Statistical Significance Testing From Scratch

Karl Stratos

1 Fundamental Notions

Let $Z_\mu \sim \mathcal{N}(\mu, 1)$ with unknown μ. Our "alternative" hypothesis is

$$H_1 : \ \mu \neq 0$$

We want to decide if we should accept H_1 based on a single sample z_μ of Z_μ. Since we don't care about the exact nonzero value of μ that makes H_1 true, we may consider a proof by contradiction. A **null hypothesis** is a statement that is false iff H_1 is true, in this case

$$H_0 : \ \mu = 0$$

Now we need to decide if we should *reject* H_0 based on z_μ. There are two possible errors.

- **Type I error**: We accept a false H_1 (i.e., reject a true H_0).
- **Type II error**: We reject a true H_1 (i.e., accept a false H_0).

We want to especially avoid a type I error. To this end, we introduce a hyperparameter $\alpha \in (0, 1)$ called a **significance level**. We will define $\texttt{RejectNull}_\alpha : \mathbb{R} \to \{0, 1\}$ that maps z_μ to 1 iff it rejects H_0 such that the associated type I error probability is α. Formally,

$$\alpha = \Pr(\texttt{RejectNull}_\alpha(Z_\mu) = 1 | H_0 \text{ is true}) \tag{1}$$

where the probability is over $Z_\mu \sim \mathcal{N}(\mu, 1)$. This justifies rejecting H_0 (i.e., accepting H_1) if $\texttt{RejectNull}_\alpha(z_\mu) = 1$ and α is sufficiently small since then the chance of rejecting a true H_0 is small. In constrast, we do *not* directly control the probability of making a type II error

$$\beta = \Pr(\texttt{RejectNull}_\alpha(Z_\mu) = 0 | H_0 \text{ is false})$$

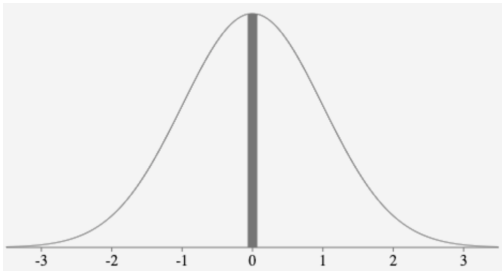
which is a complicated function of α. Thus we do *not* accept H_0 (i.e., reject H_1) if $\texttt{RejectNull}_\alpha(z_\mu) = 0$: we simply say there is not enough evidence to accept either H_0 or H_1.

Statistical power. A related quantity is the probability of rejecting a false H_0 (i.e., accepting a true H_1) called the **statistical power** of the test

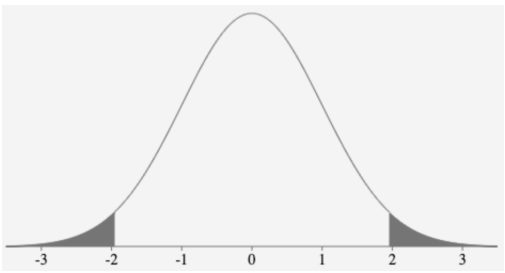
$$1 - \beta = \Pr(\texttt{RejectNull}_\alpha(Z_\mu) = 1 | H_0 \text{ is false})$$

If a test is statistically weak, it has a high chance of not accepting a true H_1. As an illustration, a conservative test that never rejects H_0 has the statistical power of 0.

Critical region. There are many valid rejection rules that satisfy (1). Specifically, $\texttt{RejectNull}_\alpha(z_\mu) = [[z_\mu \in \kappa_\alpha]]$ where $\kappa_\alpha \subset \mathbb{R}$ is any subset whose probability mass is α under $\mathcal{N}(0, 1)$. Below we show two possible choices of κ_α at $\alpha = 0.05$ (image credit: standard normal mass calculator by David Lane).



$\kappa_{0.05}$ $(-0.0626, 0.0626)$



$(-\infty, -1.96) \cup (1.96, \infty)$

(One can imagine other crazy probability mass allocations.) While both are valid, the second is more intuitive because it consists of points that "most violate" H_0 (i.e., $\mu = 0$). Henceforth we will only consider this unique choice of κ_α and call it the **critical region**. In this case, it is given by $\kappa_\alpha = (-\infty, -c_\alpha) \cup (c_\alpha, \infty)$ with c_α (so-called **critical value**) satisfying

$$1 - \Phi(c_\alpha) = \Phi(-c_\alpha) = \frac{\alpha}{2}$$

where $\Phi(z) = \Pr(Z_0 \leq z)$ is the CDF of $\mathcal{N}(0, 1)$. Note that we divide α by 2 because $\mathcal{N}(0, 1)$ is symmetric and the most violating points are evenly split across two extremes. We reject H_0 iff the sample falls in the critical region. This is equivalent to rejecting H_0 iff 0 is not trapped in a confidence interval for the mean of Z_μ with confidence $1 - \alpha$ (Appendix B).

p-value. The particular choice of κ_α as the critical region is also useful because it allows us to consider a sample-dependent quantity called the **p-value**. Specifically, p is the probability of sampling a point that violates H_0 *more than* the given observation. In this case,

$$p = \Pr(|Z_\mu| > z_\mu | H_0 \text{ is true}) = \Pr(Z_\mu > z_\mu | H_0 \text{ is true}) + \Pr(Z_\mu < -z_\mu | H_0 \text{ is true}) = 2\Phi(-z_\mu)$$

Since $p \leq \alpha$ iff $z_\mu \in \kappa_\alpha$, we can use $p \leq \alpha$ as an alternative rejection rule. If we reject H_0, the p-value tells us how strongly we reject it (the smaller p is, the stronger the rejection). If $z_\mu = 2.03$, we reject H_0 at a p-value of 0.04. If $z_\mu = -3.7$, we reject H_0 at a p-value of 0.0001.

1.1 One-Tailed Tests

Our hypothesis may be "one-tailed". Instead of claiming $\mu \neq 0$, we may claim $H_1 : \mu > 0$ (**upper-tailed**). The corresponding null hypothesis is $H_0 : \mu \leq 0$. The notion of "most violating" H_0 changes to being furthest away *to the right*, so the critical region is given by $\kappa_\alpha = (c_\alpha, \infty)$ for a critical value $c_\alpha > 0$. For any $\mu \leq 0$ the probability of a type I error is

$$\Pr(Z_\mu > c_\alpha | \mu \leq 0) \leq \Pr(Z_\mu > c_\alpha | \mu = 0) = \Phi(-c_\alpha)$$

Hence we can use a critical region (c_α, ∞) satisfying $\Phi(-c_\alpha) = \alpha$ and upper bound the probability by α. Similarly, for any $\mu \leq 0$ the probability of sampling a point that violates H_0 more than $z_\mu \sim \mathcal{N}(\mu, 0)$ is

$$\Pr(Z_\mu > z_\mu | \mu \leq 0) \leq \Pr(Z_\mu > z_\mu | \mu = 0) = \Phi(-z_\mu)$$

So $p = \Phi(-z_\mu)$ upper bounds the probability. In short, we can simply compute the critical region and the p-value assuming $\mu = 0$ because doing so will only make our conclusion more conservative. In a **lower-tailed** test, we claim that $H_1 : \mu < 0$.

1.2 Example

Let $\alpha = 0.05$ be our significance level which ensures that the chance of accepting a false hypothesis is at most 5%. We get a sample $z_\mu \sim \mathcal{N}(\mu, 1)$ and find that $z_\mu = 2.03$. Since the value is rather large, we have every reason to hypothesize that $\mu > 0$ or $\mu \neq 0$, but probably not $\mu < 0$. Below are our test results.

$z_\mu = 2.03$, $\alpha = 0.05$

H_1		H_0	κ_α	Conclusion	p
$\mu > 0$	(upper-tailed)	$\mu \leq 0$	$(1.645, \infty)$	reject H_0	0.02
$\mu < 0$	(lower-tailed)	$\mu \geq 0$	$(\infty, -1.645)$	do not reject H_0	0.98
$\mu \neq 0$	(two-tailed)	$\mu = 0$	$(-\infty, -1.96) \cup (1.96, \infty)$	reject H_0	0.04

Note that the two-tailed test is statistically weaker than one-tailed tests. Had we sampled $z_\mu = 1.87$, we can accept $\mu > 0$ but not $\mu = 0$. Even so, one-tailed tests are rarely appropriate because they fail to check if the other direction is true which might be vital information (e.g., the new drug is actually *less* effective, not more). Reporting a one-tailed p-value after rejecting the null hypothesis with a two-tailed test to make the result more significant would be dishonest. Thus we always consider a two-tailed test except in test types in which only a one-tailed test is meaningful (e.g., F-test).

