Daniel **Fitz**
(43961229)

University of Queensland

STAT2203 – Probability and Statistics for Engineering

STAT2203 Lecture Notes

Table of Contents

List of Tables

List of Figures

Sums and Extremes of Independent Random Variables

Multivariate Normal Distribution

- Jointly Gaussian Random Variables; as affine transform of vector of independent standard normals, examples
- Expectation of Vector- and Matrix-valued RVs; application to multivariate normal
- Affine combinations of independent normals; result, examples

Affine Combinations of Normals Example

Exercise: Let $X_1, ..., X_n \sim N(u,o^2)$ represent repeated measurements. Find is the distribution of the average measurement

$$Y = \frac{X_1 \cdots + X_n}{n}$$

Sums of Independent Random Variables

Law of Large Numbers and the **Central Limit Theorem**. Both theorems deal with **Sums of Independent Random Variables**. They arise for example in the following situations:

1) We flip a (biased) coin infinitely many times. Let $X_i = 1$ if the ith flip is "heads" and $X_i = 0$ otherwise. In general we do not know $p = P(X_i = 1)$. However, using the outcomes $x_1, ..., x_n$, we could estimate p by $(x_1 + ... + x_n)/n$

2) A certain machine needs to work continuously. The machine has one component that is very unreliable. This component is replaced immediately upon failure. Suppose there are n such (spare) components. If we denote the component lifetimes by $X_1, ..., X_n$, then the lifetime of the machine is given by $X_1 + ... + X_n$.

3) We weigh 20 randomly selected people. The average weight of the group is $(X_1 + ... + X_{20})/20$

Let $X_1, ..., X_n$ be independent and identically distributed random variables. For each n let

$$Sn = X1 \cdots + Xn$$

Let $EX_i = u$ and $Var(X_i) = o$^2 (assuming that these are finite).

Some easy results are:

$$\mathbb{E}S_n = n\mathbb{E}X_1 = n\mu$$

and, by the independence of the summands,

$$Var(S_n) = n\ Var(X_1) = n\sigma^2$$

If we know the pdf or pmf of X_i, then we can (in principle) determine the pdf or pmf of S_n. The easiest way is to use **transform** techniques (Laplace transform, Characteristic function, etc).
An important property of these transforms is that the transform of the **sum** of independent random variables is equal to the **product** of the individual transforms.

Example

Example: Suppose each $X_i \sim Exp(lambda)$. The Laplace transform of X_i, say L is given by

$$L(s) = \mathbb{E}e^{-sX_i} = \frac{\lambda}{\lambda + s}$$

The Laplace transform of S_n, is given by

$$\begin{aligned} \mathbb{E}e^{-sS_n} &= \mathbb{E}e^{-s(X_1 + \cdots + X_n)} \\ &= \mathbb{E}e^{-sX_1} \cdots \mathbb{E}e^{-sX_n} = (L(s))^n \\ &= \left(\frac{\lambda}{\lambda + s}\right)^n \end{aligned}$$

Using the uniqueness of Laplace transforms, this shows that S_n has a Gamma(n, lambda) distribution (Erlang distribution)

Law of Large Numbers

Consider the coin flip example. We expect that S_n/n is close to the unknown p for large n.
We know this happens "empirically".

In general, we expect S_n/n to be close to u. Does this happen in our mathematical model?
By *Chebyshev's inequality* we have for all $e > 0$,

$$\mathbb{P}\left(\mid \frac{S_n}{n} - \mu \mid > \epsilon\right) \leq \frac{Var(S_n/n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

as *n* → *infinite*.

In other words the probability that S_n/n is more than *e* away from *u* can be made arbitrarily small by choosing *n* large enough.
This is the **Weak Law of Large Numbers**.

There is also a **Strong Law of Large Numbers:**

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1$$

as *n* → *infinite*

Central Limit Theorem

The Central Limit Theorem states, roughly, this: The **sum** of a large number of **iid** random variables has **approximately** a **Gaussian** distribution.
More precisely, it states that for all *x*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x)$$

where *Phi* is the cdf of the standard normal distribution.

Approximating Binomial by Normal

Using the CLT we thus find the following important approximation:
Let *X ~ Bin(n, p)*. For large *n*, we have

$$\mathbb{P}(X \leq k) \approx \mathbb{P}(Y \leq k)$$

where *Y ~ N(np, np(1 - p))*.

As a rule of thumb, the approximation is accurate if both *np* and *n(1 - p)* are larger than 5.

We can improve on this somewhat by using a continuity correction, as illustrated by the following graph for the pmf of the *Bin(10, 1/2)* distribution.

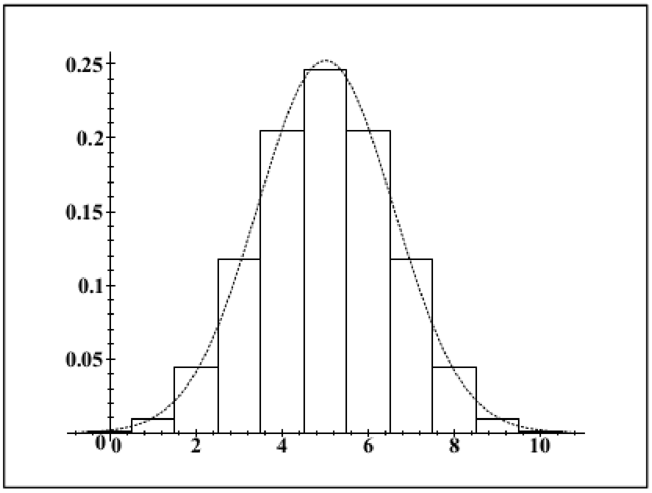

Figure 1: Approximating Binomial by Normal

For example,

$$\mathbb{P}(X = k) \approx \mathbb{P}(k - \frac{1}{2} \leq Y \leq k + \frac{1}{2})$$

Example

Example: Let *X ~ Bin(200, 0.51)*, and suppose we wish to to calculate *P(X <= 99)*.
Let *Y ~ N(200 x 0.51, 200 x 0.51 x 0.49)*, and let *Z* be standard normal. Using the CLT we have

$$\begin{aligned}\mathbb{P}(X \le 99) &\approx \mathbb{P}(Y \le 99)\\ &= \mathbb{P}\left(\frac{Y-102}{\sqrt{49.98}} \le \frac{99-102}{\sqrt{49.98}}\right)\\ &= \mathbb{P}(Z \le -0.4243) = 1 - \mathbb{P}(Z \le 0.4243)\\ &= 0.3357\end{aligned}$$

Using the continuity correction we find

$$\mathbb{P}(X \le 99) \approx \mathbb{P}(Y \le 99 + \frac{1}{2}) = 0.3618$$

Approximating via the CLT

Exercise: The number of calls *X* arriving at a call centre during an hour has a *Poi(100)* distribution.
Show, using probability generating functions, that *X* has the same distribution as $X_1 + ... + X_{100}$, where $X_1, ..., X_{100}$ are independent *Poi(1)*-distributed random variables.
Use this fact to approximate (with the CLT) the probability that there are more than 130 arrivals during an hour

Extremes of Independent Random Variables

In addition to the average behaviour of iid variates $X_1, ..., x_n$, we are often interested in the extremes – that is, how the largest (or smallest) variate behaves.

