2016 09 25

Review of Math 494 (Mathematic Statistic)

Qian Liu

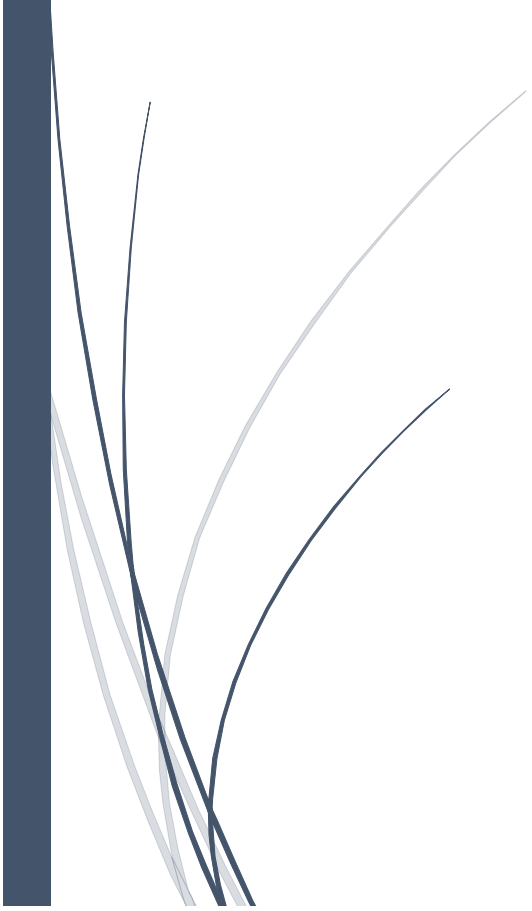


Table of Contents

Review of Math 494 (Mathematic Statistic)

Object: Define a mathematical frame work for statistics

Basic ideal: Gain an understanding about the distribution of variables based on random samples. (making inference on variable)

1. Definition: Statistics, estimators

1.1. Statistics

A **Statistic** is a function of your sample: T = f(X1;X2; : : :Xn);

- e.x. the sample mean and sample standard deviation

1.2. estimators

An "**estimator**" or "point estimate" is a statistic (that is, a function of the data) that is used to infer the value of an unknown parameter in a statistical model.

- e.x. the sample mean and sample standard deviation of a normal distribution statistical model.
- An estimator T_n is an **unbiased** estimator for μ if $E(T_n) = \mu$.
- An estimator T_n is a **consistent** estimator for μ if $Tn \xrightarrow{p} \mu \; as \; n \to \infty$(asymptotically unbiased).
- An estimator T_n is a **efficient** estimator for μ if $Var(Tn) = minimum\ variance\ bound\ (MVB)$.

2. Distributions of estimators(statistics)

The mean and the variance often tell us a great deal about the distribution, but sometimes not a lot about the shape of the distribution.

2.1. The central limit theorem

Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with $\mathbb{E}(X_i) = \mu$ and $V(X_i) = \sigma^2$, and let $S_n = X_1 + X_2 + \ldots + X_n$. Then

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1), \quad \text{as } n \to \infty.$$

In other words for large n, $S_n \overset{d}{\approx} N(n\mu, n\sigma^2)$ or $\overline{X} \overset{d}{\approx} N(\mu, \frac{\sigma^2}{n})$.

Proof using moment generating function:

The main tool we are going to use is the so-called *moment generating function*, defined as follows for a random variable X:

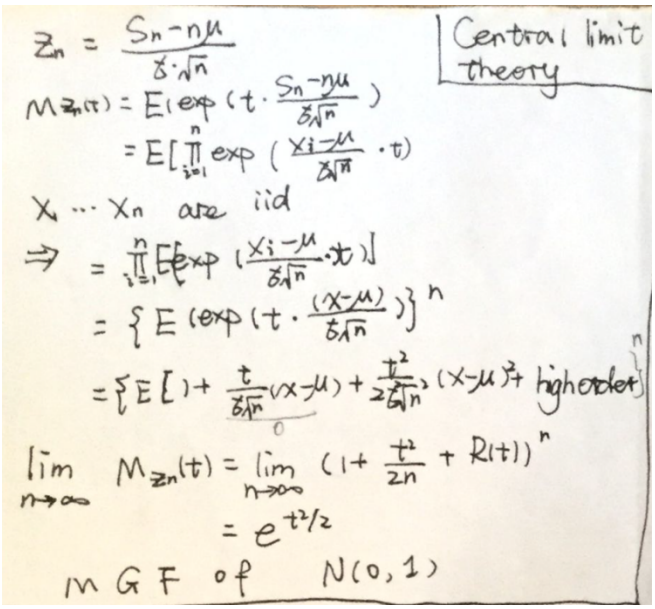
$$M_X(t) = \mathbb{E}[e^{tX}].$$

Expanding the Taylor series of e^{tX}, we discover the reason it's called the moment generating function:

$$M_X(t) = \sum_{n=0}^{\infty} \frac{\mathbb{E}[X^n]}{n!} t^n.$$

The moment generating function is thus just the exponential generating function for the moments of X. In particular,

$$M_X^{(n)}(0) = \mathbb{E}[X^n].$$



2.2. Sample mean

$$\bar{X} = \sum_{i=1}^{n} X_i$$

$$\mathrm{E}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$$

$$\mathrm{var}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}$$

2.3. Sample variance

$E(S^2) = \sigma^2$,

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right).$$

Proof:

$$\begin{aligned}
E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) &= \sum_{i=1}^{n} E(X_i{}^2) - 2E(X_i\bar{X}) + E(\bar{X}^2) \\
&= \sum_{i=1}^{n} E(X_i{}^2) - 2E(X_i\bar{X}) + E(\bar{X}^2) \\
&= \sum_{i=1}^{n} E(X_i{}^2) - E(\bar{X}^2) \\
&= \sum_{i=1}^{n} E(X_i{}^2) - E(\bar{X}^2) \\
&= \sum_{i=1}^{n} E(X_i{}^2 - \mu^2) - E(\bar{X}^2 - \mu^2) \\
&= \sum_{i=1}^{n} \sigma^2 - \frac{\sigma^2}{n} \\
&= (n-1)\sigma^2
\end{aligned}$$

It can also be shown (with a great deal of effort) that

$$\mathrm{Var}(S^2) = \frac{\nu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)},$$

where $\nu_4 = \mathbb{E}((X-\mu)^4)$.

Note that there isn't an asymptotic result for the sample variance similarto the central limit theorem for the sample mean. The only case where we can say something about the distribution of S2 is when sampling from a normal distribution.

2.4. Statistics have a normal distribution

$X_1, X_2, X_3 \ldots X_n$ are i.i.d. random variables with distribution $N(\mu, \sigma^2)$.

$$\bar{X} \xrightarrow{d} N(\mu, \frac{\sigma^2}{n})$$

2.5. Statistics have a ${\chi_n}^2$ distribution

$Z_1, Z_2, Z_3 \quad Z_n$ are all iid standard normal random variables ($Z_i \xrightarrow{d} N(0,1)$), and $U = \sum_{i=1}^{n} {Z_i}^2$, then we say U has a χ^2distribution with n degrees of freedom.

$$U \xrightarrow{d} {\chi_n}^2$$

Noting that it can be shown that S^2 and $\bar{X}$ are independent, we can show

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 - \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2,$$

and hence using moment generating functions

$$\frac{(n-1)S^2}{\sigma^2} \overset{d}{=} \chi^2_{n-1}$$

2.6. Statistics have a t_v distribution

If $U \overset{d}{\to} {\chi_v}^2$, and $Z \overset{d}{\to} N(0,1)$, and U, Z are independent then $T = \frac{Z}{\sqrt{U/v}}$ is said to follow the t-distribution with v degrees of freedom.We write

$$T \overset{d}{\to} t_v$$

The most useful fact about the t distribution is that

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \overset{d}{=} t_{n-1},$$

and hence we can use the t-distribution to construct confidence intervals.