1.3 Test Statistic

One way to categorize significance tests is by the distribution type of the **test statistic**, defined as whatever quantity derived from the sample that we use to accept or reject the hypothesis. In this toy test, the test statistic *is* the sample $z_\mu \sim \mathcal{N}(\mu, 1)$. In general, the test statistic is a nontrivial function of the sample.

2 Z-Test

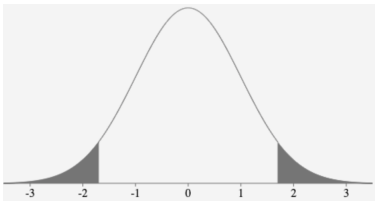
This is the simplest non-identity test statistic. Let $X_1 \dots X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where we know the value of σ^2. Our hypothesis is

$$H_1 : \ \mu \neq \tilde{\mu}$$

for some $\tilde{\mu} \in \mathbb{R}$. The corresponding null hypothesis is $H_0 : \mu = \tilde{\mu}$. Under the null hypothesis, $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i \sim \mathcal{N}(\tilde{\mu}, \sigma^2/N)$, therefore

$$\frac{\bar{X}_N - \tilde{\mu}}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1) \tag{2}$$

so we can again use the standard normal distribution to calculate critical values and the p-value. For instance, if the test statistic is 1.7, the two-tailed p-value is 0.0891 corresponding to the shaded area under $\mathcal{N}(0, 1)$ and we fail to reject H_0 at $\alpha = 0.05$.[1]



The toy test in Section 1 is a special case with $\tilde{\mu} = 0$, $\sigma^2 = 1$, and $N = 1$. Note that this is specifically designed to take advantage of the central limit theorem (CLT): even if the sample distribution is not normal, $\bar{X}_N \overset{\text{approx.}}{\sim} \mathcal{N}(\tilde{\mu}, \sigma^2/N)$ as $N \to \infty$ so that (2) holds approximately. However, since we rarely assume that we know the true variance this is rarely used.

3 T-Test

3.1 One-Sample T-Test

Let $X_1 \dots X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $N \geq 2$. Our hypothesis is again

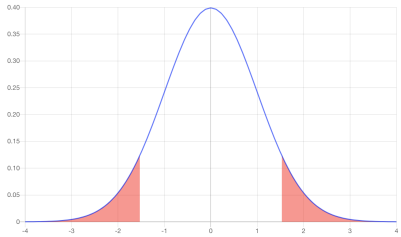
$$H_1 : \ \mu \neq \tilde{\mu}$$

for some $\tilde{\mu} \in \mathbb{R}$. The corresponding null hypothesis is $H_0 : \mu = \tilde{\mu}$. Under the null hypothesis, $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and $\bar{S}_N^2 = \frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X}_N)^2$ satisfy (Lemma H.2)

$$\frac{\bar{X}_N - \tilde{\mu}}{\bar{S}_N/\sqrt{N}} \sim \tau(N-1) \tag{3}$$

where $\tau(\nu)$ is the t-distribution with ν degrees of freedom, which we can use calculate critical values and the p-value (Appendix H.2). For instance, if $N = 5$ and the test statistic is -1.533, the two-tailed p-value is 0.2 corresponding to the shaded area under $\tau(4)$

[1] In a lower-tailed z-test, the p-value would be a half 0.0446 and we reject H_0. But as discussed earlier we should not consider one-tailed tests when a two-tailed test is meaningful.



As $N \to \infty$ (3) becomes equivalent to (2) since $\frac{\bar{X}_N - \tilde{\mu}}{\bar{S}_N/\sqrt{N}}$ converges to $\frac{\bar{X}_N - \tilde{\mu}}{\sigma/\sqrt{N}}$ by Slutsky's theorem (Theorem C.1) and $\tau(N-1)$ converges to $\mathcal{N}(0,1)$ (Appendix H.2). This makes the t-test applicable on non-normal samples if N is sufficiently large.

3.2 Paired Two-Sample T-Test

This test compares the means of two arbitrarily dependent (i.e., paired) normal variables. We can reduce it to the one-sample t-test as follows. Let $(X_1, Y_1) \dots (X_N, Y_N) \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Sigma)$ where $\mu = (\mu_X, \mu_Y) \in \mathbb{R}^2$ and $\Sigma \in \mathbb{R}^{2\times 2}_{\succ 0}$. Our hypothesis is

$$H_1 : \ \mu_X \neq \mu_Y$$

The corresponding null hypothesis is $H_0 : \mu_X = \mu_Y$. Under the null hypothesis, $X_i - Y_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = \Sigma_{11} + \Sigma_{22} + 2\Sigma_{12}$ so that

$$\frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)}{\bar{S}_N/\sqrt{N}} \sim \tau(N-1)$$

where $\bar{S}_N^2$ is the sample estimator of σ^2. Again, even if the samples are not normal, the test is asymptotically exact as $N \to \infty$ by the CLT.

3.3 Independent Two-Sample T-Test

An alternative two-sample test can be derived if the variables are independent. Let $X_1 \dots X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu_X, \sigma^2)$ and $Y_1 \dots Y_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu_Y, \sigma^2)$. We assume equal variance σ^2 and sample size $N \geq 2$ for simplicity. Our hypothesis is

$$H_1 : \ \mu_X \neq \mu_Y$$

The corresponding null hypothesis is $H_0 : \mu_X = \mu_Y$. Under the null hypothesis, we have (Lemma H.3)

$$\frac{\bar{X}_N - \bar{Y}_N}{\bar{S}_{\text{pooled}}\sqrt{\frac{2}{N}}} \sim \tau(2N-2) \tag{4}$$

where $\bar{S}^2_{\text{pooled}} = (\bar{S}^2_X + \bar{S}^2_Y)/2$ is the pooled estimator of σ^2 (37). The denominator $\bar{S}_{\text{pooled}}\sqrt{\frac{2}{N}}$ is called the **standard error of the difference between two means**.

3.4 Regression Slope T-Test

Let $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ denote any full-rank matrix with $d \geq 2$ and first column $\mathbf{1}_N$ (i.e., bias dimension) and let $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}w_{\text{true}}, \sigma^2 I_{N\times N})$. Pick any $j \in \{1 \dots d\}$. Our hypothesis is

$$H_1 : \ [w_{\text{true}}]_j \neq 0$$

That is, we hypothesize that the response variable is correlated with the j-th feature. The corresponding null hypothesis is $H_0 : [w_{\text{true}}]_j = 0$. Under the null hypothesis, we have (Corollary D.18)

$$\frac{[\hat{w}]_j}{\sqrt{\frac{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{j,j}}{N-d}}\,||\hat{\boldsymbol{\epsilon}}||} \sim \tau(N-d)$$

where $\hat{w} = \arg\min_{w\in\mathbb{R}^d} \ ||\boldsymbol{y} - \boldsymbol{X}w||^2$ denote the LSE parameter with residuals $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{w}$. The denominator $\sqrt{\frac{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{j,j}}{N-d}}\,||\hat{\boldsymbol{\epsilon}}||$ is called the **standard error of the slope coefficient**.

4 F-Test

In a single-comparison test like z-test or t-test, the null hypothesis is of the form $a = b$ which can be violated as either $a > b$ or $a < b$. In contrast, many F-tests are multiple-comparison ("omnibus") tests in which the null hypothesis is of the form $a_1 = a_2 = \cdots = a_K$ for $K \geq 2$ where it does not make sense to make a two-way distinction. Instead, the test stastistic is designed in such a way that the more violated the null hypothesis is, the bigger the statistic is. Thus multiple-comparison F-tests are always upper-tailed. A single-comparison F-test (e.g., comparing two variances) remains two-tailed.