If $M = max\{X_1, ..., X_n\}$, we have seen (by example) that

$$\begin{aligned}F_M(m) &= \mathbb{P}(M \le m) = \mathbb{P}(X_1 \le m, \ldots, X_n \le m)\\ &= \mathbb{P}(X_1 \le m)^n = (F_X(m))^n\end{aligned}$$

What distribution does *M* have, as *n* → *infinite*

Remark: It turns out that, when *M* is suitably shifted and scaled, there are essentially three possibilities (listed here for completeness). The Gumbel distribution (*u element R, o > 0*):

$$f(x) = \frac{1}{\sigma} exp\left[-\frac{x-\mu}{\sigma} - exp\left[-\frac{x-\mu}{\sigma}\right]\right], x \in \mathbb{R}$$

The Frechet distribution (*u element R, o > 0, alpha > 0*):

$$f(x) = \frac{\alpha}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{-\alpha-1} exp\left[-\left(\frac{x-\mu}{\alpha}\right)^{-\alpha}\right], x > 0$$

The reversed Weibull distribution (*u element R, o > 0, alpha > 0*):

$$f(x) = \frac{\alpha}{\sigma}\left(\frac{\mu-x}{\sigma}\right)^{\alpha-1} exp\left[-\left(\frac{\mu-x}{\sigma}\right)^{\sigma}\right], x < \mu$$

Similarly, if $M = min\{X_1, ..., X_n\}$, we have that

$$\begin{aligned}F_M(m) &= \mathbb{P}(M \le m) = 1 - \mathbb{P}(M > m)\\ &= 1 - \mathbb{P}(X_1 > m, \ldots, X_n > m)\\ &= 1 - \mathbb{P}(X_1 > m)^n = 1 - (1 - F_X(m))^n\end{aligned}$$

Remark: It turns out that, when *M* is suitably shifted and scaled, there are again essentially three possibilities as *n* → *infinite*, being the distribution of *Y = -X*, where *X* is one of the three listed for the largest extreme value.

Summary

- Law of Large Numbers: statement, weak, strong
- Central Limit Theorem: statement, approximation of sums via CLT, examples
- Extreme Value Distributions: calculation (finite *n*), limiting behaviour (statement)

Statistics, Likelihood, and Estimation

Sums and Extremes of Independent Random Variables

- Law of Large Numbers; statement, weak, strong
- Central Limit Theorem; statement, approximation of sums via CLT, examples
- Extreme Value Distributions; calculation (finite *n*), limiting behaviour (statement)

Statistics

Data x is viewed as the outcome of a random variable X described by a probabilistic model. Usually, model is specified up to a (multidimensional) parameter: $X \sim F(.;\theta)$ for some element in Θ. In classical (frequentist) statistics, purely concerned with the model and in particular with the parameter $\varnothing$.
For example, given data, we may wish to

- estimate the parameter,
- perform statistical tests on that parameter, or
- validate the model

In Bayesian statistics, concerned with distribution of parameter $\theta \sim F(\theta)$.

Any real- or vector-valued function of data x or X is called a **statistic** of the data.
For example, the sample mean is a statistic:

$$T = T(x) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

given an outcome of X, or as a random variable

$$T = T(X) = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Often, we will view data as a series of independent outcomes from the same random experiment: $X = (X_1, \ldots, X_N)$, where $X_1, \ldots, X_N$ are iid from $F(.;\theta)$. $\{X_1, \ldots, X_N\}$ is called a **random sample** (from $F(.;\theta)$ or from X).
Therefore, the joint cdf of a random sample is given by

$$F(x;\theta) = \prod_{k=1}^{N} F(x_k;\theta)$$

and so the joint pdf/pmf is of the same form:

$$f(x;\theta) = \prod_{k=1}^{N} f(x_k;\theta)$$

Likelihood

When viewed as a function of θ, then point pdf/pmf of a random sample is called the **Likelihood**:

$$L(\theta;x) = f(x;\theta)$$

The (natural) logarithm of the likelihood

$$l(\theta;s) = ln L(\theta;x)$$

is called the **log-likelihood**

Likelihood Example

Example: Model X_{1}, $_{2}$, ..., $X_N \sim iid\ Bin(m, p)$; m known, p unknown, in $\Theta = (0, 1)$
pmf:

$$f(x;p) = \binom{m}{x} p^x (1-p)^{m-x}, x \in \{0, 1, \ldots, m\}$$

Therefore, the likelihood can be written as

$$L(p;X) = \prod_{i=1}^{N} \binom{m}{x_i} p^{x_i}(1-p)^{m-x_i}$$

$$= p^{\sum_{i=1}^{N} x_i}(1-p)^{Nm-\sum_{i=1}^{N} x_i} \prod_{i=1}^{N} \binom{m}{x_i}$$

Maximum Likelihood Estimation

How do we find "good" estimators for model parameters? Given data and a parametric model, how to find a member of that family (point estimate) from which the data is "most likely" to have come?
Given data x, one approach is to maximize the likelihood in θ – that is, find

$$\hat{\theta} \in \Theta$$

for which

$$L(\hat{\theta};x) \geq L(\theta;x), \theta \in \Theta$$

A maximizer

$$\hat{\theta} \equiv \hat{\theta}(x)$$

of L is called a **maximum likelihood estimate** (MLE). The corresponding random variable &hat;θ(X) is called a maximum likelihood estimator (also MLE).
Remark: A maximiser of *l* equivalent to a maximiser of *L*

ML Estimation Example: Binomial Probability

Example: Continuing our example, recall that we found

$$L(p; X) = \prod_{i=1}^{N} \binom{m}{x_i} p^{x_i}(1-p)^{m-x_i}$$

$$= p^{\sum_{i=1}^{N} x_i}(1-p)^{Nm-\sum_{i=1}^{N} x_i} \prod_{i=1}^{N} \binom{m}{x_i}$$

How do we find an MLE?
Maximisation Strategy: Since *L* is a continuous function of *p*, find *p* such that

$$\frac{d}{dp}L(p; x) = 0$$

Working directly with *L* appears cumbersome; obtain the log-likelihood and work with that instead.
Taking the natural logarithm of *L*, we obtain the log-likelihood:

$$l(p; X) = ln(p) \sum_{i=1}^{N} x_i ln(1-p) \left(Nm - \sum_{i=1}^{N} x_i \right) + \sum_{i=1}^{N} ln \left(\binom{m}{x_i} \right)$$

