# Statistics Review Sheet

Michael Li

February 7, 2020

**Estimation**
Review of distribution and density functions, parametric families. Examples: binomial, Poisson, gamma. Sufficiency, minimal sufficiency, the Rao-Blackwell theorem. Maximum likelihood estimation. Confidence intervals. Use of prior distributions and Bayesian inference. [5]

**Hypothesis testing**
Simple examples of hypothesis testing, null and alternative hypothesis, critical region, size, power, type I and type II errors, Neyman-Pearson lemma. Significance level of outcome. Uniformly most powerful tests. Likelihood ratio, and use of generalised likelihood ratio to construct test statistics for composite hypotheses. Examples, including $t$-tests and $F$-tests. Relationship with confidence intervals. Goodness-of-fit tests and contingency tables. [4]

**Linear models**
Derivation and joint distribution of maximum likelihood estimators, least squares, Gauss-Markov theorem. Testing hypotheses, geometric interpretation. Examples, including simple linear regression and one-way analysis of variance. Use of software. [7]

# Contents

# 1 Statistics

## 1.1 Estimators, MSE and Sufficiency

Now let us have a probability distribution $f(x, \theta)$, where we know $f(x)$ but not $\theta$. Like we know the family of distribution but not the parameter. Then we have:

**Definition** (Statistic)**.** A *statistic* is an estimate of $\theta$. It is a function $T$ of the data. If we write the data as $\mathbf{x} = (x_1, \cdots, x_n)$, then our estimate is written as $\hat{\theta} = T(\mathbf{x})$. $T(\mathbf{X})$ is an *estimator* of $\theta$.

The distribution of $T = T(\mathbf{X})$ is the *sampling distribution* of the statistic.

We have the following criteria to know if the statistic is good or bad:

**Definition** (Bias)**.** Let $\hat{\theta} = T(\mathbf{X})$ be an estimator of $\theta$. The *bias* of $\hat{\theta}$ is the difference between its expected value and true value.

$$\text{Bias}(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta}) - \theta.$$

If 0, the estimator is *unbiased*. If the bias tends to 0 as $n \to \infty$ (number of samples increase), then we call it *asymptotically unbiased*.

**Definition** (Mean squared error (MSE))**.** The *mean squared error* of an estimator $\hat{\theta}$ is $\mathbb{E}_\theta[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) - \text{Bias}(\hat{\theta})^2$.

The last relation comes directly from expanding the definition of MSE. Now we can see that if we want the estimator to have lower variance, we may want it to have higher bias.

**Note.** Unbiased estimators are *NOT* always the best estimators. Most often they are not.

We also have the following criteria for finding a good estimator:

**Definition** (Sufficient statistic)**.** A statistic $T$ is *sufficient* for $\theta$ if the conditional distribution of $\mathbf{X}$ given $T$ does not depend on $\theta$.

What does this actually mean? This just says that if we know $T$, we know everything about $\theta$. All information about $\theta$ can be seen as information for $T$, which is really desirable (and needed). Using this, we have the following theorem:

**Theorem** (The factorization criterion)**.** $T$ is sufficient for $\theta$ if and only if

$$f_{\mathbf{X}}(\mathbf{x} \mid \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$$

for some functions $g$ and $h$.

*Discrete.* Suppose $f_{\mathbf{X}}(\mathbf{x} \mid \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$. If $T(\mathbf{x}) = t$, then

$$f_{\mathbf{X}|T=t}(\mathbf{x}) = \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t)}{\mathbb{P}_\theta(T = t)} = \frac{g(T(\mathbf{x}), \theta)h(\mathbf{x})}{\sum_{\{\mathbf{x}:T(\mathbf{x})=t\}} g(T(\mathbf{x}), \theta)h(\mathbf{x})} = \frac{g(t, \theta)h(\mathbf{x})}{g(t, \theta)\sum h(\mathbf{x})} = \frac{h(\mathbf{x})}{\sum h(\mathbf{x})}$$

which doesn't depend on $\theta$. So $T$ is sufficient.

Now suppose $T$ is sufficient so that the conditional distribution of $\mathbf{X} \mid T = t$ does not depend on $\theta$. Then

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

The first factor does not depend on $\theta$ by assumption; call it $h(\mathbf{x})$. Let the second factor be $g(t, \theta)$, and so we have the required factorization. □

From the formula we can see that any invertible function of a sufficient estimator is sufficient. But we want the one that contains all the information of $\theta$ but minimal information needed for $T$, which brings:

**Definition** (Minimal sufficiency)**.** A sufficient statistic $T(\mathbf{X})$ is *minimal* if it is a function of every other sufficient statistic, ie. if $T'(\mathbf{X})$ is also sufficient, then $T'(\mathbf{X}) = T'(\mathbf{Y}) \Rightarrow T(\mathbf{X}) = T(\mathbf{Y})$.

Again, we can find minimal statistics easily:

**Theorem.** Suppose $T = T(\mathbf{X})$ is a statistic that satisfies

$$\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{y};\theta)} \text{ is a constant as a function of } \theta \text{ if and only if } T(\mathbf{x}) = T(\mathbf{y}).$$

Then $T$ is minimal sufficient for $\theta$.

*Proof.* First we have to show sufficiency. We will use the factorization criterion to do so.

Firstly, for each possible $t$, pick a favorite $\mathbf{x}_t$ such that $T(\mathbf{x}_t) = t$.

Now let $\mathbf{x} \in \mathcal{X}^N$ and let $T(\mathbf{x}) = t$. So $T(\mathbf{x}) = T(\mathbf{x}_t)$. By the hypothesis, $\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{x}_t;\theta)}$ does not depend on $\theta$. Let this be $h(\mathbf{x})$. Let $g(t,\theta) = f_{\mathbf{X}}(\mathbf{x}_t,\theta)$. Then

$$f_{\mathbf{X}}(\mathbf{x};\theta) = f_{\mathbf{X}}(\mathbf{x}_t;\theta)\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{x}_t;\theta)} = g(t,\theta)h(\mathbf{x}).$$

So $T$ is sufficient for $\theta$.

To show that this is minimal, suppose that $S(\mathbf{X})$ is also sufficient. By the factorization criterion, there exist functions $g_S$ and $h_S$ such that

$$f_{\mathbf{X}}(\mathbf{x};\theta) = g_S(S(\mathbf{x}),\theta)h_S(\mathbf{x}).$$

Now suppose that $S(\mathbf{x}) = S(\mathbf{y})$. Then

$$\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{y};\theta)} = \frac{g_S(S(\mathbf{x}),\theta)h_S(\mathbf{x})}{g_S(S(\mathbf{y}),\theta)h_S(\mathbf{y})} = \frac{h_S(\mathbf{x})}{h_S(\mathbf{y})}.$$

This means that the ratio $\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{y};\theta)}$ does not depend on $\theta$. By the hypothesis, this implies that $T(\mathbf{x}) = T(\mathbf{y})$. So we know that $S(\mathbf{x}) = S(\mathbf{y})$ implies $T(\mathbf{x}) = T(\mathbf{y})$. So $T$ is a function of $S$. So $T$ is minimal sufficient. □

Below is an often useful and handy example:

**Example.** Suppose $X_1, \cdots, X_n$ are iid $N(\mu,\sigma^2)$. Then

$$\begin{aligned}\frac{f_{\mathbf{X}}(\mathbf{x} \mid \mu,\sigma^2)}{f_{\mathbf{X}}(\mathbf{y} \mid \mu,\sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right\}}{(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_i (y_i-\mu)^2\right\}} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_i x_i^2 - \sum_i y_i^2\right) + \frac{\mu}{\sigma^2}\left(\sum_i x_i - \sum_i y_i\right)\right\}.\end{aligned}$$

This is a constant function of $(\mu,\sigma^2)$ iff $\sum_i x_i^2 = \sum_i y_i^2$ and $\sum_i x_i = \sum_i y_i$. So $T(\mathbf{X}) = (\sum_i X_i^2, \sum_i X_i)$ is minimal sufficient for $(\mu,\sigma^2)$.

Now, since this is *minimal sufficient*, we can use this to improve any estimator:

**Theorem** (Rao-Blackwell Theorem)**.** Let $T$ be a sufficient statistic for $\theta$ and let $\tilde{\theta}$ be an estimator for $\theta$ with $\mathbb{E}(\tilde{\theta}^2) < \infty$ for all $\theta$. Let $\hat{\theta}(\mathbf{x}) = \mathbb{E}[\tilde{\theta}(\mathbf{X}) \mid T(\mathbf{X}) = T(\mathbf{x})]$. Then for all $\theta$,

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \leq \mathbb{E}[(\tilde{\theta} - \theta)^2].$$

The inequality is strict unless $\tilde{\theta}$ is a function of $T$.

*Proof.* By the conditional expectation formula, we have $\mathbb{E}(\hat{\theta}) = \mathbb{E}[\mathbb{E}(\tilde{\theta} \mid T)] = E(\tilde{\theta})$. So they have the same bias.