4.1 One-Way ANOVA

For $k = 1 \dots K \geq 2$, let $Y_{k,1} \dots Y_{k,N_k} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_k, \sigma^2)$ where $N_k \geq 2$ and let $\bar{Y}_k$ and $\bar{S}_k^2$ denote the sample mean and variance. Denote the total number of samples by $N = \sum_{k=1}^K N_k$ and the "grand mean" by $\bar{Y} = \frac{1}{N}\sum_{k=1}^K \sum_{i=1}^{N_k} Y_{k,i}$. We may define an unbiased and a biased estimator of σ^2 called the pooled variance (37) and the between-group variance (41) (Appendix F):

$$\begin{aligned} \bar{S}^2_{\text{pooled}} &= \frac{1}{N-K}\sum_{k=1}^K \sum_{i=1}^{N_k} (Y_{k,i} - \bar{Y}_k)^2 & \mathbf{E}\left[\bar{S}^2_{\text{pooled}}\right] &= \sigma^2 \\ \bar{S}^2_{\text{between}} &= \frac{1}{K-1}\sum_{k=1}^K N_k(\bar{Y}_k - \bar{Y})^2 & \mathbf{E}\left[\bar{S}^2_{\text{between}}\right] &= \sigma^2 + \frac{1}{K-1}\sum_{k=1}^K (\mu_k - \mu)^2 \end{aligned} \tag{5}$$

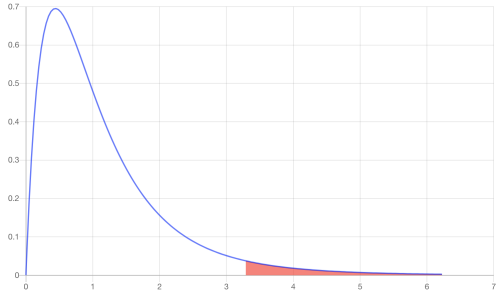
Our hypothesis is

$$H_1 : \ \mu_i \neq \mu_j \text{ for some } i \neq j$$

The corresponding null hypothesis is $H_0 : \mu_1 = \cdots = \mu_K$. Under the null hypothesis, we have (Corollary F.6)

$$\frac{\bar{S}^2_{\text{between}}}{\bar{S}^2_{\text{pooled}}} \sim F(K-1, N-K) \tag{6}$$

where $F(d_1, d_2)$ is the F-distribution with (d_1, d_2) degrees of freedom (Appendix H.3). The statistic is always positive. Importantly, the bigger it is, the more H_0 is violated (5), so the test is upper-tailed. For instance, if $K = 5$ and $N = 20$, and the test statistic is 3.29, the upper-tailed p-value is 0.04 corresponding to the shaded area under $F(4, 15)$



When $K = 2$. For simplicity assume the same sample size $M = N_1 = N_2$. In this case $\bar{S}^2_{\text{between}} = (M/2)(\bar{Y}_1 - \bar{Y}_2)^2$ (Lemma K.2) and, using the fact that $F(1, d) = \tau^2(d)$,

$$\frac{\bar{S}^2_{\text{between}}}{\bar{S}^2_{\text{pooled}}} \sim F(1, 2M-2) \qquad \Leftrightarrow \qquad \frac{\bar{Y}_1 - \bar{Y}_2}{\bar{S}_{\text{pooled}}\sqrt{\frac{2}{M}}} \sim \tau(2M-2)$$

Thus when $K = 2$, the one-way ANOVA (6) equivalent to the independent two-sample t-test (4).

4.2 Equality of Two Variances

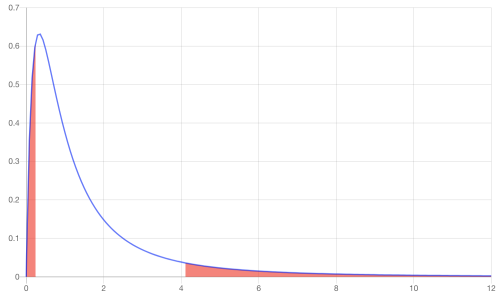
Let $X_1 \dots X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1 \dots Y_M \overset{\text{iid}}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$ where $N, M \geq 2$. Our hypothesis is

$$H_1 : \ \sigma_X^2 \neq \sigma_Y^2$$

The corresponding null hypothesis is $H_0 : \sigma_X^2 = \sigma_Y^2$. Under the null hypothesis, we have (Corollary H.5)

$$\frac{\bar{S}_X^2}{\bar{S}_Y^2} \sim F(N-1, M-1)$$

where $\bar{S}_X^2$ and $\bar{S}_Y^2$ are sample variances. Since H_0 can be violated in either direction $\sigma_X^2 > \sigma_Y^2$ or $\sigma_X^2 < \sigma_Y^2$, the test should be two-tailed. For instance, if $N = M = 5$ and the test statiatic is 0.18, the two-tailed p-value is 0.2 corresponding to the shaded area under $F(4, 4)$



4.3 Regression F-Test

Let $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ denote any full-rank matrix with $d \geq 2$ and first column $\mathbf{1}_N$ and $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X} w_{\text{true}}, \sigma^2 I_{N \times N})$. Pick any $Q \subset \{1 \dots d\}$. Our hypothesis is

$$H_1 : \ \text{There exists some } j \notin Q \text{ such that } [w_{\text{true}}]_j \neq 0.$$

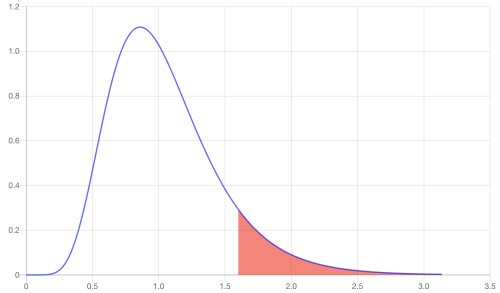
This can be seen as an omnibus version of the regression slope t-test (Section 3.4) where we hypothesize that the response variable is correlated with *some* feature outside Q. The corresponding null hypothesis is

$$H_0 : \ [w_{\text{true}}]_j = 0 \text{ for all } j \notin Q.$$

Let $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{X}^+\boldsymbol{y}$ denote the least-squares prediction by the "full" model. For any subset $S \subseteq \{1 \dots d\}$, we will write $\hat{\boldsymbol{y}}_S = \boldsymbol{X}_S \boldsymbol{X}_S^+ \boldsymbol{y}$ to denote the prediction by a "partial" model using *only* the columns of $\boldsymbol{X}$ indexed by S. Under the null hypothesis, we have (Theorem D.15)

$$\frac{\left|\left|\hat{\boldsymbol{y}}_P - \hat{\boldsymbol{y}}_Q\right|\right|^2 / (|P| - |Q|)}{||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2 / (N - d)} \sim F(|P| - |Q|, N - d) \qquad \forall P \supset Q \tag{7}$$

Since this holds for any $P \supset Q$, we can devise multiple tests for the same hypothesis by varying P. Intuitively, (7) measures the predictive gain when features in $P \backslash Q$ are added while considering the performance of the full model (e.g., the gain is meaningless if $||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2$ is huge). The regression F-test is upper-tailed since the bigger the statistic is, the more violated H_0 is. For instance, if $N = 100$, $d = 58$, $|P| = 30$, and $|Q| = 10$, and the test statistic is 1.596, the upper-tailed p-value is 0.1 corresponding to the shaded area under $F(20, 42)$



A popular application of (7) is "ablating" discrete variables in regression. This is because encoding $A \in \{1 \dots K\}$ requires $K - 1$ dimensions (Appendix G), so asking "Does A matter?" requires an omnibus test "Does any of the $K - 1$ features generated by A matter?".

4.3.1 One-way ANOVA (revisited)

Let $A \in \{1 \dots K\}$. Our hypothesis is that A is correlated with Y. The full regression model is

$$Y = w_{\text{true}} \begin{bmatrix} 1 \\ C_A(A) \end{bmatrix} + Z_{\sigma^2} \qquad\qquad Z_{\sigma^2} \sim \mathcal{N}(0, \sigma^2)$$

where $C_A(A) \in \mathbb{R}^{K-1}$ is an encoding of A. Under the null hypothesis (there is no correlation between A and Y), (7) holds with $Q = \{1\}$ and $P = \{1 \dots K\}$. Note that the test distribution $F(K-1, N-K)$ is the same as in the one-way ANOVA. In fact, (7) and (6) are equivalent under dummy coding (Appendix F.3.1).

4.3.2 Multi-way ANOVA.

The regression view of the one-way ANOVA naturally leads to a multivariate generalization. For instance, let $A \in \{1 \dots K\}$ and $B \in \{1 \dots L\}$ (i.e., **two-way ANOVA**). Our hypothesis considers A, B, or their interaction $A{:}B \in \{1 \dots KL\}$, and claims that it is correlated with Y. The full regression model is

$$Y = w_{\text{true}} \begin{bmatrix} 1 \\ C_A(A) \\ C_B(B) \\ C_A(A) \otimes C_B(B) \end{bmatrix} + Z_{\sigma^2} \qquad\qquad Z_{\sigma^2} \sim \mathcal{N}(0, \sigma^2)$$

where $C_A(A) \in \mathbb{R}^{K-1}$ and $C_B(B) \in \mathbb{R}^{L-1}$ are encodings of A, B and $C_A(A) \otimes C_B(B) \in \mathbb{R}^{(K-1)(L-1)}$ is their kronecker product. Let $\mathcal{I}_A$, $\mathcal{I}_B$, $\mathcal{I}_{A:B}$ denote subsets of dimensions corresponding to A, B, $A{:}B$. We can apply (7) to ablate the impact of A, B, or $A{:}B$ under the full model. But this requires choosing appropriate subsets $Q \subset P$. For instance, to test the impact of A, we need to select Q, P such that $P \backslash Q = \mathcal{I}_A$. There are three standard "types".