First Derivative with respect to *p*:

$$\frac{d}{dp}l(p; X) = \frac{1}{p}\sum_{i=1}^{N} x_i - \frac{1}{1-p}\left(Nm - \sum_{i=1}^{N} x_i \right)$$

Set to zero and rearrange to find critical point:

$$(1-p)\sum_{i=1}^{N} x_i = p\left(Nm - \sum_{i=1}^{N} x_i \right)$$

Unique solution:

$$\hat{p} = \frac{1}{Nm}\sum_{i=1}^{N} x_i$$

What type of critical point is this?
Second Derivative with respect to *p*:

$$h(p) = \frac{d^2}{dp^2}l(p; X) = -\frac{1}{p^2}\sum_{i=1}^{N} x_i - \frac{1}{(1-p)^2}\left(Nm - \sum_{i=1}^{N} x_i \right) < 0$$

Therefore &hat;p is a local maximiser.
Moreover, *l(p;X)* → ∞ as *p* → *0* or *p* → *1* (boundary of Θ). Thus &hat;p is in fact a global maximiser. Therefore, we have the Maximum Likelihood Estimator:

$$\hat{p} = \frac{1}{Nm}\sum_{i=1}^{N} X_i$$

Summary

- Statistics; definition, example
- Likelihood and log-likelihood; definition, binomial example
- Maximum Likelihood Estimation; definition, examples, bias, consistency

Confidence Intervals and Hypothesis Testing

Statistics, Likelihood, and Estimation

- Statistics; definition, example
- Likelihood and log-likelihood; definition, binomial example
- Maximum Likelihood Estimation; definition, examples, bias, consistency

Confidence Intervals

Last time, we were introduced to maximum likelihood estimation, which provided a systematic way of obtaining estimates and estimators &hat;θ of unknown parameters contained in θ ∈ Θ
How can we gauge the accuracy of &hat;θ?
Confidence intervals (sometimes called interval estimates) provide a precise way of describing the uncertainty of &hat;θ

Formally, given random variables $X_1, ..., X_n$ whose joint distribution depends on some unknown θ ∈ Θ, a **(1 - α) stochastic confidence interval** is a pair of statistics

$$T_1(X_1, \ldots, X_n) \text{and} T_2(X_1, \ldots, X_n)$$

with the property that

$$\mathbb{P}(T_1 < \theta < T_2) \geq 1 - \alpha, \text{ for all } \theta \in \Theta$$

for some α ∈ [0, 1]
That is, (T_1, T_2) is a random interval, based only on the (as yet to be observed) outcomes $X_1, ..., X_n$, that contains the unknown θ with probability at least *1* - α.

A realisation of the random interval, say (t_1, t_2), is called a **(1 - α) numeric confidence interval** for θ.

Remark: Whilst stochastic confidence intervals contain the unknown θ with probability at least *1* - α, their numerical counterparts either contain θ or they do not. It may be helpful to think of a Bernoulli analogy, where "success" occurs with probability (at least) *1* - α – then outcomes are either "successes" or "failures"

Confidence Interval Example

Example: Model: $X_1, X_2, ..., X_n$ *~iid N*(μ, σ^2); σ^2 known, μ unknown, in Θ = R.
We have seen that

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad \sim N(\mu, \frac{\sigma^2}{N})$$

Therefore,

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \sim N(0, 1)$$

Hence,

$$\mathbb{P}\left(z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

where z_γ is the γ-quantile of the standard normal distribution.

Rearranging, we have

$$\mathbb{P}\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right) = 1 - \alpha$$

Note that, by symmetry, the quantiles satisfy $-z_{\alpha/2} = z_{1-\alpha/2}$. Hence a stochastic *1* - α confidence interval for μ in this case is

$$\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}, \bar{X} z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}\right)$$

which is often abbreviated to

$$\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}$$

Approximate Confidence Intervals

When

$$\mathbb{P}(T_1 < \theta < T_2) \geq 1 - \alpha, \text{ or all } \theta \in \Theta$$

only holds approximately, we call (T_1, T_2) an **approximate (1 - α) confidence interval** for θ.

Remark: We can often employ the central limit theorem to construct such approximate confidence intervals, as we shall see next.

Approximate Confidence Interval Example

Example: Model $X_1, X_2, ..., X_N$ *~iid Bin(m, p)*; *m* known, *p* unknown, in Θ = (0, 1), with MLE for *p*:

$$\hat{p} = \frac{1}{Nm}\sum_{i=1}^{N} X_i$$

Notice that $Y = \Sigma^{N}_{i=1} X_i$ can be thought of as $Y \sim Bin(Nm, p)$, and so by the central limit theorem,

$$Y \;_{approx}\mathbf{N}(Nmp, Nmp(1-p))$$

or equivalently

$$\hat{p} \;_{approx}\mathbf{N}\left(p, \frac{p(1-p)}{Nm}\right)$$

Therefore, we have

$$\mathbb{P}\left(z_{\alpha/2} \leq \frac{\hat{p}-p}{\frac{\sqrt{p(1-p)}}{\sqrt{Nm}}} \leq z_{1-\alpha/2}\right) \approx 1-\alpha$$

By the law of large numbers, &hat;p&approx;p, so we may replace *p* in the denominator to obtain

$$\mathbb{P}\left(z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}/\sqrt{Nm}} \leq z_{1-\alpha/2}\right) \approx 1-\alpha$$

Rearranging, and using the symmetry of standard normal quantiles, we have

$$\mathbb{P}\left(\hat{p} - z_{1-\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{Nm}} \leq p \leq \hat{p} z_{1-\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{Nm}}\right) \approx 1-\alpha$$

which is an approximate $1 - \alpha$ confidence interval for *p*:

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{Nm}}$$

Hypothesis Testing

Closely related to the notion of confidence intervals is that of hypothesis tests. In hypothesis testing, given data, we wish to determine which of two competing hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ holds true. H_0 is called the **null hypothesis** and contains the "status quo" statement, whereas H_1 is called the **alternative hypothesis** which is unlikely to have occurred if H_0 were true.
Remark: Usually, $\Theta_0 \cap \Theta_1 = \varnothing$

Outcomes of hypothesis tests are decisions as to whether to accept the "status quo" H_0, or reject the "status quo" in favour of the alternative H_1. As such, we seek a decision rule based on the outcome of a statistic *T*.

- Decision Rule 1: Reject H_0 if *T* falls in some critical region *C*
- Decision Rule 2: Reject H_0 if $P(T \in C)$ is less than some critical p-value p_c.

