d)$ where $W \sim \chi^2(n_1 + n_2 - 2)$.

Table 6.4
Hypothesis Tests for
Two Sample Gaussian Model

Model	Hypothesis	Test Statistic	$p-value$
$G(\mu_1,\sigma_1)$ $G(\mu_2,\sigma_2)$ σ_1, σ_2 known	$H_0:\mu_1=\mu_2$	$\frac{\lvert\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$2P\left(Z\geq\frac{\lvert\bar{y}_1-\bar{y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right)$ $Z\sim G(0,1)$
$G(\mu_1,\sigma)$ $G(\mu_2,\sigma)$ σ unknown	$H_0:\mu_1=\mu_2$	$\frac{\lvert\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)\rvert}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$2P\left(T\geq\frac{\lvert\bar{y}_1-\bar{y}_2-(\mu_1-\mu_2)\rvert}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)$ $T\sim t(n_1+n_2-2)$
$G(\mu_1,\sigma)$ $G(\mu_2,\sigma)$ μ_1, μ_2 unknown	$H_0:\sigma=\sigma_0$	$\frac{(n_1+n_2-2)S_p^2}{\sigma_0^2}$	$\min(2P\left(W\leq\frac{(n_1+n_2-2)s_p^2}{\sigma_0^2}\right),$ $2P\left(W\geq\frac{(n_1+n_2-2)s_p^2}{\sigma_0^2}\right))$ $W\sim\chi^2(n_1+n_2-2)$
$G(\mu_1,\sigma_1)$ $G(\mu_2,\sigma_2)$ $\sigma_1\neq\sigma_2$ σ_1, σ_2 unknown	$H_0:\mu_1=\mu_2$	$\frac{\lvert\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$	approximate $p-value$ $2P\left(Z\geq\frac{\lvert\bar{y}_1-\bar{y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\right)$ $Z\sim G(0,1)$

Comparison of Means Using Paired Data

There are two types of Gaussian models which can be used to model paired data. The first involves a Bivariate Normal distribution for (Y_{1i}, Y_{2i}) where $\sigma^2 = Var(Y_{1i}) + Var(Y_{2i}) - 2Cov(Y_{1i}, Y_{2i})$. We can analyze the within-pair differences $Y_i = Y_{1i} - Y_{2i}$

$$Y_i = Y_{1i} - Y_{2i} \sim G(\mu_1 - \mu_2, \sigma),\ i = 1, \ldots, n \text{ independently}$$

or

$$Y_i \sim G(\mu, \sigma)$$

where $\mu = \mu_1 - \mu_2$. The methods for single parameters testing can be used.
The second Gaussian model used with paired data assumes

$$Y_{1i} \sim G(\mu_1 + \alpha_i, \sigma_1^2), Y_{2i} \sum G(\mu_2 + \alpha_i, \sigma_2^2) \text{ independently}$$

where α_i's are unknown constants. This model has the same Gaussian distribution as $Y_i \sim G(\mu, \sigma)$ with

$$E(Y_{1i} - Y_{2i}) = \mu_1 - \mu_2 = \mu$$

notice α_i cancel, and

$$Var(Y_{1i} - Y_{2i}) = \sigma_1^2 + \sigma_2^2 = \sigma^2$$

Pairing and Experimental Design

The condition for pairing is that the association or correlation between Y_{1i} and Y_{2i} be positive. To see why the pairing is helpful in estimating the mean difference $\mu_1 - \mu_2$ suppose that $Y_{1i} \sim G(\mu_1, \sigma_1^2)$ and $Y_{2i} \sim G(\mu_2, \sigma_2^2)$ but that Y_{1i} and Y_{2i} are not necessarily independent. The estimator of $\mu_1 - \mu_2$ is

$$\overline{Y}_1 - \overline{Y}_2$$

We have

$$E(\overline{Y}_1 - \overline{Y}_2) = \mu_1 - \mu_2$$

and

$$Var(\overline{Y}_1 - \overline{Y}_2) = Var(\overline{Y}_1) + Var(\overline{Y}_2) - 2Cov(\overline{Y}_1, \overline{Y}_2) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2\frac{\sigma_{12}}{n}$$

Regression through the origin: Consider the model $Y_i \sim G(\beta x_i, \sigma), i = 1, 2, \ldots, n$ independently.