Ablated variable	**Type I**	**Type II**	**Type III**
$A \in \{1 \dots K\}$	$Q = \{1\}$ $P = \{1\} \cup \mathcal{I}_A$	$Q = \{1\} \cup \mathcal{I}_B$ $P = \{1\} \cup \mathcal{I}_A \cup \mathcal{I}_B$	$Q = \{1\} \cup \mathcal{I}_B \cup \mathcal{I}_{A:B}$ $P = \{1\} \cup \mathcal{I}_A \cup \mathcal{I}_B \cup \mathcal{I}_{A:B}$
$B \in \{1 \dots L\}$	$Q = \{1\} \cup \mathcal{I}_A$ $P = \{1\} \cup \mathcal{I}_A \cup \mathcal{I}_B$	$Q = \{1\} \cup \mathcal{I}_A$ $P = \{1\} \cup \mathcal{I}_A \cup \mathcal{I}_B$	$Q = \{1\} \cup \mathcal{I}_A \cup \mathcal{I}_{A:B}$ $P = \{1\} \cup \mathcal{I}_A \cup \mathcal{I}_B \cup \mathcal{I}_{A:B}$
$A{:}B \in \{1 \dots KL\}$	$Q = \{1\} \cup \mathcal{I}_A \cup \mathcal{I}_B$ $P = \{1\} \cup \mathcal{I}_A \cup \mathcal{I}_B \cup \mathcal{I}_{A:B}$	$Q = \{1\} \cup \mathcal{I}_A \cup \mathcal{I}_B$ $P = \{1\} \cup \mathcal{I}_A \cup \mathcal{I}_B \cup \mathcal{I}_{A:B}$	$Q = \{1\} \cup \mathcal{I}_A \cup \mathcal{I}_B$ $P = \{1\} \cup \mathcal{I}_A \cup \mathcal{I}_B \cup \mathcal{I}_{A:B}$

Type I is "sequential" and requires an ordering of variables (here, $A \to B \to A{:}B$). Type II and III are "simultaneous" (i.e., no ordering necessary) but differ in that Type II only uses terms of the same or lower interaction order while Type III always sets P to be the full model including higher-order interactions. Caution: Type III requires a mean-centered encoding of discrete variable (e.g., sum coding when data is balanced, see Appendix G.2.2).[2]

Balanced data. The good news is that when the data is balanced (i.e., the number of samples for each configuration of discrete variables is the same, see Appendix G.1), type I/II/III produce the same result. Intuitively, this is because it removes collinearity: knowing about the value of one variable tells nothing about the value of other variable, so how we define the nested models is irrelevant.

5 Studentized Range Test

For $k = 1 \dots K \geq 2$, let $Y_{k,1} \dots Y_{k,M} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_k, \sigma^2)$ where $M \geq 2$ and let $\bar{Y}_k$ and $\bar{S}_k^2$ denote the sample mean and variance. Our hypothesis is

$$H_1 : \ \mu_i \neq \mu_j \text{ for some } i \neq j$$

[2] To see why, let Z, Z' be two independent scalar variables. The covariance between Z and ZZ' is

$$\text{Cov}\left(Z, ZZ'\right) = \mathbf{E}\left[Z^2 Z'\right] - \mathbf{E}\left[Z\right]\mathbf{E}\left[ZZ'\right] = \mathbf{E}\left[Z^2\right]\mathbf{E}\left[Z'\right] - \mathbf{E}\left[Z\right]^2\mathbf{E}\left[Z'\right] = \mathbf{E}\left[Z'\right]\text{Var}\left(Z\right)$$

So even if variables are independent, the covariance between a variable and a higher-order interaction term involving that variable is nonzero unless the variable has zero mean. Because model Q in Type III uses higher-order interaction terms, it is not a true "subset" of model P if the encoding is not mean-centered.

The corresponding null hypothesis is $H_0 : \mu_1 = \cdots = \mu_K$. Under the null hypothesis, we have

$$\frac{\max_{k=1}^K \bar{Y}_k - \min_{k=1}^K \bar{Y}_k}{\bar{S}_{\text{pooled}}\sqrt{\frac{2}{N}}} \sim \textbf{srange}(K, MK - K) \tag{8}$$

where $\bar{S}^2_{\text{pooled}} = (1/K)\sum_{k=1}^K \bar{S}_k^2$ is the pooled estimator of σ^2 (37) and $\textbf{srange}(K, \nu)$ is the studentized range distribution with K groups and ν degrees of freedom (Appendix H.4). The test is upper tailed since the statistic is bigger when H_0 is more violated. Note that the statistic reduces to (4) in the independent two-sample t-test if $K = 2$ (with the difference that we consider the absolute difference between the means).

Tukey's range test. A notable aspect of (8) is that the numerator considers the absolute difference between a *pair* of means, even though the test itself is an omnibus test. This allows for a *pairwise comparison* test known as Tukey's range test in which we compute

$$\frac{\bar{Y}_l - \bar{Y}_k}{\bar{S}_{\text{pooled}}\sqrt{\frac{2}{N}}} \qquad \forall 1 \le k < l \le K \tag{9}$$

where WLOG we assume $\bar{Y}_1 \le \cdots \le \bar{Y}_K$. We reject that $\bar{Y}_l = \bar{Y}_k$ if (9) is larger than the critical value at a given significance level α, where the critical value is computed using (8). This trick allows us to test multiple hypotheses (namely $H_1^{(k,l)} : \mu_k \ne \mu_l$ for $k < l$) without increasing the familiy-wise error rate (Appendix A). In constrast with the Bonferroni correction which divides α by the number of hypotheses, Tukey's range test uses the same α, so it is more statistically powerful when the number of hypotheses is large (as is the case in all pairwise comparisons).

6 TODO: Chi-Square Tests

7 Non-Parametric Tests

Most of the tests assume normality. While we can invoke the CLT to argue that this assumption is benign when the sample size is nontrivial, it is possible to devise **non-parameteric** tests that make no assumption on the type of the data distribution.

7.1 Wilcoxon Signed-Rank Test

This is a non-parametric version of the paired two-sample t-test (Section 3.2). Let $(X_1, Y_1) \dots (X_N, Y_N) \in \mathbb{R}^2$ be N iid pairs of random variable such that $X_i - Y_i \sim \textbf{Sym}(\mu)$ where $\textbf{Sym}(\mu)$ is some symmetric distribution over $\mathbb{R}$ centered at μ. Our hypothesis is

$$H_1 : \ \mu \ne 0$$

The corresponding null hypothesis is $H_0 : \mu = 0$. Let $N' \le N$ denote the number of samples such that $X_i \ne Y_i$. Under the null hypothesis, we have

$$W = \sum_{i=1:\ X_i \ne Y_i}^{N} \textbf{sign}(X_i - Y_i) R_i \sim \textbf{Wilcoxon}(N')$$

where $\textbf{sign} : \mathbb{R} \to \{\pm 1\}$ is the sign function, $R_i \in \mathbb{N}$ is the rank of $|X_i - Y_i|$ from smallest to largest, and $\textbf{Wilcoxon}(\nu)$ is some complex distribution with parameter ν (centered at zero). Since H_0 can be violated by either $\mu > 0$ or $\mu < 0$, the test is two-tailed. For instance, with $N' = 10$, the lower and upper critical values at $\alpha = 0.05$ are 8 and 47, so we reject H_0 if $W \le 8$ or $W \ge 47$. When $N' \ge 20$, W is approximately normally distributed by the CLT, so we may standardize it and use the standard normal distribution instead. While the Wilcoxon signed-rank test doesn't require normality, it is statistically weaker than the paired t-test.

8 Discussions

8.1 Practical Issues

Significance level. The significance level is almost always fixed to be $\alpha = 0.05$. That is, we are usually content with a 5% chance of accepting a false H_1.

Software. In practice we never compute these statistics by hand. Instead we use software like R or statsmodels (for Python). We design an alternative hypothesis *a priori* (and never change it after the test *a posteriori*!), choose an appropriate test, then feed our data to a function that executes the test.

Single hypothesis. If we compare a single pair of means, we can use the paired t-test. We can also use the Wilcoxon signed-rank test if we want to be conservative.