By the conditional variance formula,

$$\mathrm{Var}(\tilde{\theta}) = \mathbb{E}[\mathrm{Var}(\tilde{\theta} \mid T)] + \mathrm{Var}[\mathbb{E}(\tilde{\theta} \mid T)] = \mathbb{E}[\mathrm{Var}(\tilde{\theta} \mid T)] + \mathrm{Var}(\hat{\theta}).$$

Hence $\mathrm{Var}(\tilde{\theta}) \geq \mathrm{Var}(\hat{\theta})$. So $\mathrm{MSE}(\tilde{\theta}) \geq \mathrm{MSE}(\hat{\theta})$, with equality only if $\mathrm{Var}(\tilde{\theta} \mid T) = 0$. □

**Note.** During the test, if it asks you to find an estimator that has lower variance/"variance of *some specific expression*", the first try is always Rao-Blackwell.

## 1.2 Likelihood and Confidence Intervals

**Definition** (Likelihood)**.** For any given $\mathbf{x}$, the *likelihood* of $\theta$ is $\mathrm{like}(\theta) = f_{\mathbf{X}}(\mathbf{x} \mid \theta)$, regarded as a function of $\theta$. The *maximum likelihood estimator* (mle) of $\theta$ is an estimator that picks the value of $\theta$ that maximizes $\mathrm{like}(\theta)$.

**Note.** What you should really note is:

- The likelihood is really just the probability for getting the data if $\theta = \theta$. Like any value.
- We find the MLE by taking the derivative of $\mathrm{loglike}(\theta)$ and setting that to 0 because it is usually easier. Second derivative check is not necessary unless it is actually required.

Now we have another important concept:

**Definition.** A $100\gamma\%$ $(0 < \gamma < 1)$ *confidence interval* for $\theta$ is a random interval $(A(\mathbf{X}), B(\mathbf{X}))$ such that $\mathbb{P}(A(\mathbf{X}) < \theta < B(\mathbf{X})) = \gamma$, no matter what the true value of $\theta$ may be.

The main takeaway for Confidence intervals are:

- Note the boundary is a random quantity. So we are not just randomly writing down a *fixed* probability expression and says $\gamma$ has to be in there.
- The method for calculating confidence interval is:
  - Find a function of $\mathbf{X}$ and $\theta$, so that $R(mbX, \theta)$ has a probability distribution *not* dependent on anything random. The confidence interval just says that "if we keep making these intervals, $\gamma\%$ of them will contain $\theta$".
  - Find the appropriate bound so that $P(a < R(\mathbf{X}, \theta) < b) = \gamma$.
  - Rearrange to get it out.

## 1.3 Bayesian Estimation

Now this is where everyone gets confused. WHat is Bayesian inference? It is basically two things:

(i) You have a *prior distribution* in which you assume the value of some statistic $\theta$ to take before you do any testing. This is like knowing by experience the chance it would rain tomorrow is $30\%$.

(ii) After testing you have some values of $\mathbf{X}$ to test it. For example, I tested that out of 10 days like this, only 2 days had a raining next day. Then I should update this information into my distribution for $\theta$, the random variable indicating raining (1) or not (0).

By conditional probability, if we denote $\pi(\theta)$ as the *prior* and $\pi(\theta|\mathbf{x})$ as the posterior, we have that:

$$\pi(\theta) = \frac{f_X(\mathbf{x}|\theta)\pi(\theta)}{f_X(x)}$$

The concept of Bayesian inference is intuitive if you understand this is just using information to update probabilities. Like the example below:

**Example.** Suppose I have 3 coins in my pocket. One is $3:1$ in favour of tails, one is a fair coin, and one is $3:1$ in favour of heads.

I randomly select one coin and flip it once, observing a head. What is the probability that I have chosen coin 3?

Let $X = 1$ denote the event that I observe a head, $X = 0$ if a tail. Let $\theta$ denote the probability of a head. So $\theta$ is either 0.25, 0.5 or 0.75.

Our prior distribution is $\theta(\theta = 0.25) = \pi(\theta = 0.5) = \pi(\theta = 0.75) = 1/3$.

The probability mass function $f_X(x \mid \theta) = \theta^x(1-\theta)^{1-x}$. So we have to following results:

| $\theta$ | $\pi(\theta)$ | $f_X(x=1 \mid \theta)$ | $f_X(x=1 \mid \theta)\pi(\theta)$ | $\pi(x)$ |
|---|---|---|---|---|
| 0.25 | 0.33 | 0.25 | 0.0825 | 0.167 |
| 0.50 | 0.33 | 0.50 | 0.1650 | 0.333 |
| 0.75 | 0.33 | 0.75 | 0.2475 | 0.500 |
| Sum | 1.00 | 1.50 | 0.4950 | 1.000 |

So if we observe a head, then there is now a $50\%$ chance that we have picked the third coin.

So now we have a posterior distribution. But what about the estimator? We surely need an estimator, and we need some criteria for it:

**Definition** (Bayes estimator)**.** The *Bayes estimator* $\hat{\theta}$ is the estimator that minimises the expected posterior loss:

$$h(a) = \int L(\theta, a)\pi(\theta \mid \mathbf{x})\, \mathrm{d}\theta.$$

The posterior loss is the "loss" incurred by estimating the value to be $a$ when it is actually $\theta$. But how do we quantify the loss? Common functions include quadratic loss ($L(\theta, a) = (\theta - a)^2$) and absolute loss ($L(\theta, a) = |\theta - a|$).

# 2 Hypothesis Testing

## 2.1 Simple Hypotheses

Now in real life, we often want to compare some theories or hypotheses about the value of the unknown parameter $\theta$. We do this by setting a *null hypothesis* (the hypothesis that we test against), against an *alternative hypothesis* (a hypothesis that we test for). We would "reject" the null hypothesis if the data doesn't look like its from the null one. But first, some definitions:

**Definition** (Simple and composite hypotheses)**.** A *simple hypothesis* $H$ specifies $f$ completely (eg. $H_0 : \theta = \frac{1}{2}$). Otherwise, $H$ is a *composite hypothesis*.

**Definition** (Critical/Acceptance region)**.** For testing $H_0$ against an alternative hypothesis $H_1$, a test procedure has to partition $\mathcal{X}^n$ into two disjoint exhaustive regions $C$ and $\bar{C}$, such that if $\mathbf{x} \in C$, then $H_0$ is rejected, and if $\mathbf{x} \in \bar{C}$, then $H_0$ is not rejected. $C$ is the *critical region*. The complement of the critical region is called the *acceptance region*.

**Definition** (Type I and II error, $p$ value)**.**

(i) *Type I error/False Positive*: reject $H_0$ when $H_0$ is true. The probability of a Type I error is called $\alpha$, the *size* of the test. *$p$-value* is the actual possibility of getting a Type I error, and we compare it to our criteria, which is $\alpha$.

(ii) *Type II error/False Negative*: not rejecting $H_0$ when $H_0$ is false. The probability of a Type II Error is denoted $\beta$, where $1 - \beta$ is the power of the test.

As the alternative hypothesis is usually the more interesting one, we really care about false positives (the ones that gives us false hopes) more than false negatives.

Now we can finally define our likelihood test, which is:

**Definition** (Likelihood Ratio Test)**.** The *likelihood ratio* of two simple hypotheses $H_0, H_1$ given data $\mathbf{x}$ is

$$\Lambda_{\mathbf{x}}(H_0; H_1) = \frac{L_{\mathbf{x}}(H_1)}{L_{\mathbf{x}}(H_0)}.$$

A *likelihood ratio test* (LR test) is one where the critical region $C$ is, for some $k$:

$$C = \{\mathbf{x} : \Lambda_{\mathbf{x}}(H_0; H_1) > k\}$$

As we said, we care about false positives more, so we want to control false positive to some $\alpha$, and then find the one with the largest power. Turns out the simple one above is the best one:

**Lemma** (Neyman-Pearson lemma)**.** Suppose $H_0 : f = f_0$, $H_1 : f = f_1$, where $f_0$ and $f_1$ are continuous densities nonzero on the same regions. Then among all tests of size $\leq \alpha$, the test with the largest power is the likelihood ratio test of size $\alpha$.

Also, one can see this easily expands to tests of $H_0 : \theta = a$ against $H_0 : \theta > a$, because the critical region does not depend on $f_1$!

*Proof.* Under the likelihood ratio test, our critical region is

$$C = \left\{\mathbf{x} : \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} > k\right\},$$

where $k$ is chosen such that $\alpha = \mathbb{P}(\text{reject } H_0 \mid H_0) = \mathbb{P}(\mathbf{X} \in C \mid H_0) = \int_C f_0(\mathbf{x})\,\mathrm{d}\mathbf{x}$. The probability of Type II error is given by

$$\beta = \mathbb{P}(\mathbf{X} \notin C \mid f_1) = \int_{\bar{C}} f_1(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

Let $C^*$ be the critical region of any other test with size less than or equal to $\alpha$. Let $\alpha^* = \mathbb{P}(X \in C^* \mid f_0)$ and $\beta^* = \mathbb{P}(\mathbf{X} \notin C^* \mid f_1)$. We want to show $\beta \leq \beta^*$.