Remark: Common critical regions are one-sided (C = (-∞, c], C = [c, ∞)), or two-sided (C = (-∞, c_1] ∪ [c_2, ∞), c_1 <= c_2)

Regardless of which type of decision rule is employed, we can make two types of error.

Decision	H_0 True	H_1 True
Retain H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

Remark: We can think of Type I error as a "false positive" and Type II error as a "false negative".
In classical statistics, Type I error is considered more serious, and so decision rules are designed to control this type of error.

We will denote the probability of a Type I error by α, and the probability of a Type II error by β.
Remark: The **power** of a statistical test is the probability of correctly rejecting the null, 1 - β
We will design our decision rules around a predetermined **significance level** α, which describes the acceptable level of Type I error for our test. In this framework, the two types of decision rule are equivalent:

- Decision Rule 2: Reject H_0 if $P(T \in C_\alpha) <= \alpha$
- Decision Rule 1: Reject H_0 if *T* falls in C_α

Hypothesis Testing Example

Example: Model $X_1, X_2, ..., X_N \sim iid\ N(\mu, \sigma^2)$; σ^2 known, μ unknown, in Θ = R. We can readily adapt our previous work to form a hypothesis test together with a decision rule about the unknown μ.
Let H_0: $\mu = \mu_0$ and H_1: $\mu \neq \mu_0$;
Under the null hypothesis H_0,

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}} \quad N(0,1)$$

and so

$$C_\infty = (-\infty, z_{\alpha/2}] \cup [z_{1-\alpha/2}, \infty)$$

is a critical region satisfying

$$\mathbb{P}_{H_0}(T \in C_\alpha) \leq \alpha$$

Therefore, we reject H_0 if our observed statistic *t* falls in C_α

Summary

- Confidence intervals; definition, stochastic, numerical, approximate, examples
- Hypothesis testing; decision rules, null and alternative hypotheses, Type I and II error, significance level, power, critical region, critical *p*-value, one- and two-sided regions (hence tests).

Confidence Intervals and Hypothesis Testing II

Sample Variance

For a single normal random sample with known variance σ^2, we have seen that the sample mean (X bar) is normally distributed, and can therefore construct confidence intervals and hypothesis tests for the unknown mean μ

How can we proceed when σ^2 is unknown?
First, we will determine an appropriate estimator for σ^2, and state its distribution for a normal random sample.

Recall that we defined the sample variance of data as

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2$$

For a random sample, is the associated random variable an unbiased estimator for σ^2?
We have

$$\begin{aligned}
\mathbb{E}\hat{\sigma^2} &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[X_i^2] - \mathbb{E}[\bar{X}^2] \\
&= \mathbb{E}[X_1^2] - \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)^2\right] \\
&= \mathbb{E}[X_1^2] - \mathbb{E}\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} X_i X_j\right] \\
&= \mathbb{E}[X_1^2] - \frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^{N} X_i^2 + \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} X_i X_j\right] \\
&= \mathbb{E}[X_1^2] - \frac{1}{N}\mathbb{E}[X_1^2] - \frac{N(N-1)}{N^2}\mathbb{E}[X_1]^2 \\
&= \frac{N-1}{N}(\mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2) = \frac{N-1}{N}\sigma^2
\end{aligned}$$

Therefore, &hat;σ^2 is a biased (but consistent) estimator for σ^2.
Remark: &hat;σ^2 is the MLE of σ^2 for a normal random sample.

We can easily correct for the bias in the **(bias corrected) sample variance:**

$$\begin{aligned}
S^2 &= \frac{N}{N-1}\hat{\sigma^2} \\
&= \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N} X_i^2 - \frac{N}{N-1}\bar{X}^2
\end{aligned}$$

For a normal random sample, it turns out that

$$(N-1)\frac{S^2}{\sigma^2} \sim X^2_{N-1} \equiv \text{Gamma}(\frac{N-1}{2}, \frac{1}{2})$$

Remark: The fact that the degrees of freedom is *N - 1* comes from the fact that there are only *N - 1* linearly independent elements of

$$\begin{pmatrix} X_1 - \bar{X} \\ \cdot \\ \cdot \\ \cdot \\ X_N - \bar{X} \end{pmatrix}$$

Sample Variance Example

Example: For a normal random sample $X_1, ..., X_N$ *~iid* $N(\mu, \sigma^2)$ with unknown mean μ and variance σ^2, find a $1 - \alpha$ (stochastic) confidence interval for σ^2.

Since $(N - 1)S^2/\sigma^2 \sim X_{N-1}^{\ \ 2}$, we have by definition

$$\mathbb{P}\left(X^2_{N-1;\alpha/2} \leq (N-1)\frac{S^2}{\sigma^2} \leq X^2_{N-1;1-\alpha/2}\right) = 1 - \alpha$$

where $X_{N-1;\gamma}^{\ \ 2}$ denotes the γ-quantile of this chi-squared distribution

Since $\sigma^2 > 0$ and $S^2 > 0$, we rearrange as follows:

$$\mathbb{P}\left(\frac{1}{X^2_{N-1;\alpha/2}} \geq \frac{\sigma^2}{(N-1)S^2} \geq \frac{1}{X^2_{N-1;1-\alpha/2}}\right) = 1 - \alpha$$

giving

$$\mathbb{P}\left(\frac{(N-1)S^2}{X^2_{N-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{(N-1)S^2}{X^2_{N-1;\alpha/2}}\right) = 1 - \alpha$$

Hence, a stochastic $1 - \alpha$ confidence interval for σ^2 for a normal random sample is

$$\left(\frac{(N-1)S^2}{X^2_{N-1;1-\alpha/2}}, \frac{(N-1)S^2}{X^2_{N-1;\alpha/2}}\right)$$

We can easily construct hypothesis tests at significance level α.

If $H_0 : \sigma^2 = \sigma_0^{\ 2}$ and $H_1 : \sigma^2 \neq \sigma_0^{\ 2}$, then our test statistic is

$$T = (N-1)\frac{S^2}{\sigma_0^2}$$

which (under H_0) has a $X^2_{\ N-1}$ distribution

Therefore, we reject H_0 in favour of H_1 if *T* falls in the (two-sided) critical region

$$(-\infty, X^2_{N-1;\alpha/2}] \cup [X^2_{N-1;1-\alpha/2}, \infty)$$

Similarly, if $H_0 : \sigma^2 = \sigma_0^{\ 2}$ and $H_1 : \sigma^2 > \sigma_0^{\ 2}$, we reject H_0 in favour of H_1 if *T* falls in the (right one-sided) critical region

$$[X^2_{N-1;1-\alpha}, \infty)$$

and if $H_0 : \sigma^2 = \sigma_0^{\ 2}$ and $H_1 : \sigma^2 < \sigma_0^{\ 2}$, we reject H_0 in favour of H_1 if *T* falls in the (left one-sided) critical region

$$(-\infty, X^2_{N-1;\alpha}]$$

Sample Mean with Unknown Variance

We have seen how to construct confidence intervals and hypothesis tests for a normal random sample with known variance σ^2. How does this change when σ^2 is unknown, and must instead be replaced by an estimate?