Multiple hypotheses. If we have multiple pairwise comparisons, we want to correct for the FWER to avoid accepting some false hypothesis. We can use either the paired t-test with the Bonferroni correction (i.e., divide α by the number of hypotheses, then conduct each test separately) or the Tukey's range test. The latter is preferred if we are doing *all-pair* comparisons among many distributions since the Bonferroni correction becomes statistically weak with too many hypotheses.

Omnibus tests followed by post-hoc tests. We can consider a two-stage test procedure in which we first show that there exists a pair of means that differ under multiple populations through ANOVA, *then* follow up with the Tukey's range test as a post-hoc test to identify which means differ. The benefit of doing this instead of just applying Tukey immediately is that we can specify a more detailed model in ANOVA which may also have more statistical power (but this seems a bit controversial). For instance, Ehud Reiter proposes to first check if there's any difference between systems for text generation by a three-way ANOVA where the full model is, in the formula language for specifying a regression model,

$$\text{Likert} \sim \text{Subject} * \text{Scenario} * \text{System}$$

Here, $\text{Likert} \in \mathbb{R}$ is the Likert score assigned on the text from $\text{System} \in \{1 \dots K\}$ by $\text{Subject} \in \{1 \dots L\}$ (i.e., human annotator) in a certain $\text{Scenario} \in \{1 \dots G\}$, and * denotes full interactions (so there will be 3 second-order interaction variables and 1 third-order interaction variable) under some discrete-variable encoding. We may start with all-way interactions and gradually remove higher order interactions that are not significant according to ANOVA to arrive at a final model that only retains significant interactions, for instance

$$\text{Likert} \sim \text{Subject} + \text{Scenario} + \text{System} + \text{System:Subject}$$

where $\text{System:Subject} \in \{1 \dots KL\}$ denotes a pairwise interaction. If the means of System differ according to the final model, then we use Tukey to make all $K(K-1)/2$ pairwise comparisons to find which systems differ.

Useful resources.

- Tutorial by Python for Data Science: This shows how to conduct a multi-way ANOVA test and a post-hoc test in Python.
- Blog post by Mattan S. Ben-Shachar: This gives more details of the different types of ANOVA and how to replicate them by hand.

References

Driscoll, M. F. and Gundberg Jr, W. R. (1986). A history of the development of craig's theorem. *The American Statistician*, **40**(1), 65–70.

Geary, R. (1936). The distribution of" student's" ratio for non-normal samples. *Supplement to the Journal of the Royal Statistical Society*, **3**(2), 178–184.

A Multiple Hypotheses

Often we need to test multiple hypotheses simultaneously. For instance, for $i = 1 \dots m$ we may claim

$$H_1^{(i)} : \mu^{(i)} \neq \nu^{(i)}$$

for some parameters $\mu^{(i)}, \nu^{(i)}$ based on samples from their distributions. We test each hypothesis with significance level $\alpha^{(i)}$. The **family-wise error rate (FWER)** is defined as the probability of incorrectly accepting *some* false hypothesis. As m increases, the FWER increases.[3] Thus we need to control for FWER if we wish to avoid accepting some false hypothesis. A simplest approach is called the **Bonferroni correction** which upper bounds the FWER by the union bound:

$$\textbf{FWER} = \Pr(\exists i \in \{1 \dots m\} : H_0^{(i)} \text{ is falsely rejected}) \leq \sum_{i=1}^{m} \Pr(H_0^{(i)} \text{ is falsely rejected}) = \sum_{i=1}^{m} \alpha^{(i)}$$

This means that to ensure $\textbf{FWER} \leq \alpha$, we can set $\alpha^{(i)} = \alpha/m$. However, the union bound can be loose and the test is statistically weak if m is large (i.e., we may fail to reject many false $H_0^{(i)}$).

B Hypothesis Tests and Confidence Intervals

We can express a hypothesis test as an equivalent statement about a confidence interval (CI). A CI for the parameter θ (mostly the mean) of a distribution with confidence γ is a *random* interval I_θ^γ that traps θ with probability γ. For instance, a CI for the mean of $Z_\mu \sim \mathcal{N}(\mu, 0)$ with $\gamma = 0.95$ is $I_\mu^{0.95} = [Z_\mu \pm 1.96]$ since

$$\Pr(-1.96 \leq Z_0 \leq 1.96) = 1 - 2\Phi(-1.96) = 0.95 \qquad \Leftrightarrow \qquad \Pr(Z_\mu - 1.96 \leq \mu \leq Z_\mu + 1.96) = 0.95$$

where we use the fact that $Z_\mu = \mu + Z_0$. Note that other choices are possible (e.g., assymetric or segmented), but this is the most "natural" choice of CI for trapping the mean of a symmetric distribution. Each sample z_μ of Z_μ yields a sample $i_\mu^{0.95} = [z_\mu \pm 1.96]$ of $I_\mu^{0.95}$. For any $\tilde{\mu} \in \mathbb{R}$, we reject the hypothesis $\mu = \tilde{\mu}$ iff $\tilde{\mu} \notin i_\mu^{0.95}$. A sample yields a range of hypotheses to reject. For instance, if $z_\mu = 2.03$, we have $i_\mu^{0.95} = [0.07, 3.99]$ and reject $\mu = \tilde{\mu}$ for all $\tilde{\mu} \leq 0.07$ and $\tilde{\mu} \geq 3.99$. The criterion for rejecting $\mu = 0$ can be written as

$$\begin{aligned} 0 \notin i_\mu^{0.95} \quad &\Leftrightarrow \quad 0 \notin [z_\mu \pm 1.96] \\ &\Leftrightarrow \quad z_\mu \in (-\infty, -1.96) \cup (1.96, \infty) \\ &\Leftrightarrow \quad z_\mu \in \kappa_{0.05} \end{aligned}$$

The last expression is exactly the rejection rule for the null hypothesis $\mu = 0$ with significance level $\alpha = 0.05$. More generally, the correspondence between the two-tailed test in Section 1 and a CI is

$$\text{reject } \mu = 0 \text{ iff } z_\mu \in \kappa_\alpha \qquad \Leftrightarrow \qquad \text{reject } \mu = 0 \text{ iff } 0 \notin i_\mu^{1-\alpha}$$

B.1 Common Symmetric Form

A common situation is we have N iid random variables $X_1 \dots X_N \in \mathbb{R}$ from an unknown distribution with mean μ, and for some scaled standard deviation estimate $S > 0$ we have

$$\frac{\bar{X}_N - \mu}{S} \sim \textbf{Sym}$$

where $\bar{X}_N = (1/N) \sum_{i=1}^{N} X_i$ and **Sym** is a known symmetric distribution centered at 0. In this case

$$\Pr\left(-F_{\textbf{Sym}}^{-1}\left(1 - \frac{\alpha}{2}\right) \leq \frac{\bar{X}_N - \mu}{S} \leq F_{\textbf{Sym}}^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

[3] For instance, if each hypothesis is independent then

$$\textbf{FWER} = 1 - (1 - \alpha^{(i)})^m$$

where $(1 - \alpha^{(i)})^m$ is the probability of making no error in m tests.

where $F^{-1}_{\mathbf{Sym}}(\gamma)$ is the quantile function of **Sym**.[4] Note that $F^{-1}_{\mathbf{Sym}}(1-\frac{\alpha}{2})$ is the upper critical value for **Sym** at significance level α. (We could also use $-F^{-1}_{\mathbf{Sym}}(\alpha/2)$ by the symmetry of the distribution.) Thus a $(1-\alpha)$-CI for μ is given by

$$\Pr\left(\mu \in \left[\bar{X}_N \pm F^{-1}_{\mathbf{Sym}}\left(1-\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{N}}\right]\right) = 1-\alpha \tag{10}$$

For instance, if $X_1 \dots X_N \sim \mathcal{N}(\mu, \sigma^2)$, then $\bar{X}_N \sim \mathcal{N}(\mu, \sigma^2/N)$ and thus $\frac{\bar{X}_N-\mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)$. In this case

$$\Pr\left(\mu \in \left[\bar{X}_N \pm F^{-1}_{\mathcal{N}(0,1)}\left(1-\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{N}}\right]\right) = 1-\alpha \tag{11}$$

where we can get the value of $F^{-1}_{\mathcal{N}(0,1)}(\gamma)$ from the z-table. Similarly, we also know that $\frac{\bar{X}_N-\mu}{\bar{S}_N/\sqrt{N}} \sim \tau(N-1)$ where $\bar{S}_N$ is the sample standard deviation (Lemma H.2), thus

$$\Pr\left(\mu \in \left[\bar{X}_N \pm F^{-1}_{\tau(N-1)}\left(1-\frac{\alpha}{2}\right)\frac{\bar{S}_N}{\sqrt{N}}\right]\right) = 1-\alpha \tag{12}$$

where we can get the value of $F^{-1}_{\tau(N-1)}(\gamma)$ from the t-table.