We know $\alpha^* \leq \alpha$, ie

$$\int_{C^*} f_0(\mathbf{x})\,\mathrm{d}\mathbf{x} \leq \int_C f_0(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

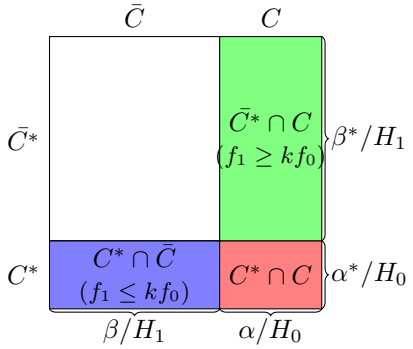
Also, on $C$, we have $f_1(\mathbf{x}) > kf_0(\mathbf{x})$, while on $\bar{C}$ we have $f_1(\mathbf{x}) \leq kf_0(\mathbf{x})$. So

$$\begin{aligned}
\int_{\bar{C}^* \cap C} f_1(\mathbf{x})\,\mathrm{d}\mathbf{x} &\geq k \int_{\bar{C}^* \cap C} f_0(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
\int_{\bar{C} \cap C^*} f_1(\mathbf{x})\,\mathrm{d}\mathbf{x} &\leq k \int_{\bar{C} \cap C^*} f_0(\mathbf{x})\,\mathrm{d}\mathbf{x}.
\end{aligned}$$

Hence

$$\begin{aligned}
\beta - \beta^* &= \int_{\bar{C}} f_1(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int_{\bar{C}^*} f_1(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \int_{\bar{C} \cap C^*} f_1(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int_{\bar{C} \cap \bar{C}^*} f_1(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int_{\bar{C}^* \cap C} f_1(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int_{\bar{C} \cap \bar{C}^*} f_1(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \int_{\bar{C} \cap C^*} f_1(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int_{\bar{C}^* \cap C} f_1(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&\leq k \int_{\bar{C} \cap C^*} f_0(\mathbf{x})\,\mathrm{d}\mathbf{x} - k \int_{\bar{C}^* \cap C} f_0(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= k \left\{ \int_{\bar{C} \cap C^*} f_0(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int_{C \cap C^*} f_0(\mathbf{x})\,\mathrm{d}\mathbf{x} \right\} d - k \left\{ \int_{\bar{C}^* \cap C} f_0(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int_{C \cap C^*} f_0(\mathbf{x})\,\mathrm{d}\mathbf{x} \right\} \\
&= k(\alpha^* - \alpha) \\
&\leq 0.
\end{aligned}$$

Courtesy of Dexter to provide the following illuminating graph:




□

**Note.** Remember this proof! This has appeared multiple times in the test and may as well appear in the future.This is really just algebraic manipulation of the sectors drawn above. Good way of remembering it:

- The difference of power, which is $f_1$ under $\bar{C} \times (\bar{C}^* + C^*)$ minus $f_1$ under $\bar{C}^* \times (\bar{C} + C)$ is just $f_1$ under blue minus $f_1$ under green.
- Now due to the inequalities under green and blue, this is greater or equal than $k$ times $f_0$ under blue minus $f_0$ under green.
- Now we add blue to red and green to red. Blue+red is $\alpha^*$ under $f_0$ and green+red is $\alpha$ under $f_0$. So their difference, by hypothesis, is assumed to be negative. Done!

## 2.2 Composite Hypotheses

If only life was just simple…. To evaluate composite hypotheses like those that bound $\theta$ by an inequality, we need some more definitions:

**Definition** (Power function)**.** The *power function* is

$$W(\theta) = \mathbb{P}(\mathbf{X} \in C \mid \theta) = \mathbb{P}(\text{reject } H_0 \mid \theta),$$

**Definition** (Size)**.** The *size* of the test is

$$\alpha = \sup_{\theta \in \Theta_0} W(\theta),$$

**Definition** (Uniformly most powerful test)**.** A test specified by a critical region $C$ is *uniformly most powerful* (UMP) size $\alpha$ test for test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ if

(i) $\sup_{\theta \in \Theta_0} W(\theta) = \alpha$.

(ii) For any other test $C^*$ with size $\leq \alpha$ and with power function $W^*$, we have $W(\theta) \geq W^*(\theta)$ for all $\theta \in \Theta_1$.

Note that these may not exist.

**Note.** This, in layman terms, is really:

- The worst size we can get is $\alpha$.
- The power *function* is larger than any other possible power function for *all* values. Hence uniformly.

Now we can still similarly define likelihood of a composite hypothesis:

**Definition** (Likelihood of a composite hypothesis)**.** The *likelihood* of a composite hypothesis $H : \theta \in \Theta$ given data $\mathbf{x}$ to be

$$L_{\mathbf{x}}(H) = \sup_{\theta \in \Theta} f(\mathbf{x} \mid \theta).$$

And the ratio test is still defined as the ratio of the likelihoods! Now we introduce a powerful asymptotic theorem, that (fortunately) would not be proven here:

**Theorem** (Generalized likelihood ratio theorem)**.** Suppose $\Theta_0 \subseteq \Theta_1$ and $|\Theta_1| - |\Theta_0| = p$. Let $\mathbf{X} = (X_1, \cdots, X_n)$ with all $X_i$ iid. Then if $H_0$ is true, as $n \to \infty$,

$$2 \log \Lambda_{\mathbf{X}}(H_0 : H_1) \sim \chi^2_p.$$

If $H_0$ is not true, then $2 \log \Lambda$ tends to be larger. We reject $H_0$ if $2 \log \Lambda > c$, where $c = \chi^2_p(\alpha)$ for a test of approximately size $\alpha$.

Now what is $|\Theta_1|$? This is basically the number of parameters that you, as the conductor of the test, has control over. Like if your alternative hypothesis is that the probability follows a binomial distribution of size $n$, then you have one parameter (the probability $p$) in which you have control over. Why basically? Very technically this is not precise. But this is beyond the course.

## 2.3 Tests of Goodness-of-Fit and Independence

The test of Goodness-of-Fit is just using the likelihood ratio test to test against the approximate chi-squared values. What's more interesting is the Pearson Chi-squared test. Now consider a set of data in which the sum of the observed/expected is fixed. Then in general if the proportionality is a product of powers of the individually observed data, then we have $2 \log \Lambda = 2 \sum o_i \log \left( \frac{o_i}{e_i} \right)$. Now we approximate the statistic $2 \log \Lambda$ in the generalized likelihood ratio theorem:

$$\begin{aligned} 2 \log \Lambda = 2 \sum o_i \log \left( \frac{o_i}{e_i} \right) &= 2 \sum (e_i + \delta_i) \log \left( 1 + \frac{\delta_i}{e_i} \right) \\ &= 2 \sum (e_i + \delta_i) \left( \frac{\delta_i}{e_i} - \frac{\delta_i^2}{2e_i^2} + O(\delta_i^3) \right) \\ &= 2 \sum \left( \delta_i + \frac{\delta_i^2}{e_i} - \frac{\delta_i^2}{2e_i} + O(\delta_i^3) \right) \end{aligned}$$

We know that $\sum \delta_i = 0$ since $\sum e_i = \sum o_i$. So

$$\approx \sum \frac{\delta_i^2}{e_i} = \sum \frac{(o_i - e_i)^2}{e_i}.$$

This is known as the *Pearson's Chi-squared test.* Let's look at an example:

**Example.** Mendel crossed 556 smooth yellow male peas with wrinkled green peas. From the progeny, let

(i) $N_1$ be the number of smooth yellow peas,

(ii) $N_2$ be the number of smooth green peas,

(iii) $N_3$ be the number of wrinkled yellow peas,

(iv) $N_4$ be the number of wrinkled green peas.

We wish to test the goodness of fit of the model

$$H_0 : (p_1, p_2, p_3, p_4) = \left( \tfrac{9}{16}, \tfrac{3}{16}, \tfrac{3}{16}, \tfrac{1}{16} \right).$$

Suppose we observe $(n_1, n_2, n_3, n_4) = (315, 108, 102, 31)$.

We find $(e_1, e_2, e_3, e_4) = (312.75, 104.25, 104.25, 34.75)$. The actual $2\log\Lambda = 0.618$ and the approximation we had is $\sum \frac{(o_i - e_i)^2}{e_i} = 0.604$.

Here $|\Theta_0| = 0$ and $|\Theta_1| = 4 - 1 = 3$. So we refer to test statistics $\chi^2_3(\alpha)$.

Since $\chi^2_3(0.05) = 7.815$, we see that neither value is significant at $5\%$. So there is no evidence against Mendel's theory. In fact, $\mathbb{P}(\chi^2_3 > 0.6) \approx 0.96$.

## 2.4 Testing Independence in Contingency Tables

There is really nothing new in contingency tables. This is just a matrix of data with constraints on each row and each column, so the $\chi^2$ statistic is tested on $(m-1) \times (n-1)$ d.f.. We can work the likelihood ratio test:

### 2.4.1 Derivation of the Test

Consider a two-way contingency table with $r$ rows and $c$ columns. For $i = 1, \cdots, r$ And $j = 1, \cdots, c$, let $p_{ij}$ be the probability that an individual selected from the population under consideration is classified in row $i$ and column $j$. (ie. in the $(i, j)$ cell of the table).

Let $p_{i+} = \mathbb{P}(\text{in row } i)$ and $p_{+j} = \mathbb{P}(\text{in column } j)$. Then we must have $p_{++} = \sum_i \sum_j p_{ij} = 1$.

Suppose a random sample of $n$ individuals is taken, and let $n_{ij}$ be the number of these classified in the $(i, j)$ cell of the table.