Recall that $X_1, X_2, ..., X_N$ *~iid* $N(\mu, \sigma^2)$, and consider the hypothesis test $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$. Our test statistic in this case simply replaces the known σ with its unbiased estimator *S* = &sqrt;S^2, giving

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{N}}$$

If H_0 is true, then it turns out that *T* has a **(Student's) t** distribution, with *N - 1* degrees of freedom, which we will write as t_{N-1}. We will not concern ourselves with the particulars of this distribution, other than to note a few salient points:

- A *t*-distribution random variable is continuous, symmetric around zero, and has non-zero pdf over R (just like the standard normal distribution)

- As with any other distribution, we may compute γ-quantiles for a t_N-distributed random variable, which we will denote by $t_{N;\gamma}$.
  - Like the standard normal distribution, we will rely on tables or numerical computation for quantiles and probabilities.
- As $N \to \infty$, t_N converges in distribution to N(0, 1). (Moreover, t_1 is the Cauchy distribution)

Accepting that $T \sim t_{N-1}$, we construct a two-sided critical region at significance level α:

$$(-\infty, t_{N-1;\alpha/2}] \cup [t_{N-1;1-\alpha/2}, \infty)$$

and we reject H_0 if the outcome of our test statistic falls in this region. Similarly, critical regions for one-sided tests are easily constructed:

- $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$. Critical region: $[t_{N-1;1-\alpha}, \infty)$
- $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$. Critical region: $(-\infty, t_{N-1;\alpha}]$

Moreover, confidence intervals for the mean are straight-forwardly constructed from T:

$$\left(\bar{X} - t_{N-1;1-\alpha/2}\frac{S}{\sqrt{N}}, \bar{X} - t_{N-1;\alpha/2}\frac{S}{\sqrt{N}}\right)$$

or more compactly, by the symmetry of this distribution around zero:

$$\bar{X} \pm t_{N-1;1-\alpha/2}\frac{S}{\sqrt{N}}$$

Summary

- Sample variance; bias and correction, confidence intervals and hypothesis tests for normal population.
- Sample mean with unknown variance; Student's *t* distribution (briefly), confidence intervals and hypothesis tests for normal population

Confidence Intervals and Hypothesis Testing III

- Sample variance; bias and correction, confidence intervals and hypothesis tests for normal population
- Sample mean with uknown variance; Student's *t* distribution (briefly), confidence intervals and hypothesis tests for normal population

Two Sample Inference

Previously, we have seen how to construct confidence intervals and hypothesis tests for unknown parameters for a single random sample. However, in many cases we are interested in inference regarding the unknown parameters of two random samples. How does the construction of confidence intervals and hypothesis tests extend to this situation?

Two Sample Inference Example

Example: Model $X_1, \ldots, X_M \sim iid\ N(\mu_X, \sigma^2_X)$ independent of $Y_1, \ldots, Y_N \sim iid\ N(\mu_Y, \sigma^2_Y)$, with known variances σ^2_X and σ^2_Y, but unknown means μ_X and μ_Y.
Construct a *1* - α stochastic confidence interval for the difference in means, $\mu_X - \mu_Y$
Firstly, notice that *&bar;X ~ N(*μ_X, σ^2_X*/M)* independent of *&bar;Y ~ N(*μ_Y, σ^2_Y*/N)*.
Therefore, *&bar;X - &bar;Y ~ N(*$\mu_X - \mu_Y$, σ_X^2*/M +* σ_Y^2*/N)*, and so

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2_X}{M} + \frac{\sigma^2_Y}{N}}} \sim N(0, 1)$$

Hence, by definition,

$$\mathbb{P}(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$$

Rearranging as usual, we obtain an output which can be put more compactly (and using the symmetry of normal quantiles)

$$(\bar{X} - \bar{Y}) \pm z_{1-\alpha/2}\sqrt{\sigma^2_X/M + \sigma^2_Y/N}$$

as a *1* - α stochastic confidence interval for the difference in means.

Remark: If each random sample has a common known variance $\sigma_X^{\ 2} = \sigma_Y^{\ 2} = \sigma^2$, then this confidence interval reduces to

$$(\bar{X} - \bar{Y}) \pm z_{1-\alpha/2}\sigma\sqrt{\frac{1}{M} + \frac{1}{N}}$$

This work can be extended to create hypothesis tests in the usual way, as follows. For the two-sided test, with a pair of normal random samples with known variances σ_X^2 and σ_Y^2, we have $H_0 : (\mu_X - \mu_Y) = \delta_0$ and $H_1 : (\mu_X - \mu_Y) \neq \delta_0$.
Under H_0

$$T = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{\sigma_X^2}{M} + \frac{\sigma_Y^2}{N}}} \sim N(0,1)$$

and so the critical region for a test with significance level α is

$$C_\alpha = (-\infty, z_{\alpha/2}] \cup [z_{1-\alpha/2}, \infty)$$

One-sided tests, and tests with common variance σ^2 can be constructed in the same way.

Two Sample Inference with Unknown Variance

How does this change when the variances of the samples are unknown?
There are two possibilities:

- The unknown variances are not assumed to be the same
- The unknown variances are assumed to be the same

In the first case, we may estimate σ_X^2 by S_X^2, and σ_Y^2 by S_Y^2.
Then we may construct the same intervals and tests as before, replacing each variance by its estimator. This will yield approximate confidence intervals, and approximate hypothesis tests, which become more exact as both of the sample sizes become large.
In the second case, we need to estimate the common variance. The (uncorrected) pooled sample variance would just be

$$\hat{\sigma_p^2} = \frac{1}{M+N}\left(\sum_{i=1}^{M}(X_i - \bar{X})^2 + \sum_{j=1}^{N}(Y_j - \bar{Y})^2\right)$$

However, as we have seen before, this is a biased estimator. Here, we can easily compute

$$\mathbb{E}[\hat{\sigma_p^2}] = \frac{M-1+N-1}{M+N}\sigma^2$$

so the (bias corrected) pooled sample variance is just

$$S_p^2 = \frac{1}{M+N-2}\left(\sum_{i=1}^{M}(X_i - \bar{X})^2 + \sum_{j=1}^{N}(Y_j - \bar{Y})^2\right)$$

Therefore, we can use our previous work, and note that

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{M} + \frac{1}{N}}} \sim t_{M+N-2}$$

Hence, a $1 - \alpha$ stochastic confidence interval for $(\mu_X - \mu_Y)$ with unknown common variance is

$$(\bar{X} - \bar{Y}) \pm t_{M+N-2;1-\alpha/2} S_p \sqrt{\frac{1}{M} + \frac{1}{N}}$$

For the two-sided test, with a pair of normal random samples with unknown common variance σ^2, we have $H_0 : (\mu_X - \mu_Y) = \delta_0$ and $H_1 : (\mu_X - \mu_Y) \neq \delta_0$. Under H_0,

$$T = \frac{(\bar{X} - \bar{Y}) - \delta_0}{S_p\sqrt{\frac{1}{M} + \frac{1}{N}}} \sim t_{M+N-2}$$

and so the critical region for a test with significance level α is

$$C_\alpha = (-\infty, t_{M+N-2;\alpha/2}] \cup [t_{M+N-2;1-\alpha/2}, \infty)$$

One-sided tests are simply constructed as seen previously

Approximate Intervals and Tests

We can readily adapt the confidence intervals and tests described so far to give approximate results by appealing to the central limit theorem.