B.1.1 Estimation

Given iid samples $X_1 \dots X_N$ from an unknown distribution $\mathbf{Unk}(\mu, \sigma^2)$ with mean μ and variance σ^2, how we estimate $I^{1-\alpha}_{\mu}$ depends on what assumptions we make.

- **Normal, known variance.** If $\mathbf{Unk}(\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2)$ and we know σ^2, we can calculate the exact $[\bar{X}_N \pm F^{-1}_{\mathcal{N}(0,1)}\left(1-\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{N}}]$ in (11).
- **Normal, unknown variance.** If $\mathbf{Unk}(\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2)$ and we don't know σ^2, we can calculate the exact $[\bar{X}_N \pm F^{-1}_{\tau(N-1)}\left(1-\frac{\alpha}{2}\right)\frac{\bar{S}_N}{\sqrt{N}}]$ in (12).
- **Known variance.** If we know σ^2 and N is large enough, we can use $[\bar{X}_N \pm F^{-1}_{\mathcal{N}(0,1)}\left(1-\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{N}}]$ in (11) as an approximation. By the CLT, $\bar{X}_N$ is approximately distributed as $\mathcal{N}(\mu, \frac{\sigma^2}{N})$ as $N \to \infty$ no matter what the underlying distribution of X_i is, so $\frac{\bar{X}_N-\mu}{\sigma/\sqrt{N}}$ is approximately distributed as $\mathcal{N}(0,1)$.
- **Unknown variance.** If we don't know σ^2, we can get the sample estimate $\bar{S}_N$ and use $[\bar{X}_N \pm F^{-1}_{p}\left(1-\frac{\alpha}{2}\right)\frac{\bar{S}_N}{\sqrt{N}}]$. Should p be $\mathcal{N}(0,1)$ or $\tau(N-1)$? If N is large it doesn't matter. Since $\bar{S}_N \xrightarrow{p} \sigma$ by the law of large numbers and $\bar{X}_N \xrightarrow{d} Z_{\mu, \frac{\sigma^2}{N}} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$ by the CLT, by Slutsky's theorem (Theorem C.1)

  $$\frac{\bar{X}_N-\mu}{\bar{S}_N/\sqrt{N}} \xrightarrow{d} \frac{Z_{\mu,\frac{\sigma^2}{N}}-\mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)$$

  so we can use $p = \mathcal{N}(0,1)$. But $\tau(N-1)$ converges to $\mathcal{N}(0,1)$ as $N \to \infty$ so we could also just use $p = \tau(N-1)$. But a typical choice is $p = \tau(N-1)$ (especially for small N) since it has the benefit of incorporating the uncertainty due to finite sample size.
- **Unknown variance (bootstrap).** If we don't know σ^2, there is another way of constructing a CI called bootstraipping that doesn't appealing to the CLT and is more flexible (e.g., it can handle statistics other than the mean like the median more easily). The idea is to "pretend $\bar{X}_N$ is μ" and draw B samples $\tilde{X}^1_N \dots \tilde{X}^B_N$ of $\bar{X}_N$ from the empirical distribution.[5] This way we have the ability to simulate the distribution of $\tilde{X}_N - \bar{X}_N \approx \bar{X}_N - \mu$, and this information is all we need to estimate a CI.

[4] Recall that for any distribution p, the quantile function F^{-1}_p is the mapping $\gamma \mapsto x_\gamma$ such that $\Pr_{x\sim p}(x \le x_\gamma) = \gamma$.

[5] More specifically, each $\tilde{X}^i_N$ is computed by drawing N iid samples from $\mathrm{Unif}(X_1 \dots X_N)$ with replacement and averaging them. It is important that we use the same sample size N to preserve the variance of the statistic.

We can also use non-asymptotic bounds to establish "loose" CIs. If $\mathbf{Unk}(\mu, \sigma^2)$ is almost surely bounded in $[a, b]$ (i.e., $\Pr(X_i \in [a, b]) = 1$) Hoeffding's inequality gives us

$$\Pr\left(\left|\bar{X}_N - \mu\right| > \epsilon\right) \leq 2\exp\left(-\frac{2N\epsilon^2}{b-a}\right) \quad \Leftrightarrow \quad \Pr\left(\mu \in \left[\bar{X}_N \pm \epsilon\right]\right) \geq 1 - \underbrace{2\exp\left(-\frac{2N\epsilon^2}{b-a}\right)}_{\alpha}$$

$$\Leftrightarrow \quad \Pr\left(\mu \in \left[\bar{X}_N \pm \sqrt{\frac{(b-a)\log\frac{2}{\alpha}}{2N}}\right]\right) \geq 1 - \alpha$$

While this generally applies to any (a.s. bounded) distribution without appealing to the CLT, it can be loose (note the inequality instead of equality).[6]

C Convergence of Random Variables

Let $Y_N, Y \in \mathbb{R}^d$ denote random vectors. Here Y can be constant (i.e., follows a point-mass distribution). We say Y_N converges to Y as $N \to \infty$ if any of the following is true.

- **Convergence in distribution.** $Y_N \xrightarrow{d} Y$ means $\lim_{N\to\infty} \text{CDF}_{Y_N}(y) = \text{CDF}_Y(y)$ for all continuity points $y \in \mathbb{R}^d$ of CDF_Y. In this note, we also write $Y_N \overset{\text{approx.}}{\sim} \text{PDF}_Y$ as $N \to \infty$.
- **Convergence in probability.** $Y_N \xrightarrow{p} Y$ means $\lim_{N\to\infty} \Pr(||Y_N - Y|| > \epsilon) = 0$ for all $\epsilon > 0$.
- **Almost sure convergence.** $Y_N \xrightarrow{\text{a.s.}} Y$ means $\Pr(\lim_{N\to\infty} Y_N = Y) = 1$. We also say "$Y_N$ converges to Y with probability 1". This is the strongest form of convergence.

For instance, let $X_1 \dots X_N \in \mathbb{R}^d$ denote iid random vectors with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}_{\succ 0}$. Define $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and $Y_N = \sqrt{N}\Sigma^{-1/2}(\bar{X}_N - \mu)$. Let $Y \in \mathbb{R}^d$ be distributed as $\mathcal{N}(0_d, I_{d\times d})$. We have[7]

$$\begin{aligned}
&Y_N \xrightarrow{d} Y && \text{(CLT)} \\
&\bar{X}_N \xrightarrow{p} \mu && \text{(weak law of large numbers)} \\
&\bar{X}_N \xrightarrow{\text{a.s.}} \mu && \text{(strong law of large numbers)}
\end{aligned}$$

We also have $\bar{X}_N \xrightarrow{d} \mu$ since $\bar{X}_N \overset{\text{approx.}}{\sim} \mathcal{N}(\mu, \frac{1}{N}\Sigma)$ as $N \to \infty$. In general,

$$\begin{aligned}
Y_N \xrightarrow{p} Y \quad &\Rightarrow \quad Y_N \xrightarrow{d} Y \\
Y_N \xrightarrow{\text{a.s.}} Y \quad &\Rightarrow \quad Y_N \xrightarrow{p} Y
\end{aligned}$$

Some properties of convergence we will use are as follows:

- **Continuous mapping theorem.** A continuous function $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ preserves convergence of any type. More generally, if f has discontinuity points $S \subset \mathbb{R}^d$ satisfying $\Pr(Y \in S) = 0$

$$Y_N \to Y \quad \Rightarrow \quad f(Y_N) \to f(Y) \tag{13}$$

- If $Y'_N \in \mathbb{R}^{d'}$ is an additional random variable and $y' \in \mathbb{R}^{d'}$ is a constant vector

$$Y_N \xrightarrow{d} Y, \;\; Y'_N \xrightarrow{d} y' \quad \Rightarrow \quad (Y_N, Y'_N) \xrightarrow{d} (Y, y') \tag{14}$$

[6] Even for an unbounded distribution we can derive a similar bound by using the knowledge of the distribution and variance. For instance, if $\mathbf{Unk}(\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2)$ and σ^2 is known we can use the fact that $\Pr(|\bar{X}_N - \mu| > \epsilon) \leq 2\exp(-N\epsilon^2/(2\sigma^2))$ (see Section 1.2 of this note) and have

$$\Pr\left(\mu \in \left[\bar{X}_N \pm \sqrt{2\log\frac{2}{\alpha}}\frac{\sigma}{\sqrt{N}}\right]\right) \geq 1 - \alpha$$

which is the same as the true CI in (11) except that we replace $F^{-1}_{\mathcal{N}(0,1)}\left(1 - \frac{\alpha}{2}\right)$ with $\sqrt{2\log\frac{2}{\alpha}}$ (hence loose).