2.7. Statistics have a F-distribution

If $U \overset{d}{\to} {\chi_m}^2$, and $V \overset{d}{\to} {\chi_n}^2$, and U, V are independent then $Z = \frac{U/m}{V/n}$ is said to follow the F-distribution with v degrees of freedom. We write

$$Z \overset{d}{\to} F_{m,n}$$

3 framework for deriving estimators (Methods of estimation)

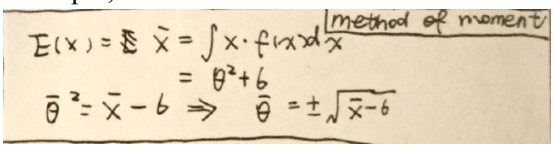
So far we have seen numerous estimators, and properties of estimators. However, where exactly do these estimators come from?

3.1 Method of moments (MM)

The method of moments essentially uses the moments of your sample to estimate the parameters of interest.

3.1.1 Single parameters

Example,



3.1.2 Multiple parameters

In the case where there are multiple parameters, we use an appropriate number of moments of the distribution in question.

Example,

Let $X_1, X_2, ..., X_n$ be normal random variables with mean μ and variance σ^2. What are the method of moments estimators of the mean μ and variance σ^2?

Solution. The first and second theoretical moments about the origin are:

$$E(X_i) = \mu \text{ and } E(X_i^2) = \sigma^2 + \mu^2$$

(Incidentally, in case it's not obvious, that second moment can be derived from manipulating the shortcut formula for the variance.) In this case, we have two parameters for which we are trying to derive method of moments estimators. Therefore, we need two equations here. Equating the first theoretical moment about the origin with the corresponding sample moment, we get:

$$E(X) = \mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

And, equating the second theoretical moment about the origin with the corresponding sample moment, we get:

$$E(X^2) = \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Now, the first equation tells us that the method of moments estimator for the mean μ is the sample mean:

$$\hat{\mu}_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

And, substituting the sample mean in for μ in the second equation and solving for σ^2, we get that the method of moments estimator for the variance σ^2 is:

$$\hat{\sigma}^2_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$$

which can be rewritten as:

$$\hat{\sigma}^2_{MM} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Again, for this example, the method of moments estimators are the same as the maximum likelihood estimators.

3.1.3 Comment on MM

Due to the law of large numbers, as long as θ has an inverse, MM estimators are consistent. However, in general they are not unbiased, and often not very efficient. The main reason to use a MM estimator is that they are usually very easily derived when other methods may be too difficult to use.

3.2 Method of Maximum Likelihood

We begin with sample of iid random variables $X_1, X_2, \ldots X_n$, with common density function $f_X(x|\theta)$ where θ is a parameter of the distribution.

Define the likelihood function

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i|\theta).$$

The maximum likelihood estimator of θ, denoted by $\hat{\theta}$ is defined such that

$$L(\hat{\theta}) \geq L(\theta)$$

for all θ.

It is usually easier to compute the logarithm of the likelihood function and maximize that.
The sample does not necessarily need to be iid. In full generality the likelihood function is the joint distribution of the sample.

3.2.1 property

- Biased: can be biased
- Consistence: appears to be asymptotically unbiased, so consistent. (under certain conditions)
- Efficiency: there exists a lower bound for the variance of any unbiased estimator, and furthermore show that the MLE attains this lower bound asymptotically.

3.2.2 Proof Rao-Cramer lower bound

Theorem: Rao-Cramer lower bound

Theorem

If $X_1, X_2, \ldots X_n$ are iid with common distribution, and everything is 'nice', then if $\mathbb{E}(T) = k(\theta)$, that is T is an unbiased estimate of some function of θ, then

$$Var(T) \geq \frac{[k'(\theta)]^2}{I(\theta)}.$$

proof: (https://en.wikipedia.org/wiki/Fisher_information)

Informal derivation of the Cramér–Rao bound [edit]

The Cramér–Rao bound states that the inverse of the Fisher information is a lower bound on the variance of any unbiased estimator of θ. H.L. Van Trees (1968) and B. Roy Frieden (2004) provide the following method of deriving the Cramér–Rao bound, a result which describes use of the Fisher information, informally:

Consider an unbiased estimator $\hat{\theta}(X)$. Mathematically, we write

$$\mathrm{E}\left[\hat{\theta}(X) - \theta \middle| \theta\right] = \int \left[\hat{\theta}(x) - \theta\right] \cdot f(x;\theta)\,\mathrm{d}x = 0.$$

The likelihood function $f(X; \theta)$ describes the probability that we observe a given sample x *given* a known value of θ. If f is sharply peaked with respect to changes in θ, it is easy to intuit the "correct" value of θ given the data, and hence the data contains a lot of information about the parameter. If the likelihood f is flat and spread-out, then it would take many, many samples of X to estimate the actual "true" value of θ. Therefore, we would intuit that the data contain much less information about the parameter.

Now, we use the product rule to differentiate the unbiased-ness condition above to get

$$\frac{\partial}{\partial\theta}\int \left[\hat{\theta}(x) - \theta\right] \cdot f(x;\theta)\,\mathrm{d}x = \int \left(\hat{\theta}(x) - \theta\right) \frac{\partial f}{\partial\theta}\,\mathrm{d}x - \int f\,\mathrm{d}x = 0.$$

We now make use of two facts. The first is that the likelihood f is just the probability of the data given the parameter. Since it is a probability, it must be normalized, implying that

$$\int f\,\mathrm{d}x = 1.$$

Second, we know from basic calculus that

$$\frac{\partial f}{\partial\theta} = f\,\frac{\partial \log f}{\partial\theta}.$$

Using these two facts in the above let us write

$$\int \left(\hat{\theta} - \theta\right) f\,\frac{\partial \log f}{\partial\theta}\,\mathrm{d}x = 1.$$

Factoring the integrand gives

$$\int \left(\left(\hat{\theta} - \theta\right)\sqrt{f}\right)\left(\sqrt{f}\,\frac{\partial \log f}{\partial\theta}\right)\,\mathrm{d}x = 1.$$

If we square the equation, the Cauchy–Schwarz inequality lets us write

$$\left[\int \left(\hat{\theta} - \theta\right)^2 f\,\mathrm{d}x\right] \cdot \left[\int \left(\frac{\partial \log f}{\partial\theta}\right)^2 f\,\mathrm{d}x\right] \geq 1.$$

The right-most factor is defined to be the Fisher Information

$$\mathcal{I}(\theta) = \int \left(\frac{\partial \log f}{\partial\theta}\right)^2 f\,\mathrm{d}x.$$

The left-most factor is the expected mean-squared error of the estimator $\hat{\theta}$, since

$$\mathrm{E}\left[\left(\hat{\theta}(X) - \theta\right)^2 \middle| \theta\right] = \int \left(\hat{\theta} - \theta\right)^2 f\,\mathrm{d}x.$$

Notice that the inequality tells us that, fundamentally,

$$\mathrm{Var}\left(\hat{\theta}\right) \geq \frac{1}{\mathcal{I}(\theta)}.$$

In other words, the precision to which we can estimate θ is fundamentally limited by the Fisher Information of the likelihood function.

Theorem

$\mathbb{E}(D_1) = 0$ *and* $\mathbb{E}(D_1^2) = Var(D_1) = -\mathbb{E}(D_2) := I(\theta)$.

proof:

Under certain regularity conditions,[4] it can be shown that the first moment of the score (that is, its expected value) is 0:

$$\mathrm{E}\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\middle|\theta\right] = \mathrm{E}\left[\frac{\frac{\partial}{\partial\theta}f(X;\theta)}{f(X;\theta)}\middle|\theta\right] = \int \frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)} f(x;\theta)\,\mathrm{d}x =$$

$$= \int \frac{\partial}{\partial\theta} f(x;\theta)\,\mathrm{d}x = \frac{\partial}{\partial\theta}\int f(x;\theta)\,\mathrm{d}x = \frac{\partial}{\partial\theta}\,1 = 0.$$

The second moment is called the Fisher information:

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\middle|\theta\right] = \int \left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 f(x;\theta)\,\mathrm{d}x\,,$$

where, for any given value of θ, the expression E[...|θ] denotes the conditional expectation over values for X with respect to the probability function $f(x;\theta)$ given θ. Note that $0 \le \mathcal{I}(\theta) < \infty$. A random variable carrying high Fisher information implies that the absolute value of the score is often high. The Fisher information is not a function of a particular observation, as the random variable X has been averaged out.