Let $n_{i+} = \sum_j n_{ij}$ and $n_{+j} = \sum_i n_{ij}$. So $n_{++} = n$.

We have

$$(N_{11}, \cdots, N_{1c}, N_{21}, \cdots, N_{rc}) \sim \text{Multinomial}(n; p_{11}, \cdots, p_{1c}, p_{21}, \cdots, p_{rc}).$$

We may be interested in testing the null hypothesis that the two classifications are independent. So we test

- $H_0$: $p_i = p_{i+}p_{+j}$ for all $i, j$, ie. independence of columns and rows.
- $H_1$: $p_{ij}$ are unrestricted.

Of course we have the usual restrictions like $p_{++} = 1, p_{ij} \geq 0$.

Under $H_1$, the MLEs are $\hat{p_{ij}} = \frac{n_{ij}}{n}$.

Under $H_0$, the MLEs are $\hat{p}_{i+} = \frac{n_{i+}}{n}$ and $\hat{p}_{+j} = \frac{n_{+j}}{n}$.

Write $o_{ij} = n_{ij}$ and $e_{ij} = n\hat{p}_{i+}\hat{p}_{+j} = n_{i+}n_{+j}/n$.

Then

$$2\log\Lambda = 2\sum_{i=1}^{r}\sum_{j=1}^{c} o_{ij}\log\left(\frac{o_{ij}}{e_{ij}}\right) \approx \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

This is really just best illustrated by an example:

**Example.** We wish to test $H_0$: the new and previous car sizes are independent. The data is:

| | | New car | | | |
|---|---|---|---|---|---|
| | | Large | Medium | Small | **Total** |
| Previous car | Large | 56 | 52 | 42 | **150** |
| | Medium | 50 | 83 | 67 | **120** |
| | Small | 18 | 51 | 81 | **150** |
| | *Total* | *124* | *186* | *190* | ***500*** |

while the expected values given by $H_0$ is

| | | New car | | | |
|---|---|---|---|---|---|
| | | Large | Medium | Small | **Total** |
| Previous car | Large | 37.2 | 55.8 | 57.0 | **150** |
| | Medium | 49.6 | 74.4 | 76.0 | **120** |
| | Small | 37.2 | 55.8 | 57.0 | **150** |
| | *Total* | *124* | *186* | *190* | ***500*** |

Note the margins are the same. It is quite clear that they do not match well, but we can find the $p$ value to be sure.

$\sum\sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 36.20$, and the degrees of freedom is $(3-1)(3-1) = 4$.

From the tables, $\chi^2_4(0.05) = 9.488$ and $\chi^2_4(0.01) = 13.28$.

So our observed value of 36.20 is significant at the $1\%$ level, ie. there is strong evidence against $H_0$. So we conclude that the new and previous car sizes are not independent.

### 2.4.2 Tests of Homogeneity

A test of homogeneity is usually a test on the if two groups are the same. The test statistic follows exactly the same as the case above, although with the added restriction that the row totals are fixed to be the same. This does not change the likelihood ratio test.

## 2.5 Confidence Intervals and Hypothesis Tests

We have the following theorem connecting hypothesis tests and confidence regions:

**Theorem.**

(i) Suppose that for every $\theta_0 \in \Theta$ there is a size $\alpha$ test of $H_0 : \theta = \theta_0$. Denote the acceptance region by $A(\theta_0)$. Then the set $I(\mathbf{X}) = \{\theta : \mathbf{X} \in A(\theta)\}$ is a $100(1-\alpha)\%$ confidence set for $\theta$.

(ii) Suppose $I(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence set for $\theta$. Then $A(\theta_0) = \{\mathbf{X} : \theta_0 \in I(\mathbf{X})\}$ is an acceptance region for a size $\alpha$ test of $H_0 : \theta = \theta_0$.

Put plainly, this just says that if you have a size $\alpha$ confidence interval for $\theta$, to check if $\theta = \theta_0$ holds at size $\alpha$, you just need to check if $\theta_0$ is in the confidence interval!

*Proof.* First note that $\theta_0 \in I(\mathbf{X})$ iff $\mathbf{X} \in A(\theta_0)$.

For (i), since the test is size $\alpha$, we have

$$\mathbb{P}(\text{accept } H_0 \mid H_0 \text{ is true}) = \mathbb{P}(\mathbf{X} \in A_(\theta_0) \mid \theta = \theta_0) = 1 - \alpha.$$

And so

$$\mathbb{P}(\theta_0 \in I(\mathbf{X}) \mid \theta = \theta_0) = \mathbb{P}(\mathbf{X} \in A(\theta_0) \mid \theta = \theta_0) = 1 - \alpha.$$

For (ii), since $I(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence set, we have

$$P(\theta_0 \in I(\mathbf{X}) \mid \theta = \theta_0) = 1 - \alpha.$$

So

$$\mathbb{P}(\mathbf{X} \in A(\theta_0) \mid \theta = \theta_0) = \mathbb{P}(\theta \in I(\mathbf{X}) \mid \theta = \theta_0) = 1 - \alpha.$$

□

## 2.6 Multivariate Normal Theory

Oohhhh, The scary stuff. Not really. We define everything in the same way, except that the random variable is now a vector, so a linear operation on it now is a matrix. Basically a dimension is added to everything:

**Definition** (Multivariate normal distribution)**.** $\mathbf{X}$ has a *multivariate normal distribution* if, for every $\mathbf{t} \in \mathbb{R}^n$, the random variable $\mathbf{t}^T\mathbf{X}$ (ie. $\mathbf{t} \cdot \mathbf{X}$) has a normal distribution. If $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$ and $\mathrm{Cov}(\mathbf{X}) = \Sigma$, we write $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$.

Now here is a review of some properties of multivariate distributions (1,3 is really just from definition, and we do not need to prove 4):

**Proposition.**

(i) If $\mathbf{X} \sim N_n(\mu, \Sigma)$, and $A$ is an $m \times n$ matrix, then $A\mathbf{X} \sim N_m(A\boldsymbol{\mu}, A\Sigma A^T)$.

(ii) If $\mathbf{X} \sim N_n(\mathbf{0}, \sigma^2 I)$, then

$$\frac{|\mathbf{X}|^2}{\sigma^2} = \frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} = \sum \frac{X_i^2}{\sigma^2} \sim \chi_n^2.$$

Instead of writing $|\mathbf{X}|^2/\sigma^2 \sim \chi_n^2$, we often just say $|\mathbf{X}|^2 \sim \sigma^2\chi_n^2$.

(iii) Let $\mathbf{X} \sim N_n(\mu, \Sigma)$. We split $\mathbf{X}$ up into two parts: $\mathbf{X} = \begin{pmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{pmatrix}$, where $\mathbf{X}_i$ is a $n_i \times 1$ column vector and $n_1 + n_2 = n$.

Similarly write

$$\boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{pmatrix}, \quad \Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix},$$

where $\Sigma_{ij}$ is an $n_i \times n_j$ matrix.

Then $\mathbf{X}_i \sim N_{n_i}(\boldsymbol{\mu}_i, \Sigma_{ii})$ and $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent iff $\Sigma_{12} = 0$

(iv) When $\Sigma$ is a positive definite, then $\mathbf{X}$ has pdf

$$f_{\mathbf{X}}(\mathbf{x}; \mu, \Sigma) = \frac{1}{|\Sigma|^2}\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right].$$

*Proof.*

(iii) $\mathbf{X}_i \sim N_{n_i}(\boldsymbol{\mu}_i, \Sigma_{ii})$ can be just seen by applying $(i)$ with the appropriate matrix. Now we prove the second result:

Note that by symmetry of $\Sigma$, $\Sigma_{12} = 0$ if and only if $\Sigma_{21} = 0$.

From $(\dagger)$, $M_{\mathbf{X}}(\mathbf{t}) = \exp(\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t})$ for each $\mathbf{t} \in \mathbb{R}^{\aleph}$. We write $\mathbf{t} = \begin{pmatrix}\mathbf{t}_1 \\ \mathbf{t}_2\end{pmatrix}$.

Then the mgf is equal to

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t}_1^T\boldsymbol{\mu}_1 + \mathbf{t}_2^T\Sigma_{11}\mathbf{t}_1 + \frac{1}{2}\mathbf{t}_2^T\Sigma_{22}\mathbf{t}_2 + \frac{1}{2}\mathbf{t}_1^T\Sigma_{12}\mathbf{t}_2 + \frac{1}{2}\mathbf{t}_2^T\Sigma_{21}\mathbf{t}_1\right).$$

From (i), we know that $M_{\mathbf{X}_i}(\mathbf{t}_i) = \exp(\mathbf{t}_i^T\boldsymbol{\mu}_i + \frac{1}{2}\mathbf{t}_i^T\Sigma_{ii}\mathbf{t}_i)$. So $M_{\mathbf{X}}(\mathbf{t}) = M_{\mathbf{X}_1}(\mathbf{t}_1)M_{\mathbf{X}_2}(\mathbf{t}_2)$ for all $\mathbf{t}$ if and only if $\Sigma_{12} = 0$.

□

### 2.6.1 Normal Random Samples

This result is very useful, especially the fact that $\bar{X}$ and $S_{XX}$ are independent.