[7] The distinction between the weak vs large law of large numbers is mostly theoretical. The weak law considers the limit of a probability and is an immediate consequence of Hoeffding's inequality. The strong law considers the probability of a limit and requires tools of measure theory to prove it.

- If $Y'_N \in \mathbb{R}^{d'}$ is an additional random variable,

$$Y_N \xrightarrow{p} Y, \;\; Y'_N \xrightarrow{p} Y' \quad \Rightarrow \quad (Y_N, Y'_N) \xrightarrow{p} (Y, Y') \tag{15}$$

The limiting properties can be unintuitive and require great care (see here for a list). For instance, if Y' is a random variable with nonzero variance, in general

$$Y_N \xrightarrow{d} Y, \;\; Y'_N \xrightarrow{d} Y' \quad \not\Rightarrow \quad (Y_N, Y'_N) \xrightarrow{d} (Y, Y')$$

whereas a similar statement (15) holds when the convergence is in probability. Note that there is no restriction on Y or Y' to be *non*-constant. In particular, (15) implies that for constants $y \in \mathbb{R}^d$ and $y' \in \mathbb{R}^{d'}$,

$$Y_N \xrightarrow{p} Y, \;\; Y'_N \xrightarrow{p} y' \quad \Rightarrow \quad (Y_N, Y'_N) \xrightarrow{p} (Y, y') \tag{16}$$

$$Y_N \xrightarrow{p} y, \;\; Y'_N \xrightarrow{p} Y' \quad \Rightarrow \quad (Y_N, Y'_N) \xrightarrow{p} (y, Y') \tag{17}$$

$$Y_N \xrightarrow{p} y, \;\; Y'_N \xrightarrow{p} y' \quad \Rightarrow \quad (Y_N, Y'_N) \xrightarrow{p} (y, y') \tag{18}$$

Theorem C.1 (Slutsky's theorem)**.** Let $Y_N, Y'_N, Y \in \mathbb{R}$ be random variables satisfying $Y_N \xrightarrow{d} Y$ and $Y'_N \xrightarrow{p} y'$ for some constant $y' \in \mathbb{R}$. Then

$$Y_N + Y'_N \xrightarrow{d} Y + y' \qquad\qquad Y_N Y'_N \xrightarrow{d} y'Y$$

If $y' \neq 0$, we also have $Y_N / Y'_N \xrightarrow{d} Y/y'$.

Proof. By the premise and (14) we have $(Y_N, Y'_N) \xrightarrow{d} (Y, y')$. Since $(y, y') \mapsto y + y'$, $(y, y') \mapsto y \times y'$, $(y, y') \mapsto y/y'$ are continuous mappings the theorem follows from (13). □

Theorem C.2 (Slutsky's theorem, convergence in probability)**.** Let $Y_N, Y'_N, Y \in \mathbb{R}$ be random variables satisfying $Y_N \xrightarrow{p} Y$ and $Y'_N \xrightarrow{p} y'$ for some constant $y' \in \mathbb{R}$. Then

$$Y_N + Y'_N \xrightarrow{p} Y + y' \qquad\qquad Y_N Y'_N \xrightarrow{p} y'Y$$

If $y' \neq 0$, we also have $Y_N / Y'_N \xrightarrow{p} Y/y'$.

Proof. By the premise and (16) we have $(Y_N, Y'_N) \xrightarrow{p} (Y, y')$. Since $(y, y') \mapsto y + y'$, $(y, y') \mapsto y \times y'$, $(y, y') \mapsto y/y'$ are continuous mappings the theorem follows from (13). □

D Error Decomposition

Error decomposition refers to expressing a loss function in smaller pieces to shed insight on what it takes to minimize it. It typically involves a squared error (so regression) and a trick to use the Pythagorean theorem to simplify the expression. For instance, if θ_s is the parameter estimate using data s and $\bar{\theta} = \mathbf{E}_{s \sim \mathbf{dat}_S}[\theta_s]$, the expected squared norm of the difference between the "best" parameter θ^* and θ_s can be written as

$$\underset{s \sim \mathbf{dat}_S}{\mathbf{E}} \left[||\theta^* - \theta_s||^2 \right] = \underbrace{||\theta^* - \bar{\theta}||^2}_{\text{bias/approximation error}} + \underbrace{\underset{s \sim \mathbf{dat}_S}{\mathbf{E}} \left[||\bar{\theta} - \theta_s||^2 \right]}_{\text{variance/estimation error}}$$

This is useful because it tells us that for the estimator θ_s to be successful in recovering θ^* (in squared norm of the difference), it has to have *both* of the two properties: (1) $\theta_s \approx \theta^*$ if given enough data, and (2) the estimate shouldn't fluctuate wildly. Error decomposition is studied in specific contexts and it can feel a bit all over the place, so here's a compilation of some well-known results.

D.1 Bias-Variance Tradeoff for the Mean Squared Error

Let $\mathbf{pop}_{XY}$ denote a joint distribution over $(x, y) \in \mathbb{R}^d \times \mathbb{R}$. The **mean squared error (MSE)** of a regressor $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathrm{MSE}(f) := \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}} \left[(y - f(x))^2\right] \tag{19}$$

Lemma D.1. The unique minimizer of MSE (19) is given by

$$f^*(x) := \mathop{\mathbf{E}}_{y\sim\mathbf{pop}_{Y|X}(\cdot|x)} [y] \qquad\qquad \forall x \in \mathbb{R}^d$$

Proof. Define $J : \mathbb{R} \to \mathbb{R}$ by

$$J(z) := \mathop{\mathbf{E}}_{y\sim\mathbf{pop}_{Y|X}(\cdot|x)} \left[(y - z)^2\right]$$

We have $J'(z) = -2(f^*(x) - z)$ and $J''(z) = 2 > 0$, so it is uniquely minimized by $z = f^*(x)$. The statement follows from the observsation $\mathrm{MSE}(f) = \mathbf{E}_{x\sim\mathbf{pop}_X}[J(f(x))]$. □

Denote the smallest achievable MSE by[8]

$$\epsilon_{\min} := \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}} \left[(y - f^*(x))^2\right]$$

Let A denote a learning algorithm that yields a regressor $f_s^A : \mathbb{R}^d \to \mathbb{R}$ given a (finite) training dataset s. Let $f^A : \mathbb{R}^d \to \mathbb{R}$ denote an average regressor from A with respect to some distribution $\mathbf{dat}_S$ over training datasets, that is

$$f^A(x) := \mathop{\mathbf{E}}_{s\sim\mathbf{dat}_S} \left[f_s^A(x)\right] \qquad\qquad \forall x \in \mathbb{R}^d$$

Theorem D.2.

$$\mathop{\mathbf{E}}_{s\sim\mathbf{dat}_S} \left[\mathrm{MSE}(f_s^A)\right] = \underbrace{\epsilon_{\min}}_{\text{irreducible error}} + \mathop{\mathbf{E}}_{x\sim\mathbf{pop}_X} \left[\underbrace{(f^*(x) - f^A(x))^2}_{\text{squared bias}} + \underbrace{\mathop{\mathrm{Var}}_{s\sim\mathbf{dat}_S} \left(f_s^A(x)\right)}_{\text{variance}}\right]$$

Proof. By Lemma K.1, we can write

$$\mathop{\mathbf{E}}_{s\sim\mathbf{dat}_S} \left[\mathrm{MSE}(f_s^A)\right] = \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}} \left[(y - f^*(x))^2\right] + \mathop{\mathbf{E}}_{\substack{s\sim\mathbf{dat}_S \\ x\sim\mathbf{pop}_X}} \left[(f^*(x) - f_s^A(x))^2\right]$$

where the first term is $\epsilon_{\min}$. We further break down the second term as follows:

$$\mathop{\mathbf{E}}_{\substack{s\sim\mathbf{dat}_S \\ x\sim\mathbf{pop}_X}} \left[(f^*(x) - f_s^A(x))^2\right] = \mathop{\mathbf{E}}_{x\sim\mathbf{pop}_X} \left[(f^*(x) - f^A(x))^2\right] + \mathop{\mathbf{E}}_{\substack{s\sim\mathbf{dat}_S \\ x\sim\mathbf{pop}_X}} \left[(f^A(x) - f_s^A(x))^2\right]$$

where the cross product term disappears since

$$\mathop{\mathbf{E}}_{s\sim\mathbf{dat}_S} \left[(f^*(x) - f^A(x))(f^A(x) - f_s^A(x))\right] = (f^*(x) - f^A(x)) \mathop{\mathbf{E}}_{s\sim\mathbf{dat}_S} \left[(f^A(x) - f_s^A(x))\right] = 0$$

□

If A is a consistent estimator of the optimal regressor, the bias term vanishes and we have a simpler decomposition consisting of only the estimation error (plus $\epsilon_{\min}$):

$$\mathop{\mathbf{E}}_{s\sim\mathbf{dat}_S} \left[\mathrm{MSE}(f_s^A)\right] = \epsilon_{\min} + \mathop{\mathbf{E}}_{x\sim\mathbf{pop}_X} \left[\mathop{\mathrm{Var}}_{s\sim\mathbf{dat}_S} \left(f_s^A(x)\right)\right] \tag{20}$$

[8] A common assumption in regression is that $y = g(x) + z$ where $g : \mathbb{R}^d \to \mathbb{R}$ is the ground-truth mapping and $z \sim \nu$ is a random noise with mean zero ("white") and variance σ^2. Under this assumption we have $\epsilon_{\min} = \sigma^2$ and $f^* = g$ (since $f^*(x) = \mathbf{E}_{z\sim\nu}[g(x) + z] = g(x)$). But we don't need this assumption for the following result.