Since the expectation of the score is zero, the Fisher information is also the variance of the score.

If $\log f(x;\theta)$ is twice differentiable with respect to θ, and under certain regularity conditions, then the Fisher information may also be written as[5]

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

since

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

and

$$\mathrm{E}\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)}\middle|\theta\right] = \cdots = \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)\,\mathrm{d}x = \frac{\partial^2}{\partial\theta^2}\,1 = 0.$$

Thus, the Fisher information is the negative of the expectation of the second derivative with respect to θ of the natural logarithm of f. Information may be seen to be a measure of the "curvature" of the support curve near the maximum likelihood estimate of θ. A "blunt" support curve (one with a shallow maximum) would have a low negative expected

https://en.wikipedia.org/wiki/Fisher_information

3.2.3 Proof MLE attains this lower bound asymptotically.

Theorem

If $\hat{T}$ denotes the maximum likelihood estimator of θ based on a sample of n observations, and everything is still 'nice', then as $n \to \infty$

$$\hat{T} \xrightarrow{d} \mathrm{N}\left(\theta, \frac{1}{\mathrm{I}(\theta)}\right).$$

(proof: https://en.wikipedia.org/wiki/Maximum_likelihood_estimation)

Then the maximum likelihood estimator has asymptotically normal distribution:

$$\sqrt{n}\left(\hat{\theta}_{\text{mle}} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0,\, I^{-1}).$$

Proof, skipping the technicalities [edit]

Since the log-likelihood function is differentiable, and θ_0 lies in the interior of the parameter set Θ, in the maximum the first-order condition will be satisfied:

$$\nabla_\theta\, \hat{\ell}\,(\hat{\theta} \mid x) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ln f(x_i \mid \hat{\theta}) = 0.$$

When the log-likelihood is twice differentiable, this expression can be expanded into a Taylor series around the point $\theta = \theta_0$:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ln f(x_i \mid \theta_0) + \left[\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta\theta} \ln f(x_i \mid \tilde{\theta})\right](\hat{\theta} - \theta_0),$$

where $\tilde{\theta}$ is some point intermediate between θ_0 and $\hat{\theta}$. From this expression we can derive that

$$\sqrt{n}(\hat{\theta} - \theta_0) = \left[-\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta\theta} \ln f(x_i \mid \tilde{\theta})\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_\theta \ln f(x_i \mid \theta_0)$$

Here the expression in square brackets converges in probability to $H = \mathbb{E}\left[-\nabla_{\theta\theta} \ln f(x|\theta_0)\right]$ by the law of large numbers. The continuous mapping theorem ensures that the inverse of this expression also converges in probability, to H^{-1}. The second sum, by the central limit theorem, converges in distribution to a multivariate normal with mean zero and variance matrix equal to the Fisher information I. Thus, applying Slutsky's theorem to the whole expression, we obtain that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0,\, H^{-1} I H^{-1}\right).$$

Finally, the information equality guarantees that when the model is correctly specified, matrix H will be equal to the Fisher information I, so that the variance expression simplifies to just I^{-1}.

4 sufficient statistics.

There exist statistics that contain the maximal information about a parameter μ. A statistic of our data $X_1, X_2, \dots X_n = \mathbf{X}$, $T(X_1, X_2, \dots X_n)$ is *sufficient* if the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$ does not depend on θ. That is

$$\mathbb{P}(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t, \theta) = \mathbb{P}(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t).$$

4.1 Fisher-Neyman factorisation theorem

Let $X_1, X_2, \dots X_n$ be a random sample from a distribution with pmf/pdf $f(x|\theta)$. A statistic $Y = T(X_1, X_2, \dots X_n)$ is sufficient if and only if there exist two non-negative functions k_1 and k_2 such that

$$f(x_1|\theta)f(x_2|\theta)\dots f(x_n|\theta) = k_1(T(x_1, x_2, \dots x_n)|\theta)k_2(x_1, x_2, \dots x_n).$$

4.2 Rao-Blackwell theorem

Given a random sample which depend upon θ, a sufficient statistic Y_1 for θ and Y_2 an unbiased estimator for θ, then

$$\mathbb{E}(Y_2|Y_1) = \phi(Y_1)$$

defines a statistic that is also unbiased, and most importantly the variance of it is less than or equal to $\text{Var}(Y_2)$, that is

$$\text{Var}(Y_2) \geq \text{Var}(\phi(Y_1)).$$

If I have a sufficient statistic, then for every unbiased estimator, if we take the conditional expectation with respect to with the sufficient statistic, you get a 'better' estimate.

What this means is, that the MVB must be a function of a sufficient statistic.

4.3 Lehmann-Scheffe Theorem

Given a random sample $X_1, X_2, \dots X_n$ and Y is a sufficient statistic for θ, then if there is a function of Y, $\phi(Y)$ that is unbiased, and the family of distributions of Y is complete, then $\phi(Y)$ is the unique estimator that achieves the MVB for θ.

4.4 Exponential family and complete sufficient statistics

Consider a pmf/pdf $f(x|\theta)$ which is non-trivial on $x \in \mathcal{S}$. If there exist functions $p(\theta)$, $K(x)$, $H(x)$ and $q(\theta)$ such that

$$f(x|\theta) = \exp[p(\theta)K(x) + H(x) + q(\theta)] \quad x \in \mathcal{S},$$

where $\mathcal{S}$ does not depend upon θ and $p(\theta)$ is non-trivial, then we say f belongs to the regular exponential family.

If $X_1, X_2, \ldots X_n$ is a random sample from the regular exponential family and Y a sufficient statistic for X, then

$$Y = \sum_{i=1}^{n} K(X_i)$$

is a sufficient statistic for θ and the family of density functions for Y is complete. Hence Y is a complete sufficient statistic for θ.

If we have a sample from a the regular exponential family, then a sufficient statistic is $Y = \sum_{i=1}^{n} K(X_i)$ and

$$\mathbb{E}(Y) = -n \cdot \frac{q'(\theta)}{p'(\theta)}$$
$$\text{Var}(Y) = \frac{n}{p'(\theta)^3}\left[p''(\theta)q'(\theta) - q''(\theta)p'(\theta)\right]$$

4.5 MLEs and sufficient statistics

If a sufficient statistic Y for θ exists, and if a maximum likelihood estimator $\hat{\theta}$ for θ exists and is unique, then $\hat{\theta}$ is a function of Y.

Set E(f(Y)) $=$ θ, f(Y) is MLE

5 Make inference about statistics (inference goal of using statistics)

We can calculate the confidence interval(inference of value) based on the distribution of estimators(statistics).

5.1 Make inference about one statistics

Suppose we have a random sample on $X \overset{d}{=} \mathrm{N}(\mu, \sigma^2)$. Let X^* denote a future observation that is independent of our sample $X_1, X_2, \ldots X_n$.

Now $X^* \overset{d}{=} \mathrm{N}(\mu, \sigma^2)$ and $\bar{X} \overset{d}{=} \mathrm{N}(\mu, \sigma^2/\mathrm{n})$. Therefore

$$X^* - \bar{X} \overset{d}{=} \mathrm{N}(0, \sigma^2(1 + 1/\mathrm{n})).$$

And hence,

$$\mathbb{P}\left(\bar{X} - 1.96\sigma\sqrt{1 + \frac{1}{n}} < X^* < \bar{X} + 1.96\sigma\sqrt{1 + \frac{1}{n}}\right) = 0.95.$$

This is based on:

When we knew σ^2 we would use the fact that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{d}{=} \mathrm{N}(0, 1).$$

When we use the S as the estimator for σ, we have

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \overset{d}{=} t_{n-1},$$

and hence we can use the t-distribution to construct confidence intervals.