**Theorem** (Joint distribution of $\bar{X}$ and $S_{XX}$)**.** Suppose $X_1, \cdots, X_n$ are iid $N(\mu, \sigma^2)$ and $\bar{X} = \frac{1}{n}\sum X_i$, and $S_{XX} = \sum(X_i - \bar{X})^2$. Then

(i) $\bar{X} \sim N(\mu, \sigma^2/n)$

(ii) $S_{XX}/\sigma^2 \sim \chi^2_{n-1}$.

(iii) $\bar{X}$ and $S_{XX}$ are independent.

*Proof.* We can write the joint density as $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \sigma^2 I)$, where $\boldsymbol{\mu} = (\mu, \mu, \cdots, \mu)$.

Let $A$ be an $n \times n$ orthogonal matrix with the first row all $1/\sqrt{n}$ (orthogonal matrices have $\frac{n(n-1)}{2}$ dimensions so this is possible). Now define $\mathbf{Y} = A\mathbf{X}$. Then

$$\mathbf{Y} \sim N_n(A\boldsymbol{\mu}, A\sigma^2 I A^T) = N_n(A\boldsymbol{\mu}, \sigma^2 I).$$

We have

$$A\boldsymbol{\mu} = (\sqrt{n}\mu, 0, \cdots, 0)^T.$$

So $Y_1 \sim N(\sqrt{n}\mu, \sigma^2)$ and $Y_i \sim N(0, \sigma^2)$ for $i = 2, \cdots, n$. Also, $Y_1, \cdots, Y_n$ are independent, since the covariance matrix has every non-diagonal term 0 (remember orthogonality).

But from the definition of $A$, we have

$$Y_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i = \sqrt{n}\bar{X}.$$

So $\sqrt{n}\bar{X} \sim N(\sqrt{n}\mu, \sigma^2)$, or $\bar{X} \sim N(\mu, \sigma^2/n)$. Also

$$\begin{aligned} Y_2^2 + \cdots + Y_n^2 &= \mathbf{Y}^T\mathbf{Y} - Y_1^2 \\ &= \mathbf{X}^T A^T A\mathbf{X} - Y_1^2 \\ &= \mathbf{X}^T\mathbf{X} - n\bar{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = S_{XX}. \end{aligned}$$

So $S_{XX} = Y_2^2 + \cdots + Y_n^2 \sim \sigma^2\chi^2_{n-1}$.

Finally, since $Y_1$ and $Y_2, \cdots, Y_n$ are independent, so are $\bar{X}$ and $S_{XX}$. □

### 2.7 Student's $t$ Distribution

Now we need a motivation to introduce a new distribution. It is to solve the following question: Let $X_1, \cdots, X_n$ be iid $\mathbb{N}(\mu, \sigma^2)$. Then $\bar{X} \sim N(\mu, \sigma^2/n)$. So $Z = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$.

Also, $S_{XX}/\sigma^2 \sim \chi^2_{n-1}$ and is independent of $\bar{X}$, and hence $Z$. So

$$\frac{\sqrt{n}(\bar{X}-\mu)/\sigma}{\sqrt{S_{XX}/((n-1)\sigma^2)}} \sim \frac{Z}{\sqrt{Y/(n-1)}}$$

Where $Z \sim N(0,1)$ and $Y \sim \chi^2_{n-1}$. Now we define this distribution as $t_{n-1}$, and here are some properties for the distribution:

**Proposition.** If $k > 1$, then $\mathbb{E}_k(T) = 0$.

If $k > 2$, then $\mathrm{Var}_k(T) = \frac{k}{k-2}$.

If $k = 2$, then $\mathrm{Var}_k(T) = \infty$.

In all other cases, the values are undefined. In particular, the $k = 1$ case, this is known as the Cauchy distribution, and has undefined mean and variance.

# 3 Linear Models

## 3.1 Introduction

Linear models is a method of fitting some explanatory variables using linear combinations. The form is:

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i. \qquad (*)$$

for $i = 1, \cdots, n$. Here

- $\beta_1, \cdots, \beta_p$ are unknown, fixed parameters we wish to work out (with $n > p$)
- $x_{i1}, \cdots, x_{ip}$ are the values of the $p$ covariates for the $i$th response (which are all known). The covariates are the explanatory variables.
- $\varepsilon_1, \cdots, \varepsilon_n$ are independent/uncorrelated random variables with mean 0 and variance $\sigma^2$. This just represent noise.

Now clearly we can write them down in matrix form, which turns the equation to:

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

With $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{Y}) = \sigma^2 I$ and:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np}. \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Similar to how a Bayesian estimator can minimize quadratic loss, we define the least squares estimator:

**Definition** (Least squares estimator)**.** In a linear model $\mathbf{Y} = X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, the *least squares estimator* $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ minimizes

$$S(\boldsymbol{\beta}) = \|\mathbf{Y} - X\boldsymbol{\beta}\|^2 = (\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta}) = \sum_{i=1}^{n}(Y_i - x_{ij}\beta_j)^2$$

with implicit summation over $j$. Alternatively, this means it minimizes the sum of squares of the distances from the fitted line to the points on the graph.

Now taking the derivative and equating it equal to 0, it is easy to get that:

**Proposition.** The least squares estimator satisfies

$$X^TX\hat{\boldsymbol{\beta}} = X^T\mathbf{Y} \qquad \Rightarrow \qquad \hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{Y}$$

The second implication follows from the fact that $X^TX$ is positive definite as $X$ has full rank. Also, we have:

- $\mathbb{E}(\hat{\boldsymbol{\beta}}) = (X^TX^{-1})X^T\mathbb{E}[\mathbf{Y}] = (X^TX)^{-1}X^TX\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\operatorname{Cov}(\hat{\boldsymbol{\beta}}) = (X^TX)^{-1}X^T\operatorname{Cov}(\mathbf{Y})X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$

But how do we know it is actually good? The Gauss-Markov Theorem comes to the rescue:

**Theorem** (Gauss Markov theorem)**.** In a full rank linear model, let $\hat{\boldsymbol{\beta}}$ be the least squares estimator of $\boldsymbol{\beta}$ and let $\boldsymbol{\beta}^*$ be any other unbiased estimator for $\boldsymbol{\beta}$ which is linear in the $Y_i$'s. Then

$$\operatorname{Var}(\mathbf{t}^T\hat{\boldsymbol{\beta}}) \leq \operatorname{Var}(\mathbf{t}^T\boldsymbol{\beta}^*).$$

for all $\mathbf{t} \in \mathbb{R}^p$. We say that $\hat{\boldsymbol{\beta}}$ is the *best linear unbiased estimator* of $\boldsymbol{\beta}$ (BLUE).

**Note.** The following proof is really algebraic manipulation. The best way to remember it? Note that you want to relate $\operatorname{Cov}(\hat{\beta})$ to this. So manipulate until you can have $(X^TX)^{-1}$ term in your expression, which comes from the $\hat{\beta}$. So get there.

*Proof.* Since $\boldsymbol{\beta}^*$ is linear in the $Y_i$'s, $\boldsymbol{\beta}^* = A\mathbf{Y}$ for some $p \times n$ matrix $A$.

Since $\boldsymbol{\beta}^*$ is an unbiased estimator, we must have $\mathbb{E}[\boldsymbol{\beta}^*] = \boldsymbol{\beta}$. As $\beta^* = A\mathbf{Y}$, $\mathbb{E}[\boldsymbol{\beta}^*] = A\mathbb{E}[\mathbf{Y}] = AX\boldsymbol{\beta}$. So we must have $\boldsymbol{\beta} = AX\boldsymbol{\beta}$. Since this holds for any $\boldsymbol{\beta}$, we must have $AX = I_p$. Now

$$\begin{aligned}\operatorname{Cov}(\boldsymbol{\beta}^*) &= \mathbb{E}[(\boldsymbol{\beta}^* - \boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta})^T] \\ &= \mathbb{E}(AX\boldsymbol{\beta} + A\boldsymbol{\varepsilon} - \boldsymbol{\beta})(AX\boldsymbol{\beta} + A\boldsymbol{\varepsilon} - \boldsymbol{\beta})^T \\ &= \mathbb{E}(A\boldsymbol{\varepsilon}(A\boldsymbol{\varepsilon})^T) = A(\sigma^2 I)A^T = \sigma^2 AA^T.\end{aligned}$$

Now let $\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} = (A - (X^TX)^{-1}X^T)\mathbf{Y} = B\mathbf{Y}$, for some $B$. Then

$$BX = AX - (X^TX)^{-1}X^TX = I_p - I_p = 0.$$

By definition, we have $A\mathbf{Y} = B\mathbf{Y} + (X^TX)^{-1}X^T\mathbf{Y}$, and this is true for all $\mathbf{Y}$. So $A = B + (X^TX)^{-1}X^T$. Hence

$$\begin{aligned}\operatorname{Cov}(\boldsymbol{\beta}^*) &= \sigma^2 AA^T \\ &= \sigma^2(B + (X^TX)^{-1}X^T)(B + (X^TX)^{-1}X^T)^T \\ &= \sigma^2(BB^T + (X^TX)^{-1}) = \sigma^2 BB^T + \operatorname{Cov}(\hat{\boldsymbol{\beta}}).\end{aligned}$$

Note that in the second line, the cross-terms disappear since $BX = 0$.