Exercise: If $X \sim Bin(M, p_X)$ independently of $Y \sim Bin(N, P_Y)$, show that an approximate $1 - \alpha$ stochastic confidence interval for $p_X - p_Y$ is

$$(\hat{p_X} - \hat{p_Y}) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p_X}(1-\hat{p_X})}{M} + \frac{\hat{p_Y}(1-\hat{p_Y})}{N}}$$

where

$$\hat{p_X} = \frac{X}{M}, \qquad \hat{p_Y} = \frac{Y}{N}$$

Two Sample Inference for Variances

How can we construct confidence intervals and hypothesis tests for the unknown variances of two random samples?

Last time, we stated that for a normal random sample, $X_1, \ldots, X_M$ ~*iid* $N(\mu_X, \sigma_X^2)$,

$$(M-1)\frac{S_X^2}{\sigma_X^2} \sim X_{M-1}^2 \equiv \text{Gamma}(\frac{M-1}{2}, \frac{1}{2})$$

This time, we will state that if we have two independent normal random samples $X_1, \ldots, X_M$ ~*iid* $N(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_N$ ~*iid* $N(\mu_Y, \sigma_Y^2)$,

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{M-1,N-1}$$

where $F_{m,n}$ is the *F*-distribution with *m* and *n* degrees of freedom

Remark: As with the *t*-distribution, we will not go into details regarding the *F*-distribution, but simply accept this and rely on numerical computation or tabulation of its quantiles.

Using this fact, we may write by definition

$$\mathbb{P}\left(F_{N-1,M-1;\alpha/2} \leq \frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \leq F_{N-1,M-1;1-\alpha/2}\right) = 1-\alpha$$

Rearranging, we have a stochastic $1 - \alpha$ confidence interval for the ratio of the unknown population variances:

$$\mathbb{P}\left(F_{N-1,M-1;\alpha/2}\frac{S_X^2}{S_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq F_{N-1,M-1;1-\alpha/2}\frac{S_X^2}{S_Y^2}\right) = 1-\alpha$$

We may use this to construct hypothesis tests: $H_0 : \sigma_X^2 = \sigma_Y^2$ vs $H_1 : \sigma_X^2 \neq \sigma_Y^2$.

Under H_0,

$$\frac{S_X^2}{S_Y^2} \sim F_{M-1,N-1}$$

and so an appropriate critical region at the α significance level is

$$C_\alpha = (-\infty, F_{M-1,N-1;\alpha/2}] \cup [F_{M-1,N-2;1-\alpha/2}, \infty)$$

One-sided tests can be constructed as seen before.

Summary

- Two-sample difference of means; confidence intervals and hypothesis tests for normal population, known and unknown (common and not) variance.
- Two-sample ratio of variances; *F* distribution (briefly), confidence intervals and hypothesis tests for normal population.

Confidence Intervals and Hypothesis Testing IV

- Two-sample difference of means; confidence intervals and hypothesis tests for normal population, known and unknown (common and not) variance

Two Sample Inference for Variances

How can we construct confidence intervals and hypothesis tests for the unknown variances of two random samples?

Last time, we stated that for a normal random sample, $X_1, \ldots, X_M$ ~*iid* $N(\mu_X, \sigma_X^2)$,

$$(M-1)\frac{S_X^2}{\sigma_X^2} \sim X_{M-1}^2 \equiv \text{Gamma}(\frac{M-1}{2}, \frac{1}{2})$$

This time, we will state that if we have two independent normal random samples $X_1, \ldots, X_M$ ~*iid* $N(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_N$ ~*iid* $N(\mu_Y, \sigma_Y^2)$,

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{M-1,N-1}$$

where $F_{m,n}$ is the *F*-distribution with *m* and *n* degrees of freedom

Remark: As with the *t*-distribution, we will not go into details regarding the *F*-distribution, but simply accept this and rely on numerical computation or tabulation of its quantiles.

Using this fact, we may write by definition

$$\mathbb{P}\left(F_{N-1,M-1;\alpha/2} \leq \frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \leq F_{N-1,M-1;1-\alpha/2}\right) = 1-\alpha$$

Rearranging, we have a stochastic *1* - α confidence interval for the ratio of the unknown population variances:

$$\mathbb{P}\left(F_{N-1,M-1;\alpha/2}\frac{S_X^2}{S_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq F_{N-1,M-1;1-\alpha/2}\frac{S_X^2}{S_Y^2}\right) = 1-\alpha$$

We may use this to construct hypothesis tests: $H_0 : \sigma_X^2 = \sigma_Y^2$ vs $H_1 : \sigma_X^2 \neq \sigma_Y^2$
Under H_0

$$\frac{S_X^2}{S_Y^2} \sim F_{M-1,N-1}$$

and so an appropriate critical region at the α significance level is

$$C_\alpha = (\infty, F_{M-1,N-1;\alpha/2}] \cup [F_{M-1,N-1;1-\alpha/2}, \infty)$$

One-sided tests can be constructed as seen before

Goodness of Fit

Goodness of Fit refers to assessing the quality of a model in light of data. We have seen graphical goodness of fit procedures – namely quantile-quantile plots. We can also approach goodness of fit from a statistical viewpoint, by devising statistical tests based on the data directly (through the empirical cdf), or through first binning the data, and comparing expected bin values (based on our probabilistic model) to observed bin values from data.