5.2 Compare two statistics

Suppose we have independent samples of two normal random variables

n_1 observations on $X_1 \overset{d}{=} \mathrm{N}(\mu_1, \sigma_1^2)$: $\quad X_{11}, X_{12}, \ldots, X_{1n_1}$
n_2 observations on $X_2 \overset{d}{=} \mathrm{N}(\mu_2, \sigma_2^2)$: $\quad X_{21}, X_{22}, \ldots, X_{2n_2}$,

based upon these observations, we wish to make comparisons between the variances σ_1^2 and σ_2^2, and the means μ_1 and μ_2.

5.2.1 Compare two variances

Now recalling that

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \overset{d}{=} \chi^2_{n_1-1}, \quad \text{and} \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \overset{d}{=} \chi^2_{n_2-1}$$

Hence, from the definition of the F-distribution,

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \overset{d}{=} F_{n_1-1,n_2-1},$$

and we can use this result to find confidence intervals for σ_1^2/σ_2^2.

Example:

Given $n_1 = 25, s_1^2 = 1.27, n_2 = 7, s_2^2 = 2.92$, find a two sided 95% confidence interval for σ_1^2/σ_2^2.

Well, we have $\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \overset{d}{=} F_{24,6}$. Hence we can find

$$\mathbb{P}\left(0.33 < \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} < 5.11\right) = 0.95$$

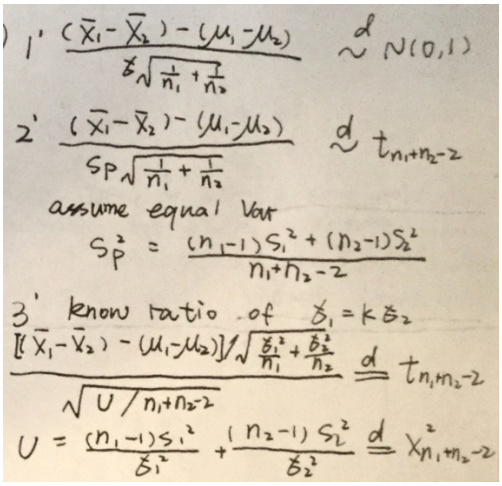
And therefore,

$$\mathbb{P}\left(\frac{1}{5.11}\frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{0.33}\frac{S_1^2}{S_2^2}\right) = 0.95$$

5.2.2 compare two means

The first thing to note is

$$\bar{X}_1 - \bar{X}_2 \overset{d}{=} \mathrm{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{\mathrm{n}_1} + \frac{\sigma_2^2}{\mathrm{n}_2}\right)$$



6 hypothesis testing (Decision making goal of using statistics)

6.1 definition of hypothesis

A statistical hypothesis is usually a statement about the distribution of a random variable X. We tend to assume that the distribution of X is specified except for a particular parameter. The hypothesis under test is called the null hypothesis, and we denote this with H_0. It usually represents some kind of known standard, what we expect the results should be.

We will usually assume that the null hypothesis always takes the form

$$H_0 : \theta = \theta_0$$

Alternative hypothesis:

$$H_1 : \theta = \theta_1$$

$$\text{Or, } H_1 : \theta \neq \theta_0$$

Reject H_0 using P-values.

$$P = \mathbb{P}(\text{test statistic is as extreme as the value obtained} | H_0),$$

where extreme is defined by the alternative hypothesis.

The test procedure is then to reject H_0 if $P < \alpha$.

The P-value tells you how `likely' the observed data is given the null hypothesis.

6.1.1 P-value can be awful:

People, particularly people who are bad at statistics seem to think of P-values as universal cure alls. If you just quote the P-value, then that's it, question answered.

The main reason why P-values are awful is that confidence intervals are better in every single way.

Furthermore, just because we do not reject H_0 does not necessarily mean that H_0 is the right thing to base our decisions on.

	accept H_0	reject H_0
H_0 true	correct decision probability $1-\alpha$	error of type I with probability α aka *size*
H_1 true	error of type II probability β	correct decision probability $1-\beta$ aka power

We would typically like the error probabilities α (false negative) and β (false possitive) to be small, though however typically we focus on making sure α is small, β being small is of lesser importance. This is due to the special importance of H_0.

6.2 Deriving tests

The method used to derive the best test is called the likelihood ratio criterion.

Theorem

The likelihood ratio test is the best test, that is given all tests with the same size, the likelihood ratio test has the greatest power.

6.2.1 Simple H_0 V.S simple H_1

Very easy, the likelihood ratio is fixed given two parameters.

Suppose $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$.

Define

$$L_1 = \text{likelihood function under } H_1 = L(\theta_1),$$
$$L_0 = \text{likelihood function under } H_0 = L(\theta_0).$$

The likelihood ratio test is then given by,

$$\text{Test: reject } H_0 \text{ if } \frac{L_1}{L_0} > k.$$

6.2.2 Composite H_0 V.S composite H_1

Suppose $H_0 : \theta \in A_0$ and $H_1 : \theta \in A_1$. Unsurprisingly this likelihood test ratio takes the form:

$$\text{reject } H_0 \text{ if } \frac{L_1}{L_0} > k$$

where $L_1 = \max_{\theta \in A_1} L(\theta)$ and $L_0 = \max_{\theta \in A_0} L(\theta)$. These sorts of hypotheses typically occur in the case where θ is a vector. There is one case of particular interest.

6.2.3 Simple H_0 V.S composite H_1: $\theta \neq \theta_0$

We choose to test null hypothesis against the most plausible alternative, that is we take θ_1 to be value that maximises $L(\theta)$ in the set of possible values (A_1) defined by the alternate hypothesis. Thus we define:

$$L_1 = \max_{\theta \in A_1} L(\theta),$$

and the likelihood ratio test is

$$\text{Reject } H_0 \text{ if } \frac{L_1}{L_0} > k.$$

Note that typically such a test simply becomes comparing against the maximum likelihood estimate.

Example:

Consider a random sample of n observations from $\mathrm{N}(\theta, 1)$ and we wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Find the exact distribution of the log-likelihood ratio.

In this case,

$$\frac{L_1}{L_0} = \exp\left(\frac{1}{2} n(\bar{x} - \theta_0)^2\right).$$

And hence

$$\log \frac{L_1}{L_2} = \frac{1}{2} n(\bar{X} - \theta_0)^2 = \frac{1}{2}\left(\frac{\bar{X} - \theta_0}{1/\sqrt{n}}\right)^2 \overset{d}{=} \frac{1}{2}\chi_1^2.$$

if the variance is not given, the test statistic is t-distribution. The likelihood ratio test corresponds exactly to the standard t-test. the test statistic has a distribution, and the p-value take place of k.

we have always needed to assume something about the underlying distribution.

6.3 Distribution-free methods (don't need to find the distribution of MLE)

Good compared to previous methods because we don't need to assume anything about underlying distributions.

These tests tend to have less statistical power though.

Usually much more robust to outliers.

6.3.1 χ^2 test.

6.3.1.1 Test a claim(expect)

Example:

Suppose we then observed the following data over 100 days.

x	0	1	2	3	4	≥ 5
$\mathbb{P}(X = x) = p_x$	0.1	0.3	0.2	0.2	0.1	0.1
obs. freq. f_x	0	11	25	22	28	14

How does Jeff's original hypothesis hold up?

The test statistic we use to assess the goodness of fit is

$$U = \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}} = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}.$$

To decide whether or not the original hypothesis was reasonable, we are simply testing if U is too large or not.

As long as n is large-ish, then what is the approximate distribution of $\frac{(f_i - np_i)}{\sqrt{np_i q_i}}$?

The central limit theorem says this is approximately normal.

If we do a little rescaling, some 'magic', it can be shown that as $n \to \infty$,

$$U = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} \xrightarrow{d} \chi^2_{k-1}.$$

A χ^2 test is a one sided test. Why?

Implicit in the χ^2 test is the fact that we are approximating a binomial with a normal, hence we normally stipulate that $np_i > 5$ as a rule of thumb.

If this is not satisfied, we often combine classes until it is satisfied.

If U is *too* small, then the first is probably *too good* and may suggest that the experiment was rigged.

We have lost a 'degree of freedom' because we have added a constraint that the sample mean must be equal to the fitted mean.

In general in fitting a distribution in this manner,

$$U \stackrel{d}{=} \chi^2_{k-p-1},$$

where $k =$ the number of classes and $p =$ the number of parameters estimated.