So for any $\mathbf{t} \in \mathbb{R}^p$, we have

$$\begin{aligned}\operatorname{Var}(\mathbf{t}^T\boldsymbol{\beta}^*) &= \mathbf{t}^T \operatorname{Cov}(\boldsymbol{\beta}^*)\mathbf{t} \\ &= \mathbf{t}^T \operatorname{Cov}(\hat{\boldsymbol{\beta}})\mathbf{t} + \mathbf{t}^T BB^T\mathbf{t}\sigma^2 = \operatorname{Var}(\mathbf{t}^T\hat{\boldsymbol{\beta}}) + \sigma^2\|B^T\mathbf{t}\|^2 \quad \geq \operatorname{Var}(\mathbf{t}^T\hat{\boldsymbol{\beta}}).\end{aligned}$$

Taking $\mathbf{t} = (0, \cdots, 1, 0, \cdots, 0)^T$ with a 1 in the $i$th position, we have

$$\operatorname{Var}(\hat{\beta}_i) \leq \operatorname{Var}(\beta_i^*).$$

□

## 3.2 Simple Linear Model

As with all applied mathematics, we consider the easiest model first:

$$Y_i = a + b(x_i - \bar{x}) + \varepsilon_i.$$

Where we have re-parametrized the $x_i$ to give them zero sum. Therefore, we have:

$$X = \begin{pmatrix} 1 & (x_1 - \bar{x}) \\ \vdots & \vdots \\ 1 & (x_{24} - \bar{x}) \end{pmatrix} \qquad \hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{Y} = \begin{pmatrix} \bar{Y} \\ \frac{S_{xY}}{S_{xx}} \end{pmatrix},$$

where $S_{xY} = \sum Y_i(x_i - \bar{x})$. So the intercept is $\hat{a} = \bar{y}$ (which is intuitive), and the gradient is:

$$\begin{aligned}\hat{b} &= \frac{S_{xy}}{S_{xx}} \\ &= \frac{\sum_i y_i(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \times \sqrt{\frac{S_{yy}}{S_{xx}}} = r \times \sqrt{\frac{S_{yy}}{S_{xx}}}. \quad (*)\end{aligned}$$

Where the $(*)$ part comes from the fact that $\sum(x_i - \bar{x}) = 0$. Why do we write it in such a form? Because $r$ is called the *Pearson product-moment correlation coefficient.*

Hence from $\operatorname{Cov}(\hat{\boldsymbol{\beta}}) = (X^TX)^{-1}\sigma^2$, we have:

$$\operatorname{Var}(\hat{a}') = \operatorname{Var}(\bar{Y}) = \frac{\sigma^2}{n}, \quad \operatorname{Var}(\hat{b}) = \frac{\sigma^2}{S_{xx}}.$$

Note that these estimators are uncorrelated.

### 3.2.1 Residual Sum of Squares and Geometry

**Definition.** Let $\hat{\mathbf{Y}} = X\beta$ be the vector of fitted values. Then the residual sum of squares is just:

$$\operatorname{RSS} = \|\mathbf{R}\|^2 = \mathbf{R}^T\mathbf{R} = (\mathbf{Y} - X\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - X\hat{\boldsymbol{\beta}}).$$

Now, if we expand the definition of the fitted values, we find:$\hat{\mathbf{Y}} = X\hat{\boldsymbol{\beta}} = X(X^TX)^{-1}X^T\mathbf{Y} = P\mathbf{Y}$. Now $P$ represents an orthogonal projection of $\mathbb{R}^n$ onto the space spanned by columns of $X$. So $P$ is idempotent ($P^2 = P$) and symmetric.

## 3.3 Normal Linear Models

Let's make our lives even easier. What if they are normal?

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I), \quad \text{Rank}(X) = p < n.$$

Now we can calculate the MLEs. The log-likelihood is:

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}S(\boldsymbol{\beta}),$$

where $S(\boldsymbol{\beta}) = (\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta})$.. So the MLE of $\hat{\beta}$ is exactly the least squares estimator (as we are basically maximizing $S(\boldsymbol{\beta})$)! For $\sigma^2$ we have:

$$\hat{\sigma}^2 = \frac{1}{n}\text{RSS}.$$

Now here is the uninteresting proof: $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent. First, remember that $\hat{\boldsymbol{\beta}} = P\mathbf{Y}$. Therefore we would like to first prove properties on $P$:

**Lemma.**

(i) If $\mathbf{Z} \sim N_n(\mathbf{0}, \sigma^2 I)$ and $A$ is $n \times n$, symmetric, idempotent with rank $r$, then $\mathbf{Z}^T A\mathbf{Z} \sim \sigma^2\chi^2_r$.

(ii) For a symmetric idempotent matrix $A$, $\text{Rank}(A) = \text{tr}(A)$.

**Note.** This is basically saying

(i) This can be diagonalized and the matrix has $r$ eigenvalues of 1.

(ii) Other eigenvalues are all 0 so the number of ones, which is the trace, is the rank.

*Proof.*

(i) Since $A$ is idempotent, $A^2 = A$ by definition. So eigenvalues of $A$ are either 0 or 1 (since $\lambda\mathbf{x} = A\mathbf{x} = A^2\mathbf{x} = \lambda^2\mathbf{x}$).

Since $A$ is also symmetric, it is diagonalizable. So there exists an orthogonal $Q$ such that

$$\Lambda = Q^T A Q$$

with $r$ copies of 1 and $n - r$ copies of 0 on the diagonal.

Let $\mathbf{W} = Q^T\mathbf{Z}$. So $\mathbf{Z} = Q\mathbf{W}$. Then $\mathbf{W} \sim N_n(\mathbf{0}, \sigma^2 I)$, since $\text{Cov}(\mathbf{W}) = Q^T\sigma^2 I Q = \sigma^2 I$. Then

$$\mathbf{Z}^T A\mathbf{Z} = \mathbf{W}^T Q^T A Q\mathbf{W} = \mathbf{W}^T\Lambda\mathbf{W} = \sum_{i=1}^{r} w_i^2 \sim \chi^2_r.$$

(ii)

$$\text{Rank}(A) = \text{Rank}(\Lambda) = \text{tr}(\Lambda) = \text{tr}(Q^T A Q) = \text{tr}(AQ^T Q) = \text{tr}\, A$$

□

**Theorem.** For the normal linear model $\mathbf{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$,

(i) $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1})$

(ii) $\text{RSS} \sim \sigma^2\chi^2_{n-p}$, and so $\hat{\sigma}^2 \sim \frac{\sigma^2}{n}\chi^2_{n-p}$.

(iii) $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent.

*Proof.*

- We have $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{Y}$. Call this $C\mathbf{Y}$ for later use. Then $\hat{\boldsymbol{\beta}}$ has a normal distribution with mean

$$(X^TX)^{-1}X^T(X\boldsymbol{\beta}) = \boldsymbol{\beta}$$

and covariance

$$(X^TX)^{-1}X^T(\sigma^2 I)[(X^TX)^{-1}X^T]^T = \sigma^2(X^TX)^{-1}.$$

So

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1}).$$

- Our previous lemma says that $\mathbf{Z}^TA\mathbf{Z} \sim \sigma^2\chi^2_r$. So we pick our $\mathbf{Z}$ and $A$ so that $\mathbf{Z}^TA\mathbf{Z} = \text{RSS}$, and $r$, the degrees of freedom of $A$, is $n - p$.

Let $\mathbf{Z} = \mathbf{Y} - X\boldsymbol{\beta}$ and $A = (I_n - P)$, where $P = X(X^TX)^{-1}X^T$. We first check that the conditions of the lemma hold:

Since $\mathbf{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$, $\mathbf{Z} = \mathbf{Y} - X\boldsymbol{\beta} \sim N_n(\mathbf{0}, \sigma^2 I)$.

Since $P$ is idempotent, $I_n - P$ also is (check!). We also have

$$\text{Rank}(I_n - P) = \text{tr}(I_n - P) = n - p.$$

Therefore the conditions of the lemma hold.

To get the final useful result, we want to show that the RSS is indeed $\mathbf{Z}^TA\mathbf{Z}$. We simplify the expressions of RSS and $\mathbf{Z}^TA\mathbf{Z}$ and show that they are equal:

$$\mathbf{Z}^TA\mathbf{Z} = (\mathbf{Y} - X\boldsymbol{\beta})^T(I_n - P)(\mathbf{Y} - X\boldsymbol{\beta}) = \mathbf{Y}^T(I_n - P)\mathbf{Y}.$$

Noting the fact that $(I_n - P)X = \mathbf{0}$.

Writing $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = (I_n - P)\mathbf{Y}$, we have

$$\text{RSS} = \mathbf{R}^T\mathbf{R} = \mathbf{Y}^T(I_n - P)\mathbf{Y},$$

using the symmetry and idempotence of $I_n - P$.

Hence $\text{RSS} = \mathbf{Z}^TA\mathbf{Z} \sim \sigma^2\chi^2_{n-p}$. Then

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n} \sim \frac{\sigma^2}{n}\chi^2_{n-p}.$$

- Let $V = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \mathbf{R} \end{pmatrix} = D\mathbf{Y}$, where $D = \begin{pmatrix} C \\ I_n - P \end{pmatrix}$ is a $(p + n) \times n$ matrix.