Kolmogorov-Smirnov Test

Suppose that $X_1, ..., X_N$ is an random sample (that is, an iid sample) from some distribution with cdf *F*. If indeed $X_1, ..., X_N$ *~iid F*, then, when ordered $X_{(1)} < ... < X_{(N)}$,

$$\mathbb{P}(\frac{k-1}{N} < F(X_{(k)}) \leq \frac{k}{N}) = \frac{1}{N}, \qquad k = 1,\ldots,N$$

In other words, if $X_1, ..., X_N$ were a random sample from *F*, then $F(X_1), ..., F(X_N)$ would be random sample from *U[0, 1]*
This observation is basis of the Kolmogorov-Smirnov test, which utilizes the distribution of maximum deviation of a uniform random sample from the straight line (0, 0) - (1, 1)
The (scaled) Kolmogorov-Smirnov statistic is

$$K_N = \sqrt{N}\max_{i=1,\ldots,N}\max\left\{\left|F(X_{(i)}) - \frac{i}{N}\right|, \left|F(X_{(i)}) - \frac{(i-1)}{N}\right|\right\}$$

Figure 2: Kolmogorov-Smirnov Statistic

which, under the null hypothesis that the data is a random sample from *F* has a Kolmogorov-Smirnov distribution. Once more, we will not go into the details of this distribution, other than to note that its quantiles may be tabulated or computed numerically. In particular, for a particular outcome k_N of K_N, we can computer the *p*-value under the null hypothesis $p = P(K_N > k_N)$, and reject the null if *p* <= α, for some pre-specified significance level α.

Remark: The Kolmogorov-Smirnov test is non-parametric, in the sense that it does not test parameters of a particular distribution, but rather is applicable to any distribution form. Moreover, it can be used when the hypothesised *F* itself is an empirical cdf, and we wish to test whether certain observed data could have come from the same distribution as other known data.

X2 Goodness of Fit Tests

Suppose that we have an underlying model that $X_1, ..., X_N$ is a random sample from a distribution with cdf *F*. Then, we may consider binning the random sample into *M* mutually exclusive and exhaustive intervals, say $I_1 = (-\infty, a_1]$, $I_2 = (a_1, a_2]$, ..., $I_{M-1} = (a_{M-2}, a_{M-1}]$, $I_M = (a_{M-1}, \infty)$. If our model were true, then the counts of the number in each bin would follow a multinomial distribution: $(Y_1, ..., Y_M) \sim Mnom(N, \pi)$, where

$$\pi_k = \mathbb{P}(X_1 \in I_k)$$

The X^2 test statistic measures the discrepancy between observed counts in each bin, and the expected counts, if our model were true:

$$T = \sum_{i=1}^{K}\frac{(X_i - N\pi_i)^2}{N\pi_i}$$

It turns out that, if our model were true,

$$T \sim X_{K-1}^2$$

and so we can use this to test whether it is reasonable that observed data comes from our hypothesised distribution *F*.

Remark: A rule of thumb for the validity of this approximation is *Nπ >= 5*, for *i = 1, ..., K*.
In particular, if our observed test statistic *t* falls in the critical region

$$[X^2_{K-1;\alpha}, \infty)$$

then we would reject the hypothesis that our data is a random sample from *F*, at the α significance level.
Remark: Once again, notice that this form of testing is non-parametric, as it does not test the parameters of a particular distribution, or rely on a particular parametric distribution form for *F*.

X2 GoF Example

Example: Suppose we expect the number of hits to our website to be equally divided between Spring, Summer, Autumn, and Winter. Therefore, with *N* hits, and letting (Y_1, Y_2, Y_3, Y_4) be the number of hits per quarter, our model is ***Y*** *~ Mnom(N, (1/4, 1/4, 1/4, 1/4))*. If we had 100,000 hits last year, our expected number of hits per quarter under our model is 25,000. Suppose we observe 25,790, 25,618, 25,671, and 22,921 hits, and we wish to test our model at the α = 0.01 level. Then our test statistic is:

$$t = \frac{(25790-25000)^2}{25000} + \frac{(25618-25000)^2}{25000} + \frac{(25671-25000)^2}{25000} + \frac{(22921-25000)^2}{25000} \approx 231.1402$$

Under the null hypothesis, $T \sim_{appprox} X_3^2$, and so the *p*-value for this outcome is

$$\mathbb{P}(T > t) \approx 0$$

This is less than 0.01, so we reject H_0 and conclude that the data is not consistent with the model at the α = 0.01 significance level

Summary

- Two-sample ratio of variances; *F* distribution (briefly), confidence intervals and hypothesis tests for normal population.
- Goodness of fit; Kolmogorov-Smirnov (basis, statistic, test, illustration), X^2 (basis, statistic, test, example)

Regression

Regression

Regression models are used to describe functional relationships between explanatory variables **X** and response variables **Y**. In such models, the response variables **Y** are assumed to be a function **f** of the explanatory variables **X**, corrupted by noise from a (zero-mean) error model. Usually, the function **f** is parametric, depending on some parameter vector β, so that we may write the regression model as

$$Y = f(X;\beta)\epsilon$$

where ε is a zero-mean random variable encoding our error model.
Remark: Usually, we assume that ε ~ $N(0, \sigma^2 I)$, independent of all other random variables.
In this framework, given outcomes of the explanatory variables **X = x**, the response variables **Y** have conditional expectation

$$\mathbb{E}[Y \mid X = x] = f(x;\beta)$$

In a **linear regression** model, this relationship is linear, so that

$$\mathbb{E}[Y \mid X = x] = \beta_0\beta_1 x$$

In other words, in a linear regression model, given the outcome of the *i*-th explanatory variable $X_i = x_i$, the *i*-th response variable Y_i modelled as

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, N$$

where the usual error model is $\varepsilon_1, \ldots, \varepsilon_N$ *~iid* $N(0, \sigma^2)$

Linear Regression

The line

$$y = \beta_0 + \beta_1 x$$

is called the **regression line** (or more generally, regression curve). Notice that the linear regression model depends on the unknown coefficients β_0 and B_1, as well as the (typically unknown) error variance σ^2. Given outcomes of the explanatory variables **X = x** and response variables **Y = y**, how can we determine the unknown coefficients in $\beta = (\beta_0, \beta_1)^T$?

Least Squares Method

To do so, we first need a reasonable way of determining how well a given parameter setting β fits the data. The usual approach is to examine the **residuals** given a particular parameter setting:

$$r_i = y_i - (\beta_0 + \beta_1 x_i)$$

which is just the residual value of observed response y_i once the model involving the explanatory variables has been removed. A typical measure of overall model fit is then the sum of the squared residuals:

$$\sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N}(y_i - (\beta_0 + \beta_1 x_i))^2$$

The usual approach for finding the best parameters is to minimise the sum of the squared residuals. This approach is called the **method of least squares**.
Formally, we seek to minimise

$$\sum_{i=1}^{N} r_i^2 = || r ||^2$$

with respect to the parameter vector β. Whenever we may write a **linear model**

$$Y = A\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I)$$

for some parameter vector β and **design matrix** *A* (whose elements may depend on the outcomes of explanatory variables **X = x**), for given outcomes **Y = y** we have

$$|| r ||^2 = || y - A\beta ||^2$$

Therefore, for a linear model, we seek a parameter vector β that solves

$$\nabla_\beta || y - A\beta ||^2 = 0$$

or in other words

$$A^T(y - A\beta) = 0$$

This set of linear equations in β are called the **normal equations.** Rearranging, we have

$$A^T A\beta = A^T y$$

so that if (A^TA) is invertible,

$$\beta = (A^T A)^{-1} A^T y$$

Remark: The design matrix *A* can always be chosen so that (A^TA) is invertible. However, in practice, we never explicitly compute its inverse, but rather solve the set of linear equations numerically (for example via Gaussian elimination).