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

and

$$\tilde{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right)$$

where $s_e^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \hat{\beta} x_i)^2$ is an unbiased estimate of σ^2. The pivotal quantity is

$$\frac{\tilde{\beta} - \beta}{\frac{S_e}{\sqrt{\sum_{i=1}^n x_i^2}}} \sim t(n-1)$$

7 Multinomial Models and Goodness of Fit Tests

7.1 Likelihood Ratio Test for the Multinomial Model

Suppose data arise from a Multinomial distribution with joint probability function

$$f(y_1, y_2, \ldots, y_k; \theta_1, \theta_2, \ldots, \theta_k) = \frac{n!}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k}$$

where $y_j = 0, 1, \ldots,$ $\sum_{j=1}^{k} y_j = n$ and $\sum_{j=1}^{k} \theta_j = 1$. The likelihood function is

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

or simply

$$L(\vec{\theta}) = \prod_{j=1}^{k} \theta_j^{y_j}$$

$L(\vec{\theta})$ is maximized by $\hat{\vec{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ where $\hat{\theta}_j = \frac{y_j}{n}$.

Note: There are only $k-1$ parameters to be estimated since we can find the other using $\sum\limits_{j=1}^{k} \theta_j = 1$.

Suppose we want to test the hypothesis that $\theta_1, \ldots, \theta_k$ are related in some way, for example, they are all functions of a parameter $\vec{\alpha}$, such that

$$H_0 : \theta_j = \theta_j(\vec{\alpha}),\ j = 1, 2, \ldots, k$$

where $\vec{\alpha} = (\alpha_1, \ldots, \alpha_p)$ and $p < k-1$. p is equal to the number of parameters that need to be estimated in the model assuming the null hypothesis.

A likelihood ratio test of $H_0 : \theta_j = \theta_j(\vec{\alpha})$ is based on the likelihood ratio statistic

$$\Lambda = -2 \log \left[\frac{L(\tilde{\vec{\theta}}_0)}{L(\tilde{\vec{\theta}})} \right]$$

where $\tilde{\vec{\theta}}_0$ maximizes $L(\vec{\theta})$ assuming H_0 is true.

The test statistic can simplified. Let $\tilde{\vec{\theta}}_0 = (\theta_1(\tilde{\alpha}), \ldots, \theta_k(\tilde{\alpha}))$ denote the maximum likelihood estimator of $\vec{\theta}$ under the null hypothesis from before. Then

$$\Lambda = 2 \sum_{j=1}^{k} Y_j \log \left[\frac{\tilde{\theta}_j}{\theta_j(\tilde{\alpha})} \right]$$

Note that $\tilde{\theta}_j = Y_j / n$ and defining the expected frequencies under H_0 as

$$E_j = n\theta_j(\tilde{\alpha}),\ j = 1, \ldots, k$$

then

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

and observed value

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

where $e_j = n\theta_j(\hat{\alpha})$.
If n is large, none of the θ_j's is too small, and H_0 is true then $\Lambda \sim \chi^2(k-1-p)$ and

$$p-value = P(\Lambda \geq \lambda; H_0) \sim P(W \geq \lambda),\ W \sim \chi^2(k-1-p)$$

The expected frequences determined assuming H_0 is true should all be **at least 5** to use the Chi-squared approximation.

Pearson Goodness of Fit Test Statistic:

$$D = \sum_{j=1}^{k} \frac{(Y_j - E_j)^2}{E_j}$$

$$d = \sum_{j=1}^{k} \frac{(y_j - e_j)^2}{e_j}$$

where D has a limiting $\chi^2(k-1-p)$ distribution when H_0 is true. d takes on small values if the y_j's and e_j's are close in value and d is large if y_j's and e_j's differ greatly.

7.2 Goodness of Fit Tests

GoF and Poisson model:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} j y_j$$

$$H_0 : \theta_j = \frac{\theta^j e^{-\theta}}{j!}$$

Calculate the expected frequency for the Poisson model and calculate the likelihood ratio statistic or Pearson goodness of fit statistic and find $p-value$.
There is one parameter to be estimated so $p = 1$.

GoF and Exponential model: The probability that an observation lies in the j'th interval $I_j = (a_{j-1}, a_j)$ is

$$p_j(\theta) = \int_{a_{j-1}}^{a_j} f(t;\theta)\, dt = e^{-a_{j-1}/\theta} - e^{-a_j/\theta}$$

$$L(\theta) = \prod_{i=1}^{n} [p_j(\theta)]^{y_j}$$

Find $\hat{\theta} = \max(L(\theta))$ and use it to calculate expected frequency for the Exponential model, $e_j = np_j(\hat{\theta})$. Find the likelihood ratio statistic or Pearson goodness of fit statistic and find $p-value$.

7.3 Two-Way (Contingency) Tables

Cross-Classification of a Random Sample of Individuals: Suppose that individuals or items in a population can be classified according to each of two factors A and B. For A, an individual can be any of a mutually exclusive types $A_1, \ldots, A_a$ and for B an individual can be any of b mutually exclusive types $B_1, \ldots, B_b$, where $a \geq 2$ and $b \geq 2$.
If a random sample of n individuals is selected, let y_{ij} denote the number that have A-type A_i and B-type B_j.