6.3.1.2 Independence Tests

If we set H_0 to be that the classifications are independent, this is equivalent to

$$H_0 : \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B),$$
$$H_1 : \mathbb{P}(A \cap B) \neq \mathbb{P}(A)\mathbb{P}(B).$$

If H_0 is true then the expected frequencies are given by

exp freq	B	B'	
A	np_Ap_B	$np_Ap_{B'}$	np_A
A'	$np_{A'}p_B$	$np_{A'}p_{B'}$	$np_{A'}$
	np_B	$np_{B'}$	n

Under H_0 we can then calculate expected frequencies, if we make estimates for p_A and p_B.
The χ^2 test generalises to a r x c contingency table in exactly the same way as you would expect.

Given that we only need to estimate $r-1$ of the p_{A_i}s and $c-1$ of the p_{B_j}s, then the degrees of freedom in the χ^2 would be
$rc - ((r-1)+(c-1)) - 1 = (r-1)(c-1)$.

6.3.1.3 Equality of medians

Test $H_0 : m_X = m_Y$ against $H_1 : m_X \neq m_Y$ for the following:

```
X: 93 98 103 111 102 112 92 90 106 103

Y: 89 103 118 96 86 84 99 107 106 101
```

The overall median is $m^* = 101.5$.

	$\leq m^*$	$> m^*$	
X-sample			n_1
Y-sample			n_2
			N

If we can calculate the expected number of observations in each cell, then we can apply a χ^2 test.

Under H_0, we expect each cell to have the same number of observations in it.

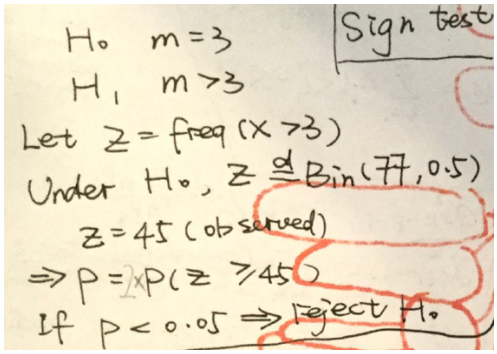
6.3.2 Sign test

6.3.2.1 Test an inference on the median

To test the hypothesis $H_0 : m = m_0$, we can use the test statistic $Z = \text{freq}(X \leq m_0)$.

Under H_0 the distribution of Z should be $\mathrm{Bi}(\mathrm{n}, 0.5)$.

This is often called the sign test because we are considering the 'sign' of the $x_i - m_0$.



Test whether within the 95% confidence interval.

It's often troublesome to work with the binomial distribution, so recall that we can always approximate the distribution of Z with a normal distribution if n is large enough,

$$Z \overset{d}{\approx} \mathrm{N}\left(\frac{\mathrm{n}}{2}, \frac{\mathrm{n}}{4}\right).$$

6.3.3 Wilcoxon Rank test

6.3.3.1 Test an inference on the median

If we are willing to assume that the distribution of X is symmetrical, then there exists a more powerful test for $H_0 : m = m_0$ based upon the 'ranks' of our observations.

We first transform our data X by considering the sample $Y_i = |X_i - m_0|$, then assign ranks to these Y_i by ordering them from smallest to largest.

Our plan is to then compare the ranks of the observations smaller than the hypothesised median m_0 with the ranks of the observations that are larger than m_0.
If we define T to be the difference of the sum of ranks above and the sum of the ranks below the hypothesised median, then it can be shown that assuming H_0 is true,

$$\mathbb{E}(T) = 0, \quad \text{Var}(T) = \frac{n(n+1)(2n+1)}{6},$$

and hence when n is large we can use a normal approximation.

It should be noted that the exact distribution for the Wilcoxon rank test is known, however we will tend to stick with normal approximations in this course.

6.3.3.2 comparative inference

We can also extend the rank test for two sample testing.

Given two random samples $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$, if $N = n_1 + n_2$ we wish to test if there is a difference between the two groups.

We simply order the data, and calculate the sum of the ranks from either group.
Example:
Our buddy Jeff the alcoholic wishes to test if his alcoholism affects his ability to play Starcraft II. Jeff (unwisely) decides to measure his ability by the number of actions per minute (APM) he can input. He plays 10 games sober (X), then plays 10 games drunk (Y) and measures his average APM per game.

```
X: 50 43 48 56 54 40 44 47 45 51

Y: 41 35 42 45 44 27 38 33 37 42
```

Calculate the ranks and see if there is a difference between his ability when sober and drunk.

If H_0 is true, then W_X (the sum of the ranks from X) should be the sum of 10 random integers between 1 and 20. So in theory we can calculate an exact P value, but considering how many combinations give us a value for the sum less than W_X.

This is a lot of effort, so we can also use a normal approximation. If n_1 and n_2 are large enough, then

$$W_1 \overset{d}{\approx} \mathrm{N}\left(\frac{1}{2}\mathrm{n}_1(\mathrm{n}_1+\mathrm{n}_2+1), \frac{1}{12}\mathrm{n}_1\mathrm{n}_2(\mathrm{n}_1+\mathrm{n}_2+1)\right).$$

7 Multiple comparisons

We consider n independent observations on k normally distributed random variances that have equal variances. There are thus $N = nk$ observations in total.

$$\begin{aligned}
X_1 &\overset{d}{=} \mathrm{N}(\mu_1, \sigma^2) \quad \text{sample: } X_{11}, X_{12}, \ldots, X_{1n}, && \bar{X}_1, S_1^2 \\
X_2 &\overset{d}{=} \mathrm{N}(\mu_2, \sigma^2) \quad \text{sample: } X_{21}, X_{22}, \ldots, X_{2n}, && \bar{X}_2, S_2^2 \\
&\ldots \\
X_k &\overset{d}{=} \mathrm{N}(\mu_\mathrm{k}, \sigma^2) \quad \text{sample: } X_{k1}, X_{k2}, \ldots, X_{kn}, && \bar{X}_k, S_k^2
\end{aligned}$$

The initial hypothesis test of

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

is merely the first step in the analysis of such data.

For starters we are usually interested in the individual parameters, and wish to be able to do inference on those.

We thus have

$$T = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{\bar{X}})^2 = (N-1)S_T^2 \overset{d}{=} \sigma^2\chi^2_{N-1} \quad \text{if } H_0 \text{ true}$$

$$W = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k}(n-1)S_i^2 \overset{d}{=} \sigma^2\chi^2_{N-k} \quad \text{always}$$

$$B = \sum_{i=1}^{k} n(\bar{X}_i - \bar{\bar{X}})^2 = n(k-1)S_B^2 \overset{d}{=} \sigma^2\chi^2_{k-1} \quad \text{if } H_0 \text{ true}$$

We call T the total sum of squares, or total SS, B the between groups SS, and W the within groups SS.

And those names call it exactly how it is.

Most importantly,

$$T = W + B.$$

The LR test said we should base our test upon

$$\frac{W}{T}.$$

However instead this is equivalent to testing

$$\frac{\frac{B}{k-1}}{\frac{W}{N-k}}.$$

which we know follows distribution $F_{k-1,N-k}$.

This approach is called Analysis of Variance, or ANOVA for short.

	df	SS	MS	F
between				
within				
Total				

compare to P-value.

We formulated our one way ANOVA using the likelihood ratio test.

To do this we assumed that our observations came from the normal distribution with common variance, but with means depending on which group they came from.

This implicitly suggests a model has been fit.

7.1 The one-way ANOVA model

The model interpretation is as follows:

$$y_{ij} = \mu_j + \epsilon_i,$$

where j indicates which group observation i belongs to, and the ϵ_i are i.i.d. $\mathrm{N}(0, \sigma^2)$.

The one-way ANOVA is used when we fit a different mean for each different level of a categorical factor. ANOVA test the hypothesis that for each factor(group), different elements(group 1,2,3) have the same mean value. The one-way ANOVA model test only one factor. We can therefore similarly formulate a hypothesis test for whether the

factor should be included in our model or not. (if all mean all the same for factor a, then no need to consider different a during linear regression)

e.x.