Since $\mathbf{Y}$ is multivariate, $V$ is multivariate with

$$\begin{aligned}\mathrm{Cov}(V) &= D\sigma^2 I D^T \\ &= \sigma^2 \begin{pmatrix} CC^T & C(I_n - P)^T \\ (I_n - P)C^T & (I_n - P)(I_n - P)^T \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} CC^T & C(I_n - P) \\ (I_n - P)C^T & (I_n - P) \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} CC^T & 0 \\ 0 & I_n - P \end{pmatrix}\end{aligned}$$

Using $C(I_n - P) = 0$ (since $(X^TX)^{-1}X^T(I_n - P) = 0$ since $(I_n - P)X = 0$ — check!).

Hence $\hat{\boldsymbol{\beta}}$ and $\mathbf{R}$ are independent since the off-diagonal covariant terms are 0. So $\hat{\boldsymbol{\beta}}$ and $\mathrm{RSS} = \mathbf{R}^T\mathbf{R}$ are independent. So $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent.

From (ii), $\mathbb{E}(\mathrm{RSS}) = \sigma^2(n-p)$. So $\tilde{\sigma}^2 = \frac{\mathrm{RSS}}{n-p}$ is an unbiased estimator of $\sigma^2$. $\tilde{\sigma}$ is often known as the *residual standard error* on $n - p$ degrees of freedom. □

## 3.4 Inference for $\boldsymbol{\beta}$

We know that $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1})$. So

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2(X^TX)^{-1}_{jj}).$$

The *standard error* of $\hat{\beta}_j$ is defined to be

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\tilde{\sigma}^2(X^TX)^{-1}_{jj}},$$

where $\tilde{\sigma}^2 = \mathrm{RSS}/(n-p)$. Unlike the actual variance $\sigma^2(X^TX)^{-1}_{jj}$, the standard error is calculable from our data.

Then

$$\frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\tilde{\sigma}^2(X^TX)^{-1}_{jj}}} = \frac{(\hat{\beta}_j - \beta_j)/\sqrt{\sigma^2(X^TX)^{-1}_{jj}}}{\sqrt{\mathrm{RSS}/((n-p)\sigma^2)}}$$

We can now recognize this as a $t$ distribution:

$$\frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)} \sim t_{n-p}.$$

So we can now do hypothesis testing on $\beta_j$.

## 3.5 $F$ Distribution

More distributions down the way....

**Definition** ($F$ distribution)**.** Suppose $U$ and $V$ are independent with $U \sim \chi^2_m$ and $V \sim \chi^2_n$. The $X = \frac{U/m}{V/n}$ is said to have an *$F$-distribution* on $m$ and $n$ degrees of freedom. We write $X \sim F_{m,n}$

Now from definition we have $X \sim F_{m,n} \Rightarrow \frac{1}{X} \sim F_{n,m}$. Why do we need this? Because finding the lower $\alpha$ point of $F_{m,n}$ just became a task of finding the upper alpha point of $F_{m,n}$ and taking the inverse. So we only need one-sided points for $F$ distributions, and that is how they are usually given.

## 3.6 Making Predictions

After performing the linear regression, we can now make *predictions* from it. Suppose that $\mathbf{x}^*$ is a new vector of values for the explanatory variables.

The expected response at $\mathbf{x}^*$ is $\mathbb{E}[\mathbf{Y} \mid \mathbf{x}^*] = \mathbf{x}^{*T}\boldsymbol{\beta}$. We estimate this by $\mathbf{x}^{*T}\hat{\boldsymbol{\beta}}$. Then we have

$$\mathbf{x}^{*T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(0, \mathbf{x}^{*T}\operatorname{Cov}(\hat{\boldsymbol{\beta}})\mathbf{x}^*) = N(0, \sigma^2\mathbf{x}^{*T}(X^TX)^{-1}\mathbf{x}^*).$$

Let $\tau^2 = \mathbf{x}^{*T}(X^TX)^{-1}\mathbf{x}^*$. Then

$$\frac{\mathbf{x}^{*T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\tilde{\sigma}\tau} \sim t_{n-p}.$$

Then a confidence interval for the *expected response* $\mathbf{x}^{*T}\boldsymbol{\beta}$ has end points

$$\mathbf{x}^{*T}\hat{\boldsymbol{\beta}} \pm \tilde{\sigma}\tau t_{n-p}\left(\frac{\alpha}{2}\right).$$

**Note.** This is a confidence interval for the *expected value*, not the actual value $\mathbf{Y}^*$, which also involves noise.

But what if we want to estimate $\mathbf{Y}^*$?

First of all, as above, the predicted *expected* response is $\hat{Y}^* = \mathbf{x}^{*T}\boldsymbol{\beta}$. This is an unbiased estimator since $\hat{Y}^* - Y^* = \mathbf{x}^{*T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \varepsilon^*$, and hence

$$\mathbb{E}[\hat{Y}^* - Y^*] = \mathbf{x}^{*T}(\boldsymbol{\beta} - \boldsymbol{\beta}) = 0,$$

To find the variance, we use that fact that $\mathbf{x}^{*T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $\varepsilon^*$ are independent, and the variance of the sum of independent variables is the sum of the variances. So

$$\begin{aligned}\operatorname{Var}(\hat{Y}^* - Y^*) &= \operatorname{Var}(\mathbf{x}^{*T}(\hat{\boldsymbol{\beta}})) + \operatorname{Var}(\varepsilon^*) \\ &= \sigma^2\mathbf{x}^{*T}(X^TX)^{-1}\mathbf{x}^* + \sigma^2. \\ &= \sigma^2(\tau^2 + 1).\end{aligned}$$

We can see this as the uncertainty in the regression line $\sigma^2\tau^2$, plus the wobble about the regression line $\sigma^2$. So

$$\hat{Y}^* - Y^* \sim N(0, \sigma^2(\tau^2 + 1)).$$

We therefore find that

$$\frac{\hat{Y}^* - Y^*}{\tilde{\sigma}\sqrt{\tau^2 + 1}} \sim t_{n-p}.$$

So the interval with endpoints

$$\mathbf{x}^{*T}\hat{\boldsymbol{\beta}} \pm \tilde{\sigma}\sqrt{\tau^2 + 1}t_{n-p}\left(\frac{\alpha}{2}\right)$$

is a $95\%$ prediction interval for $Y^*$. This is NOT a confidence interval as we are making predictions.

## 3.7 Hypothesis testing

**Note.** This whole section has not been tested extensively before. This is probably due to the fact that it is impossible to do an ANOVA or linear regression on the Part 1B test. My suggestion would be to understand 3.7.1, and focus on 3.7.2 and 3.7.3, the simplified cases in which this is actually useful.

### 3.7.1 Hypothesis testing

In real life, we are not always sure about the explanatory variables. So we would like to test, for example, $\beta_1 = 0$ against $\beta_1 \neq 0$. And the following section constructs a basis to do that. But first lets prove an uninteresting lemma:

**Lemma.** Suppose $\mathbf{Z} \sim N_n(\mathbf{0}, \sigma^2 I_n)$, and $A_1$ and $A_2$ are symmetric, idempotent $n \times n$ matrices with $A_1 A_2 = 0$ (ie. they are orthogonal). Then $\mathbf{Z}^T A_1 \mathbf{Z}$ and $\mathbf{Z}^T A_2 \mathbf{Z}$ are independent.

This is geometrically intuitive, because $A_1$ and $A_2$ being orthogonal means they are concerned about different parts of the vector $\mathbf{Z}$.

*Proof.* Let $\mathbf{X}_i = A_i \mathbf{Z}$, $i = 1, 2$ and

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix} = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \mathbf{Z}.$$

Then

$$\mathbf{W} \sim N_{2n}\left( \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \sigma^2 \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \right)$$

since the off diagonal matrices are $\sigma^2 A_1^T A_2 = A_1 A_2 = 0$.

So $\mathbf{W}_1$ and $\mathbf{W}_2$ are independent, which implies

$$\mathbf{W}_1^T \mathbf{W}_1 = \mathbf{Z}^T A_1^T A_1 \mathbf{Z} = \mathbf{Z}^T A_1 A_1 \mathbf{Z} = \mathbf{Z}^T A_1 \mathbf{Z}$$

and

$$\mathbf{W}_2^T \mathbf{W}_2 = \mathbf{Z}^T A_2^T A_2 \mathbf{Z} = \mathbf{Z}^T A_2 A_2 \mathbf{Z} = \mathbf{Z}^T A_2 \mathbf{Z}$$

are independent. □

Now we go to hypothesis testing in general linear models:

Suppose $\underset{n \times p}{X} = \begin{pmatrix} \underset{n \times p_0}{X_0} & \underset{n \times (p - p_0)}{X_1} \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \end{pmatrix}$, where $\operatorname{Rank}(X) = p, \operatorname{Rank}(X_0) = p_0$.