Least Squares Example

Example: Suppose we have the linear regression model.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, N$$

where $\varepsilon_1, ..., \varepsilon_N$ *~iid N(0, σ^2)*. Given an outcome **Y = y**, what is the least squares solution for β = *(*β_0*,* β_1*)^T*?
It is convenient to rewrite this model as

$$Y = A\beta + \epsilon$$

where the design matrix A is given by

$$\begin{pmatrix} 1 & x_1 \\ . & . \\ . & . \\ . & . \\ 1 & x_N \end{pmatrix}$$

Then, given **Y = y**, the least squares solution *&hat;*β solves

$$(A^T A)\hat{\beta} = A^T y$$

Remark: In this case, we can solve for *&hat;*β exactly, yielding

$$\hat{\beta_1} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

and

$$\hat{\beta_0} = \bar{y} - \beta_1 \bar{x}$$

where as usual *&bar;x* and *&bar;y* denote the average of outcomes *{x_i}* and *{y_i}* respectively.

Example: Suppose we have a quadratic regression model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \qquad i = 1, \ldots, N$$

where $\varepsilon_1, ..., \varepsilon_N$ *~iid N(0, σ^2)*. Given an outcome **Y = y**, what is the least squares solution for β = *(*β_0*,* β_1*,* β_2*)^T*?
Once again, it is convenient to rewrite this model as

$$Y = A\beta + \epsilon$$

where the design matrix A is given by

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & x_N & x_N^2 \end{pmatrix}$$

Then, given **Y = y**, the least squares solution *&hat;*β solves the normal equations

$$(A^T A)\hat{\beta} = A^T y$$

Remark 1: As these illustrate, linear models depend linearly on the unknown parameters β, and do not require that the form of the regression curve be linear. In contrast, whenever the regression model is not linear in β, then it is said to be a nonlinear regression model.
Remark 2: For linear models, it turns out that the least squares solution &hat;β is the maximum likelihood solution.

Summary

- Regression; model, error term, linear regression, quadratic regression, linear models, residuals
- Least squares; basis, normal equations, examples

Regression II

Linear Models

Recall that a linear model is a regression model of the form

$$Y = A\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I)$$

for some parameter vector β and **design matrix** A (whose elements may depend on the outcomes of explanatory variables **X = x**). Given observed responses **Y = y** for a linear model, we had that the least squares solution &hat;β for the unknown parameters solved

$$A^T A\hat{\beta} = A^T y$$

Remark: Geometrically, &hat;β is the projection of **y** onto the subspace spanned by the columns of the design matrix A. We saw two examples of linear models, namely the linear regression model and the quadratic regression model. What other useful regression models are linear models?

Linear Models Example

Example: Polynomial regression models seek to find the best polynomial fit to noisy data. Formally, for a polynomial model of degree *n*, each response variable Y_i is modelled as

$$Y_i = \sum_{k=0}^{n} \beta_k x_i^k + \epsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$.
This can be seen as a linear model with design matrix

$$A = \begin{pmatrix} 1 & x_i & x_i^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^n \end{pmatrix}$$

Therefore, we can fit the regression model by solving for &hat;β

Remark 1: This suggests that a sensible approach to model fitting is to successively increase model complexity until the variability seen in data is well explained by the model – and the remaining variability is consistent with our assumptions on the error model.
Remark 2: However, one pitfall is overfitting, where a more complex model will always fit data better than a simpler model nested inside it. Therefore, we always seek the simplest model that fits the data well.

Coefficient of Determination

A measure of model fit is the **coefficient of determination**, which can be constructed by scaling the residuals by the sum of squared deviations from the mean of the observed responses:

$$R^2 = 1 - \frac{\|\, r \,\|^2}{\|\, y - \bar{y} \,\|^2}$$

If R^2 is close to 0, then the model does no better than a constant model set to the sample mean *&bar;y*. On the other hand, if R^2 is close to 1, then the model well explains the variability inherent in the observed responses **y**.

R2 Example

Example (Response Surface Model): Suppose we have reponse $\{y_i\}$ with two explanatory variables $x_{i,1}$ and $x_{i,2}$. We wish to model data through a two-dimensional polynomial model

$$Y_i = \sum_{j=0}^{n}\sum_{k=0}^{n} \beta_{j,k} x_{i,1}^j x_{i,2}^k + \epsilon_i$$

We can rewrite this model as a linear model, and solve for &hat;β for successively larger *n*, until the coefficient of determination R^2 is close to 1.

Residual Testing

It seems that our final model fits well – how can we test the quality of the fit? If our model consistent with observed data, then the residuals $\{r_i\}$ should be a random sample from N(0, σ^2), where typically σ^2 is not known. Note that, when the design matrix A is of dimension *N* × *K* with *N* > *K*, there are only *N* - *K* linearly independent elements of **r**. Therefore,

$$\mathbb{E}[\| R \|^2 | X = x] = \mathbb{E}\left[\sum_{i=1}^{N}(Y_i - A\hat{\beta})^2 \mid X = x\right] = (N-K)\sigma^2$$

and so an unbiased estimator of σ^2 from the residuals is

$$S_R^2 = \frac{1}{N-K}\sum_{i=1}^{N} R_i^2 = \frac{1}{N-K}\sum_{i=1}^{N}(Y_i - A\hat{\beta})^2$$

Therefore, we can perform a hypothesis test on the residuals: $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$. Under H_0, the statistic

$$T = \frac{\bar{R}}{S_R/\sqrt{N}} \sim t_{N-K}$$

where &bar;R and S_R denote the sample mean and (unbiased) standard deviation of the residuals, respectively. Thus, provided that the assumption of common variance is reasonable, the *p*-value associated with this statistic under H_0 is a measure of the quality of our error model:

$$p = 2\max\{\mathbb{P}(T > t), \mathbb{P}(T < t)\}$$

Remark: This assumption can be checked visually by examining plots of the residuals

Remark: Up to now, we have implicitly assumed that there is only one observation of the response for the same set of explanatory variables. How multiple observations change our analysis will be subject of our next set of lectures.

Summary

- Linear models; polynomial regression, response surface models, philosophy and pitfalls.
- Coefficient of determination; definition, interpretation, example.
- Residual testing; *t*-test for residuals