Suppose I had the following data:

Group I	Group 2	Group 3
6	10	9
2	9	12
4	11	13
3	12	9

(a) * Produce the ANOVA for this data.

(b) Calculate confidence intervals for $\mu_1 - \mu_2$, $\mu_2 - \mu_3$ and $\mu_3 - \mu_1$ each with an individual 95% confidence level.

(c) Calculate the same confidence intervals using Bonferroni's method.

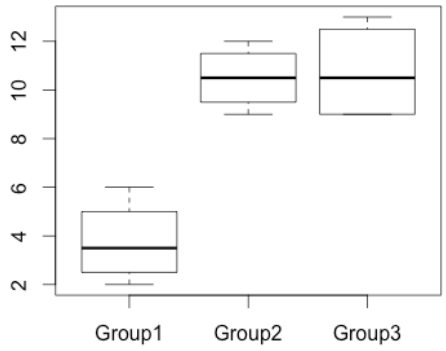
(d) * Produce the same confidence intervals using Tukey's method.

```
##one way anova
ex4 = c(6,2,4,3,10,9,11,12,9,12,13,9);
group =c(rep("Group1",4), rep("Group2",4), rep("Group3",4));
state_group = data.frame(ex4,group);
boxplot(ex4~group)
anova_4 = aov(ex4~group, data = state_group);
summary(anova_4)
TukeyHSD(anova_4)
```

```
> summary(anova_4)
            Df Sum Sq Mean Sq F value   Pr(>F)
Group(consider as a factor)       2  126.2   63.08   21.43 0.000378 *** (belief mean of different group are equal) p value low means we refuse the hypo means of different means are equal.
Residuals    9   26.5    2.94
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(anova_4)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = ex4 ~ group, data = state_group)

$group
                diff       lwr      upr     p adj
Group2-Group1 6.75  3.362315 10.137685 0.0009141
Group3-Group1 7.00  3.612315 10.387685 0.0007053
Group3-Group2 0.25 -3.137685  3.637685 0.9769280
```



7.2 Additive models

The additive model assumes that each level of a factor (row and column) simply adds something to it's expected value. We can add more factor to the model(factor 1: column=group number, factor 2: row = position in each group).

Let μ be the overall average of all the data. And let $\alpha_i = \mu_i - \mu$ (row means) and $\beta_j = \mu_j - \mu$ (column means). Then

$$y_{i,j} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

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

Does the 'row factor' have any effect?

Does the 'column factor' have any effect?

$$T = R + C + W,$$

that is, total SS is equal to the row SS + column SS + within SS.

Similarly to before, we can derive the most powerful test using the likelihood ratio test.

```
##two way anova
drunk = c(3.1, 4.2, 2.7, 4.9, 2.9, 4.9, 3.2, 4.5, 2.7, 2.9, 1.8, 3.0,
        2.9, 2.3, 2.4, 3.7, 4.0, 4.6, 3.0, 3.9, 4.4, 5.0, 2.5, 4.2);
rank =c(rep("R1",8), rep("R2",8), rep("R3",8));
rank1 =c(rep(c("C1","C2","C3","C4"),6));
boxplot(drunk~rank+rank1)
statedrunk = data.frame(drunk,rank,rank1);
result = aov(drunk~rank+rank1, data = statedrunk);
summary(result);
#TukeyHSD confidence interval
TukeyHSD(result);
```

```
> summary(result);
            Df Sum Sq Mean Sq F value   Pr(>F)
rank         2  7.298   3.649   13.58 0.000254 ***(reject H0 that rank mean are
equal for all rank)
rank1        3  8.131   2.710   10.09 0.000401 ******(reject H0 that rank1 mean
are equal for all rank1)

Residuals   18  4.838   0.269
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #TukeyHSD confidence interval
> TukeyHSD(result);
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = drunk ~ rank + rank1, data = statedrunk)

$rank
        diff       lwr       upr     p adj
R2-R1 -1.0875 -1.7490348 -0.4259652 0.0015002
R3-R1  0.1500 -0.5115348  0.8115348 0.8330813
R3-R2  1.2375  0.5759652  1.8990348 0.0004249

$rank1
           diff       lwr       upr     p adj
C2-C1  0.6500000 -0.1959220  1.4959220 0.1690003
C3-C1 -0.7333333 -1.5792554  0.1125887 0.1029726
C4-C1  0.7000000 -0.1459220  1.5459220 0.1261225
C3-C2 -1.3833333 -2.2292554 -0.5374113 0.0011086
C4-C2  0.0500000 -0.7959220  0.8959220 0.9982780
C4-C3  1.4333333  0.5874113  2.2792554 0.0007742
```

With the additive model we assumed that the main effects have zero impact on each other. However, this is often not the case. Sometimes the row and column factors may have some sort of multiplicative effect. Again, just like before we can decompose the total SS into components such that

$$T = R + C + I + W,$$

that is total SS = row SS + column SS + interaction SS + within SS.

We can therefore similarly formulate a hypothesis test for whether the interaction term should be included in our model or not.