$A \setminus B$	B_1	B_2	$\ldots$	B_b	Total
A_1	y_{11}	y_{12}	$\cdots$	y_{1b}	r_1
A_2	y_{21}	y_{22}	$\cdots$	y_{2b}	r_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_a	y_{a1}	y_{a2}	$\cdots$	y_{ab}	r_a
Total	c_1	c_2	$\cdots$	c_b	n

where $r_i = \sum\limits_{j=1}^{b} y_{ij}$ are the row totals, $c_j = \sum\limits_{i=1}^{a} y_{ij}$ are the column totals, and $\sum\limits_{i=1}^{a}\sum\limits_{j=1}^{b} y_{ij} = n$. Let θ_{ij} be the probability a randomly selected individual is combined type (A_i, B_j) and note that $\sum\limits_{i=1}^{a}\sum\limits_{j=1}^{b} \theta_{ij} = 1$. The $a \times b$ frequencies $(Y_{11}, \ldots, Y_{ab})$ follow a Multinomial distribution with $k = ab$ classes.

To test independence of the A and B classifications, we test the hypothesis

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

where $0 < \alpha_i, \beta_j < 1$, $\sum_{i=1}^{a} \alpha_i = 1$, $\sum_{j=1}^{b} \beta_j = 1$. Note

$$\alpha_i = P(\text{an individual is type } A_i), \;\; \beta_j = P(\text{an individual is type } B_j)$$

and $\theta_{ij} = \alpha_i \beta_j$ is the definition of independent events: $P(A_i \cap B_j) = P(A_i)P(B_j)$.

The number of parameters estimated under the null hypothesis is $p = (a-1) + (b-1) = a + b - 2$ and $k = ab$.

$$L_1(\vec{\alpha}, \vec{\beta}) = \prod_{i=1}^{a}\prod_{j=1}^{b} (\alpha_i \beta_j)^{y_{ij}}$$

$l(\vec{\alpha}, \vec{\beta})$ must be maximized subject to constraints that the sum α_i and β_j is 1.
The maximum likelihood estimates are

$$\hat{\alpha}_i = \frac{r_i}{n}, \;\; \hat{\beta}_j = \frac{c_j}{n} \;\; i = 1, \ldots, a, j = 1, \ldots, b$$

and expected frequencies are

$$e_{ij} = n \hat{\alpha}_i \hat{\beta}_j = \frac{r_i c_j}{n}$$

The observed likelihood ratio statistic for H_0 is

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

The degrees of freedom for Chi-squared approximation are

$$k - 1 - p = (ab - 1) - (a - 1 + b - 1) = (a - 1)(b - 1)$$

$$p - value = P(\Lambda \geq \lambda; H_0) \approx P(W \geq \lambda),\ W \sim \chi^2((a-1)(b-1))$$

8 Causal Relationships

8.1 Establishing Causation

Causation: If all other factors that affect y are held constant, let us change x (or observe different values of x) and see if some specified attribute of y changes. If the specified attribute of y changes then we say x has a causal effect on y.

Correlation $\neq$ Causation: Association (statistical dependence) between two variates x and y does not imply that a causal relationship exists.

8.2 Experimental Studies

Suppose we want to investigate whether a variate x has a causal effect on a response variate Y. In an experimental setting we can control the values of x that a unit "sees". In addition, we can use one or both of the following devices for ruling out alternative explanations for any observed changes in y that might be caused by x:

1. Hold other possible explanatory variates fixed.
2. Use randomization to control for other variates.

8.3 Observational Studies

In observational studies there are often unmeasured factors that affect the response variate y. If these factors are also related to the explanatory variate x whose (potential) causal effect we are trying to assess, then we cannot easily make any inferences about causation. For this reason, we try in observational studies to measure other important factors besides x.

Simpson's Paradox: In probabilistic terms, it says that for events A, B_1, B_2 and $C_1, \dots, C_k$, we can have

$$P(A|B_1C_i) > P(A|B_2C_i),\ \forall i = 1, \dots, k$$

but have

$$P(A|B_1) < P(A|B_2)$$

Note that $P(A|B_1) = \sum_{i=1}^{k} P(A|B_1C_i)P(C_i|B_1)$ and similarly for $P(A|B_2)$, so they depend on what $P(C_i|B_1)$ and $P(C_i|B_2)$ are.

Guidelines for causal association: In the case an experimental study cannot be conducted:

- The association between x and y must be observed in many studies of different types among different groups. This reduces the chance that an observed association is due to a defect in one type of study or a peculiarity in one group of subjects.
- The association between x and y must continue to hold when the effects of plausible confounding variates are taken into account.
- There must be a plausible explanation for the direct influence of x on y, so that a causal link does not depend on the observed association alone.
- There must be a consistent response, that is, y always increases/decreases when x increases.