We want to test $H_0 : \boldsymbol{\beta}_1 = 0$ against $H_1 : \boldsymbol{\beta}_1 \neq 0$. Under $H_0$, $X_1 \boldsymbol{\beta}_1$ vanishes and

$$\mathbf{Y} = X_0 \boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Under $H_0$, the mle of $\boldsymbol{\beta}_0$ and $\sigma^2$ are

$$\hat{\hat{\boldsymbol{\beta}}}_0 = (X_0^T X_0)^{-1} X_0^T \mathbf{Y}$$
$$\hat{\hat{\sigma}}^2 = \frac{\mathrm{RSS}_0}{n} = \frac{1}{n} (\mathbf{Y} - X_0 \hat{\hat{\boldsymbol{\beta}}}_0)^T (\mathbf{Y} - X_0 \hat{\hat{\boldsymbol{\beta}}}_0)$$

And these are independent. So the fitted values under $H_0$ are

$$\hat{\hat{\mathbf{Y}}} = X_0(X_0^T X_0)^{-1} X_0^T \mathbf{Y} = P_0 \mathbf{Y},$$

where $P_0 = X_0(X_0^T X_0)^{-1} X_0^T$.

The generalized likelihood ratio test of $H_0$ against $H_1$ is

$$\begin{aligned}
\Lambda_{\mathbf{Y}}(H_0, H_1) &= \frac{\left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\right) \exp\left(-\frac{1}{2\hat{\sigma}^2}(\mathbf{Y} - X\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - X\hat{\boldsymbol{\beta}})\right)}{\left(\frac{1}{\sqrt{2\pi\hat{\hat{\sigma}}^2}}\right) \exp\left(-\frac{1}{2\hat{\hat{\sigma}}^2}(\mathbf{Y} - X\hat{\hat{\boldsymbol{\beta}}}_0)^T(\mathbf{Y} - X\hat{\hat{\boldsymbol{\beta}}}_0)\right)} \\
&= \left(\frac{\hat{\hat{\sigma}}^2}{\hat{\sigma}^2}\right)^{n/2} \\
&= \left(\frac{\mathrm{RSS}_0}{\mathrm{RSS}}\right)^{n/2} \\
&= \left(1 + \frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}}\right)^{n/2}.
\end{aligned}$$

We reject $H_0$ when $2\log\Lambda$ is large, equivalently when $\frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}}$ is large.

Using the results in Lecture 8, under $H_0$, we have

$$2\log\Lambda = n\log\left(1 + \frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}}\right),$$

which is approximately a $\chi^2_{p_1 - p_0}$ random variable.

### 3.7.2 Exact Testing

The section above is a good approximation. But what if we demand better? Here it is:

We have previously shown that $\mathrm{RSS} = \mathbf{Y}^T(I_n - P)\mathbf{Y}$, and so

$$\mathrm{RSS}_0 - \mathrm{RSS} = \mathbf{Y}^T(I_n - P_0)\mathbf{Y} - \mathbf{Y}^T(I_n - P)\mathbf{Y} = \mathbf{Y}^T(P - P_0)\mathbf{Y}.$$

Now both $I_n - P$ and $P - P_0$ are symmetric and idempotent, and therefore $\mathrm{Rank}(I_n - P) = n - p$ and

$$\mathrm{Rank}(P - P_0) = \mathrm{tr}(P - P_0) = \mathrm{tr}(P) - \mathrm{tr}(P_0) = \mathrm{Rank}(P) - \mathrm{Rank}(P_0) = p - p_0.$$

Also,

$$(I_n - P)(P - P_0) = (I_n - P)P - (I_n - P)P_0 = (P - P^2) - (P_0 - PP_0) = 0.$$

(we have $P^2 = P$ by idempotence, and $PP_0 = P_0$ since after projecting with $P_0$, we are already in the space of $P$, and applying $P$ has no effect)

Finally,

$$\begin{aligned}
\mathbf{Y}^T(I_n - P)\mathbf{Y} &= (\mathbf{Y} - X_0\boldsymbol{\beta}_0)^T(I_n - P)(\mathbf{Y} - X_0\boldsymbol{\beta}_0) \\
\mathbf{Y}^T(P - P_0)\mathbf{Y} &= (\mathbf{Y} - X_0\boldsymbol{\beta}_0)^T(P - P_0)(\mathbf{Y} - X_0\boldsymbol{\beta}_0)
\end{aligned}$$

since $(I_n - P)X_0 = (P - P_0)X_0 = 0$.

If we let $\mathbf{Z} = \mathbf{Y} - X_0\boldsymbol{\beta}_0$, $A_1 = I_n - P$, $A_2 = P - P_0$, and apply our previous lemma, and the fact that $\mathbf{Z}^T A_i \mathbf{Z} \sim \sigma^2 \chi^2_r$, then

$$\begin{aligned} \mathrm{RSS} &= \mathbf{Y}^T(I_n - P)\mathbf{Y} \sim \chi^2_{n-p} \\ \mathrm{RSS}_0 - \mathrm{RSS} &= \mathbf{Y}^T(P - P_0)\mathbf{Y} \sim \chi^2_{p-p_0} \end{aligned}$$

and these random variables are independent.

So under $H_0$,

$$F = \frac{\mathbf{Y}^T(P - P_0)\mathbf{Y}/(p - p_0)}{\mathbf{Y}^T(I_n - P)\mathbf{Y}/(n - p)} = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/(p - p_0)}{\mathrm{RSS}/(n - p)} \sim F_{p-p_0, n-p},$$

Hence we reject $H_0$ if $F > F_{p-p_0,n-p}(\alpha)$.

$\mathrm{RSS}_0 - \mathrm{RSS}$ is the reduction in the sum of squares due to fitting $\boldsymbol{\beta}_1$ in addition to $\boldsymbol{\beta}_0$.

| Source of var. | d.f. | sum of squares | mean squares | $F$ statistic |
|---|---|---|---|---|
| Fitted model | $p - p_0$ | $\mathrm{RSS}_0 - \mathrm{RSS}$ | $\frac{\mathrm{RSS}_0 - \mathrm{RSS}}{p - p_0}$ | $\frac{(\mathrm{RSS}_0 - \mathrm{RSS})/(p-p_0)}{\mathrm{RSS}/(n-p)}$ |
| Residual | $n - p$ | RSS | $\frac{\mathrm{RSS}}{n-p}$ | |
| Total | $n - p_0$ | $\mathrm{RSS}_0$ | | |

The ratio $\frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}_0}$ is sometimes known as the *proportion of variance explained* by $\boldsymbol{\beta}_1$, and denoted $R^2$.

### 3.7.3 Simple linear regression

Now if we assume it is simple, then

$$Y_i = a' + b(x_i - \bar{x}) + \varepsilon_i,$$

where $\bar{x} = \sum x_i/n$ and $\varepsilon_i$ are $N(0, \sigma^2)$.

We can similarly see that:

| Source of var. | d.f. | sum of squares | mean squares | $F$ statistic |
|---|---|---|---|---|
| Fitted model | 1 | $\mathrm{RSS}_0 - \mathrm{RSS} = \hat{b}^2 S_{XX}$ | $\hat{b}^2 S_{xx}$ | $F = \frac{\hat{b}^2 S_{xx}}{\tilde{\sigma}^2}$ |
| Residual | $n - 2$ | $\mathrm{RSS} = \sum_i (y_i - \hat{y})^2$ | $\tilde{\sigma}^2$. | |
| Total | $n - 1$ | $\mathrm{RSS}_0 = \sum_i (y_i - \bar{y})^2$ | | |

The proposition of variance explained is $\hat{b}^2 S_{xx}/S_{yy} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2$, where $r$ is the Pearson's product-moment correlation coefficient. So this is where the $r^2$ everyone talks about comes from.

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

### 3.7.4 One way analysis of variance with equal numbers in each group

In general, suppose $J$ measurements are taken in each of $I$ groups, and that

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

where $\varepsilon_{ij}$ are independent $N(0, \sigma^2)$ random variables, and the $\mu_i$ are unknown constants.

Fitting this model gives

$$\mathrm{RSS} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i.})^2$$

on $n - I$ degrees of freedom.

Suppose we want to test the hypothesis $H_0 : \mu_i = \mu$, ie. no difference between groups.

Under $H_0$, the model is $Y_{ij} \sim N(\mu, \sigma^2)$, and so $\hat{\mu} = \bar{Y}$, and the fitted values are $\hat{Y_{ij}} = \bar{Y}$. The observed $\mathrm{RSS}_0$ is therefore

$$\mathrm{RSS}_0 = \sum_{i,j}(y_{ij} - \bar{y}_{..})^2.$$

The fitted sum of squares is therefore

$$\mathrm{RSS}_0 - \mathrm{RSS} = \sum_i \sum_j \left((y_{ij} - \bar{y}_{..})^2 - (y_{ij} - \bar{y}_{i.})^2\right) = J\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2.$$

| Source of var. | d.f. | sum of squares | mean squares | $F$ statistic |
|---|---|---|---|---|
| Fitted model | $I-1$ | $J\sum_i(\bar{y}_i - \bar{y}_{..})^2$ | $J\sum_i \frac{(\bar{y}_{i.}-\bar{y}_{..})^2}{I-1}$ | $J\sum_i \frac{(\bar{y}_{i.}-\bar{y}_{..})^2}{(I-1)\tilde{\sigma}^2}$ |
| Residual | $n-I$ | $\sum_i\sum_j(y_{ij} - \bar{y}_{i.})^2$ | $\tilde{\sigma}^2$ | |
| Total | $n-1$ | $\sum_i\sum_j(y_{ij} - \bar{y}_{..})^2$ | | |