As expected it is just a F-test where we compare mean SS for the appropriate columns in the ANOVA table.

```
> result2 = aov(drunk~rank*rank1, data = statedrunk);
> summary(result2);
              Df Sum Sq Mean Sq F value   Pr(>F)
rank           2  7.298   3.649  30.726 1.90e-05 ***
rank1          3  8.131   2.710  22.825 3.01e-05 ***
rank:rank1     6  3.413   0.569   4.789   0.0102 *
Residuals     12  1.425   0.119
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8 Estimation(regression)

We want to estimate what the value of Y would be given different values of X(X here is a factor, means are equal is beta =0), and we would seek to fit some sort of model based upon observed data (x_i , y_i).

So what we are essentially trying to estimate is

$$\mathbb{E}(Y|x) = \mu(x).$$

8.1 Linear regression

The simplest function that we can reasonably consider is a linear function That is we assume that

$$\mathbb{E}(Y|x) = \alpha + \beta x.$$

We also assume that $\text{Var}(Y_i) = \sigma^2$ and the Y_i are independent. What is the distribution of Y_i then?

8.1.1 The method of least squares

So given our function $\mu(x) = \mathbb{E}(Y|x) = \alpha + \beta x$, we wish to choose the 'best' α and β out of all possible values of α and β to minimise

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

Solutions:

Lets set $u_i = x_i - \bar{x}$.

First thing to note is that $\sum_{i=1}^{n} u_i = 0$. Now

$$E(Y_i) = \alpha + \beta x_i = \alpha + \beta\bar{x} + \beta(x_i - \bar{x}) = \alpha + \beta\bar{x} + \beta u_i = \alpha_0 + \beta u_i,$$

where $\alpha_0 = \alpha + \beta\bar{x}$.

$$\begin{aligned}
\hat{\alpha_0} &= \bar{y}, \\
\hat{\beta} &= \frac{\sum_{i=1}^{n} u_i y_i}{\sum_{i=1}^{n} u_i^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\
\hat{\alpha} &= \bar{y} - \beta\bar{x}.
\end{aligned}$$

If we are assuming the distribution of the Y_i are normal, then what are the distributions of $\hat{A}_0$ and $\hat{B}$?

Hence,

$$\hat{A}_0 \overset{d}{=} \mathrm{N}\left(\alpha_0, \frac{\sigma^2}{\mathrm{n}}\right) \quad \text{and} \quad \hat{\mathrm{B}} \overset{\mathrm{d}}{=} \mathrm{N}\left(\beta, \frac{\sigma^2}{\sum \mathrm{u}^2}\right).$$

8.1.2 Properties of the estimators

Let $\hat{A}_0$ be the estimator for α_0 and $\hat{B}$ for β,

$$\begin{aligned}
\mathbb{E}(\hat{A}_0) &= \alpha_0 \\
\mathrm{Var}(\hat{A}_0) &= \frac{\sigma^2}{n} \\
\mathbb{E}(\hat{B}) &= \beta \\
\mathrm{Var}(\hat{B}) &= \frac{\sigma^2}{\sum_{i=1}^{n} u_i^2} \\
\mathrm{Cov}(\hat{A}_0, \hat{B}) &= 0.
\end{aligned}$$

If we set $\hat{A}$ to be the estimator for α, then

$$\begin{aligned}
\mathbb{E}(\hat{A}) &= \alpha \\
\mathrm{Var}(\hat{A}) &= \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} u_i^2}\right)\sigma^2 \\
\mathrm{Cov}(\hat{A}, \hat{B}) &= -\frac{\bar{x}}{\sum_{i=1}^{n} u_i^2}\sigma^2
\end{aligned}$$

8.1.3 Estimation of variance.

We define the residual sum of squares as

$$d^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha}_0 - \hat{\beta} u_i)^2,$$

$$D^2 = \sum_{i=1}^{n} (Y_i - \hat{A}_0 - \hat{B} u_i)^2.$$

We note that

$$\sum_{i=1}^{n} (y_i - \alpha_0 - \beta u_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha}_0 - \hat{\beta} u_i)^2 + n(\hat{\alpha}_0 - \alpha_0)^2 + (\hat{\beta} - \beta)^2 \sum_{i=1}^{n} u_i^2,$$

which is proved by showing that the cross terms in

$$\sum_{i=1}^{n} \left[(y_i - \hat{\alpha}_0 - \hat{\beta} u_i) + (\hat{\alpha}_0 - \alpha_0) + (\hat{\beta} - \beta) u_i \right]^2,$$

are zero. Hence

$$\sum_{i=1}^{n} (Y_i - \alpha_0 - \beta u_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{A}_0 - \hat{B} u_i)^2 + n(\hat{A}_0 - \alpha_0)^2 + (\hat{B} - \beta)^2 \sum_{i=1}^{n} u_i^2.$$

Hence by taking expectations of both sides we get

$$n\sigma^2 = \mathbb{E}(D^2) + \sigma^2 + \sigma^2.$$

Therefore to estimate σ^2 we would use

$$S^2 = \frac{D^2}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{A}_0 - \hat{B} u_i)^2,$$

and this is an unbiased estimator for σ^2.

Hence $\frac{(n-2)S^2}{\sigma^2} \stackrel{d}{=} \chi^2_{n-2}$.

Now since $\hat{A}_0$ and $\hat{B}$ are independent of S^2, we have

$$\frac{\hat{A}_0 - \hat{\alpha}_0}{S/\sqrt{n}} \stackrel{d}{=} t_{n-2}, \qquad \frac{\hat{B} - \beta}{S/\sqrt{\sum u^2}} \stackrel{d}{=} t_{n-2}.$$

8.1.4 ANOVA approach

To test the hypothesis that factor y are means equal. (test beta = 0, Y are independent of X)

Recall that

$$\sum_{i=1}^{n}(y_i-\alpha_0-\beta u_i)^2=\sum_{i=1}^{n}(y_i-\hat{\alpha}_0-\hat{\beta}u_i)^2+n(\hat{\alpha}_0-\alpha_0)^2+(\hat{\beta}-\beta)^2\sum_{i=1}^{n}u_i^2.$$

This can be rewritten as

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2+d^2.$$

We then end up with an ANOVA of the form

	df	SS
regression	1	$\hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2$
residual	$n-2$	d^2
total	$n-1$	$\sum_{i=1}^{n}(y_i-\bar{y})^2$

From here we can apply a F-test, which turns out to be equivalent to testing for $\beta=0$.

The *sample correlation coefficient* is defined as

$$R=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}.$$

Furthermore we had

$$F=\frac{\text{regression } MS}{\text{residual } MS}=\frac{R^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{D^2/(n-2)}=\frac{(n-2)R^2}{1-R^2}\overset{d}{=}F_{1,n-2}.$$

Note that $t_{n-2}^2=F_{1,n-2}$.

So we can base our test statistic upon $\frac{R\sqrt{(n-2)}}{\sqrt{1-R^2}}$.

e.x

```
##linear regression
x = c(54,47, 69, 87, 65, 73, 83, 81, 72, 74);
y = c(61, 22, 55, 78, 45, 75, 56, 66, 59, 70);
linear_r1 = lm(y~x)
linear_r2 = lm(x~y)
plot(x,y)
summary(linear_r1)
anova(linear_r1)
cor(y,x)^2*8/(1-cor(y,x)*cor(y,x))#calculate F value using corelation
summary(linear_r2)
```

```
> summary(linear_r1)
```

```
Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q  Median     3Q    Max
-14.591  -7.023  -1.700   6.879  17.996

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.3658    22.2580  -0.376   0.7168
x             0.9513     0.3113   3.056   0.0157 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.72 on 8 degrees of freedom
Multiple R-squared:  0.5386,      Adjusted R-squared:  0.4809
F-statistic: 9.338 on 1 and 8 DF,  p-value: 0.01568

> anova(linear_r1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1 1281.9 1281.86  9.3376 0.01568 (X here refer to beta*X, is means of X is equal, beta = 0)
Residuals  8 1098.2  137.28
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> cor(y,x)^2*8/(1-cor(y,x)*cor(y,x))#calculate F value using corelation
[1] 9.337576
```

8.2 The General Linear model

The general linear model is given by

$$\underset{\sim}{Y} = A\underset{\sim}{\theta} + \underset{\sim}{E}$$

where

$\underset{\sim}{Y}$ = n x 1 vector of random observations
A = n x p matrix of known constants
$\underset{\sim}{\theta}$ = p x 1 vector of unknown parameters
$\underset{\sim}{E}$ = n x 1 vector of random errors.

It is assumed that

$$\mathbb{E}(\underset{\sim}{E}) = \underset{\sim}{0}, \quad \mathbb{D}(\underset{\sim}{E}) = \sigma^2 I$$

and

$$\underset{\sim}{Y} \overset{d}{=} \mathsf{N}_n(A\underset{\sim}{\theta}, \sigma^2 I),$$

where N_n means multivariate normal.

So the model can be interpreted as

$$y_i = A_i\underset{\sim}{\theta} + \epsilon_i,$$

or

$$\text{response} = \text{deterministic function} + \text{random error}$$

where the deterministic function must be a *linear function* of the parameters $\underset{\sim}{\theta}$.

Our aim is to estimate $\underset{\sim}{\theta}$, and to achieve this we use the method of least squares.

That is we choose the vector $\underset{\sim}{t}$ that minimises

$$(\underset{\sim}{y} - A\underset{\sim}{t})^T(y - A\underset{\sim}{t}),$$

and we denote our LS estimate with $\hat{\underset{\sim}{\theta}}$. It can be shown (and will be shown) that $\hat{\underset{\sim}{\theta}}$ satisfies

$$A^T A\hat{\underset{\sim}{\theta}} = A^T\underset{\sim}{y}$$

Solution of the general linear model:

$\underset{\sim}{y} = A\underset{\sim}{\theta} + \underset{\sim}{e}.$

LS estimates = ML estimates

$A^T A\hat{\underset{\sim}{\theta}} = A^T \underset{\sim}{y}$

$\hat{\underset{\sim}{\Theta}} \overset{d}{=} N_p(\underset{\sim}{\theta}, \sigma^2(A^T A)^{-1})$

$\hat{\underset{\sim}{E}}$ and $\hat{\underset{\sim}{\Theta}}$ are independent

$\frac{D^2}{\sigma^2} = \frac{(n-p)S^2}{\sigma^2} \overset{d}{=} \chi^2_{n-p}$

$s^2 = \frac{1}{n-p}(\underset{\sim}{y}^T \underset{\sim}{y} - \hat{\underset{\sim}{\theta}}^T A^T \underset{\sim}{y})$

E.X

2. Suppose that the independent normally distributed random variables Y_1, Y_2, Y_3, Y_4 have means given by $\mathbb{E}(Y_1) = \alpha+2\beta, \mathbb{E}(Y_2) = 3\beta, \mathbb{E}(Y_3) = \alpha-\beta, \mathbb{E}(Y_4) = 2\alpha+\beta$, and equal variances, denoted by σ^2. The following observations are made:

$$y_1 = 4, y_2 = 4, y_3 = 5, y_4 = 3.$$

(a) Express this in the form $\underset{\sim}{y} = A\underset{\sim}{\theta}+\underset{\sim}{e}$, and hence estimate α, β and σ^2 using the method of least squares.

(b) Find a 95% confidence interval for $\mathbb{E}(Y_1)$.

$$(A^TA)^{-1} = \frac{1}{65-33}\begin{pmatrix} 13 & -3 \\ -3 & 6 \end{pmatrix}$$

$$= \frac{1}{27}\begin{pmatrix} 5 & -1 \\ -1 & 2 \end{pmatrix}$$

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (A^TA)^{-1}A^Ty$$

$$= \frac{1}{27}\begin{pmatrix} 5 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0 & 1 & 2 \\ 2 & 3 & -1 & 1 \end{pmatrix}\begin{pmatrix} 4 \\ 4 \\ 5 \\ 3 \end{pmatrix}$$

$$= \frac{1}{27}\begin{pmatrix} 5 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 15 \\ 18 \end{pmatrix}$$

$$= \frac{1}{27}\begin{pmatrix} 57 \\ 21 \end{pmatrix} \approx \begin{pmatrix} 2.11 \\ 0.78 \end{pmatrix}$$

2a $\begin{pmatrix} 4 \\ 4 \\ 5 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 0 & 3 \\ 1 & -1 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \varepsilon$

$A^TA = \begin{pmatrix} 1 & 0 & 1 & 2 \\ 2 & 3 & -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 0 & 3 \\ 1 & -1 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 3 \\ 3 & 15 \end{pmatrix}$

$$s^2 = \frac{1}{n-p}(y^Ty - \hat{\theta}^TA^Ty)$$

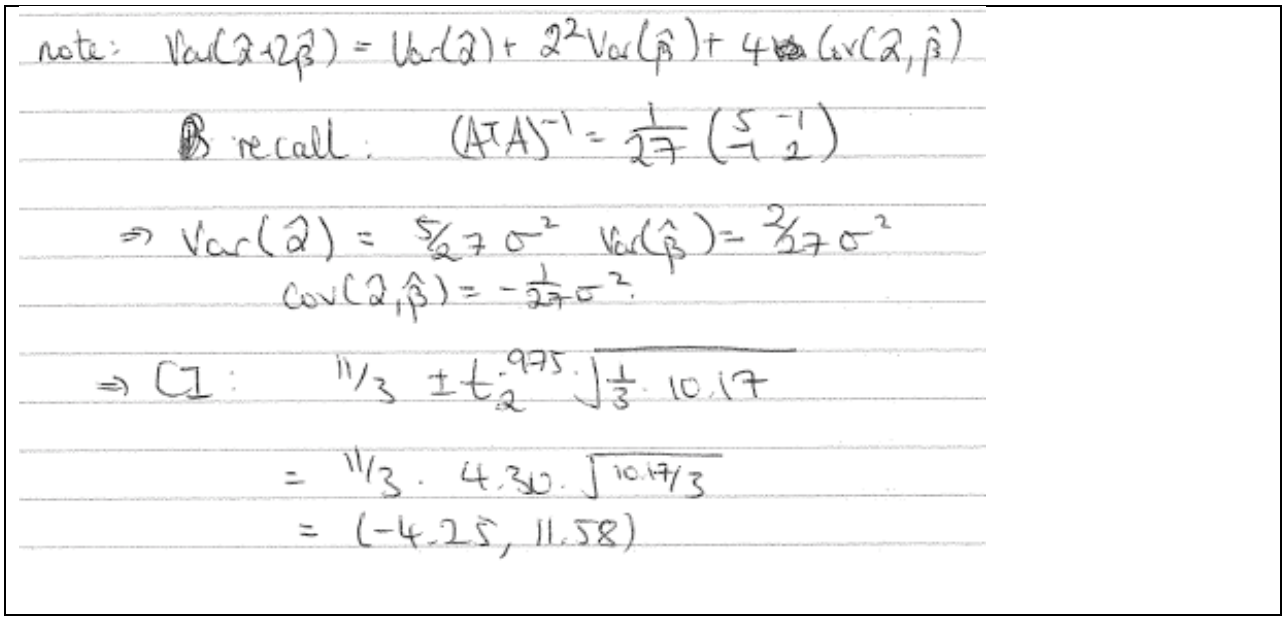
$$= \frac{1}{2}\left(66 - \begin{pmatrix} 19 \\ 7 \end{pmatrix}^T\begin{pmatrix} 15 \\ 18 \end{pmatrix}\right)$$

$$= \frac{1}{2}\cdot(66 - 45.67)$$

$$= 10.17$$

b $E(Y_1) = \alpha+2\beta$

$\hat{\alpha}+2\hat{\beta} = \frac{57+42}{27} = 11/3$



9 Model comparison

9.1 compare linear regression with one way anova

```
##compare linear regression with one way anova
x = c (3, 3, 3, 5, 5, 5, 8, 8, 8, 10, 10, 10);
y = c(4, 6, 2, 9, 12, 11, 13, 14, 14, 18, 16,15);
model.linear = lm(y~x);
model.anova = lm(y~factor(x));
summary(model.linear);
summary(model.anova);
anova(model.linear,model.anova)
```

```
> summary(model.linear);

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q  Median     3Q    Max
-3.4138 -1.0431  0.3333  0.7098  3.2989

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4828     1.4435   0.334    0.745
x             1.6437     0.2052   8.011 1.16e-05 ***(beta not 0)
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.914 on 10 degrees of freedom
Multiple R-squared:  0.8652,      Adjusted R-squared:  0.8517
F-statistic: 64.18 on 1 and 10 DF,  p-value: 1.163e-05

> summary(model.anova);

Call:
lm(formula = y ~ factor(x))

Residuals:
   Min     1Q  Median     3Q    Max
-2.0000 -0.8333  0.1667  0.5833  2.0000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.000      0.866   4.619 0.001713 ** (factor 3 value not equal)
factor(x)5     6.667      1.225   5.443 0.000614 ***
factor(x)8     9.667      1.225   7.893 4.81e-05 ***
factor(x)10   12.333       1.225  10.070 8.06e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.5 on 8 degrees of freedom
Multiple R-squared:  0.9337,      Adjusted R-squared:  0.9089
F-statistic: 37.58 on 3 and 8 DF,  p-value: 4.615e-05

> anova(model.linear,model.anova)
Analysis of Variance Table

Model 1: y ~ x(H0: model 1 is good)
Model 2: y ~ factor(x) (H1: model 2 is good)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     10 36.621
2      8 18.000  2    18.621 4.1379 0.05837 .(Accept H0)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9.2 compare linear regression with quadratic regression

```
##compare linear regression with quadratic regression
x = c(2,6,7,7,14,15,17,17,18,18);
y = c(52, 127, -61, -100, 24, 63, 110, 464, 196, 82);
```

```
model.linear = lm(y~x);
model.quadratic = lm(y~x+I(x^2));
summary(model.linear);
summary(model.quadratic)
anova(model.linear,model.quadratic)
```

```
> summary(model.linear);

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q  Median     3Q    Max
-128.91  -90.70  -60.27   72.18  304.13

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -62.754    106.466  -0.589    0.572
x             13.095      7.969   1.643    0.139

Residual standard error: 142.7 on 8 degrees of freedom
Multiple R-squared:  0.2524,      Adjusted R-squared:  0.1589
F-statistic: 2.701 on 1 and 8 DF,  p-value: 0.1389

> summary(model.quadratic)

Call:
lm(formula = y ~ x + I(x^2))

Residuals:
   Min     1Q  Median     3Q    Max
-134.81  -56.86  -36.18  -16.80  292.54

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  135.235    205.873   0.657    0.532
x            -38.693     47.003  -0.823    0.438
I(x^2)         2.401      2.149   1.117    0.301

Residual standard error: 140.6 on 7 degrees of freedom
Multiple R-squared:  0.3656,      Adjusted R-squared:  0.1843
F-statistic: 2.017 on 2 and 7 DF,  p-value: 0.2034
```

```
> anova(model.linear,model.quadratic)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x + I(x^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      8 163019
2      7 138340  1     24679 1.2488 0.3007(model 1 is good)
```