D.2 MSE decomposition for linear regression

Consider MSE (19) for linear regression. Let

$$\mathrm{MSE}_{\mathrm{lin}}(w) := \mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} \left[(y - w^\top x)^2\right] \tag{21}$$

For simplicity, assume that the second moment of $x \sim \mathbf{pop}_X$ is positive definite.

Lemma D.3. The unique minimizer of $\mathrm{MSE}_{\mathrm{lin}}$ (21) is given by

$$w^* := \mathop{\mathbf{E}}_{x\sim \mathbf{pop}_X} \left[xx^\top\right]^{-1} \mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} [yx]$$

Proof. Denoting the RHS of (21) as a function $J : \mathbb{R}^d \to \mathbb{R}$ of w, we have

$$\nabla J(w) = -2 \mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} \left[yx - xx^\top w\right]$$
$$\nabla^2 J(w) = 2 \mathop{\mathbf{E}}_{x\sim \mathbf{pop}_X} \left[xx^\top\right] \succ 0$$

where the Hessian is positive definite by premise. Thus the unique minimizer is the stationary point w^* satisfying $\nabla J(w^*) = 0$ which is

$$w^* = \mathop{\mathbf{E}}_{x\sim \mathbf{pop}_X} \left[xx^\top\right]^{-1} \mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} [yx]$$

□

Lemma D.4. For any $A \in \mathbb{R}^{m\times d}$,

$$\mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} \left[(y - (w^*)^\top x)Ax\right] = 0_m$$

Proof.

$$\begin{aligned}
\mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} \left[(y - (w^*)^\top x)Ax\right] &= \mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} \left[Axy - Axx^\top \mathop{\mathbf{E}}_{x\sim \mathbf{pop}_X} \left[xx^\top\right]^{-1} \mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} [yx]\right] \\
&= A \mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} [xy] - A \mathop{\mathbf{E}}_{x\sim \mathbf{pop}_X} \left[xx^\top\right] \mathop{\mathbf{E}}_{x\sim \mathbf{pop}_X} \left[xx^\top\right]^{-1} \mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} [yx] \\
&= 0_m
\end{aligned}$$

□

Corollary D.5. If $x_1 = 1$ for all $x \sim \mathbf{pop}_X$, then $\mathbf{E}_{(x,y)\sim \mathbf{pop}_{XY}}[y - (w^*)^\top x] = 0$. In particular, $\langle w^*, \mu_X\rangle = \mu_Y$ where $\mu_X \in \mathbb{R}^d$ and $\mu_Y \in \mathbb{R}$ are the population means of input and label.

Proof. The statement follows from Lemma D.4 with $A = [1, 0, \ldots, 0] \in \mathbb{R}^{1\times d}$. □

Corollary D.6. If $x_1 = 1$ for all $x \sim \mathbf{pop}_X$, then for any $A \in \mathbb{R}^{m\times d}$,

$$\mathop{\mathrm{Cor}}_{(x,y)\sim \mathbf{pop}_{XY}} \left(y - (w^*)^\top x, [Ax]_j\right) = 0 \qquad \forall j \in \{1 \ldots m\}$$

Proof. Since $\mathrm{Cor}\,(X, Y) = \mathrm{Cov}\,(X, Y)\,/\sigma_X/\sigma_Y$, it is sufficient to show $\mathrm{Cov}\left(y - (w^*)^\top x, [Ax]_j\right) = 0$.

$$\begin{aligned}
\mathop{\mathrm{Cov}}_{(x,y)\sim \mathbf{pop}_{XY}} \left(y - (w^*)^\top x, [Ax]_j\right) &= \mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} \left[(y - (w^*)^\top x)[Ax]_j\right] - \mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} \left[y - (w^*)^\top x\right] \mathop{\mathbf{E}}_{x\sim \mathbf{pop}_X} [[Ax]_j] \\
&= \mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} \left[(y - (w^*)^\top x)[Ax]_j\right] && \text{(Corollary D.5)} \\
&= 0 && \text{(Lemma D.4)}
\end{aligned}$$

□

Remark. Corollary D.6 says the error of the optimal linear regressor w^* is uncorrelated with (every dimension of) any linear transformation of the input. For instance, if we draw many samples $(x, y) \sim \mathbf{pop}_{XY}$ and plot the errors of w^*, we will see no linear dependence between the input values and the error values. However, this does not mean that the errors are independent of the input values.

Theorem D.7. For any $w \in \mathbb{R}^d$,

$$\text{MSE}_{\text{lin}}(w) = \text{MSE}_{\text{lin}}(w^*) + \mathop{\mathbf{E}}_{x \sim \mathbf{pop}_X}\left[((w^*)^\top x - w^\top x)^2\right] \tag{22}$$

Proof.

$$\begin{aligned}
\text{MSE}_{\text{lin}}(w) &= \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}}\left[(y - w^\top x)^2\right] \\
&= \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}}\left[(y - (w^*)^\top x)^2\right] + \mathop{\mathbf{E}}_{x\sim\mathbf{pop}_X}\left[((w^*)^\top x - w^\top x)^2\right] \\
&\qquad + 2\mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}}\left[(y - (w^*)^\top x)((w^*)^\top x - w^\top x)\right]
\end{aligned}$$

The first term is $\text{MSE}_{\text{lin}}(w^*)$. The last term is zero by Lemma D.4. □

Let $w_s \in \mathbb{R}^d$ denotes the parameter trained on a (finite) dataset $s \sim \mathbf{dat}_S$. Using w_s in (22) and taking an expectation over s, we have

$$\mathop{\mathbf{E}}_{s\sim\mathbf{dat}_S}\left[\text{MSE}_{\text{lin}}(w_s)\right] = \underbrace{\text{MSE}_{\text{lin}}(w^*)}_{\text{approximation error}} + \underbrace{\mathop{\mathbf{E}}_{\substack{s\sim\mathbf{dat}_S \\ x\sim\mathbf{pop}_X}}\left[((w^*)^\top x - w_s^\top x)^2\right]}_{\text{estimation error}} \tag{23}$$

which shows the tradeoff between how much we lose by restricting ourselves to linear models (approximation) and how easy it is to estimate the best one (estimation). We can further break down the estimation error into bias and variance

$$\mathop{\mathbf{E}}_{\substack{s\sim\mathbf{dat}_S \\ x\sim\mathbf{pop}_X}}\left[((w^*)^\top x - w_s^\top x)^2\right] = \mathop{\mathbf{E}}_{x\sim\mathbf{pop}_X}\left[((w^*)^\top x - \bar{w}^\top x)^2 + \mathop{\text{Var}}_{s\sim\mathbf{dat}_S}\left(w_s^\top x\right)\right]$$

where $\bar{w} = \mathbf{E}_{s\sim\mathbf{dat}_S}[w_s]$. Thus (23) can be seen as a special case of Theorem D.2 for linear regression with the irreducible error $\epsilon_{\min} = \text{MSE}_{\text{lin}}(w^*)$.

D.3 Sum of Squared Errors for Least-Squares Linear Regression

A linear regressor with parameter $w \in \mathbb{R}^d$ defines a mapping $\mathbb{R}^d \mapsto \mathbb{R}$ by

$$f_w(x) = w^\top x = w_1 + \sum_{i=2}^{d} w_i x_i$$

where we adopt the convention that $x_1 = 1$ so that we don't have to introduce a separate bias parameter. Let $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ denote N inputs as rows and $\boldsymbol{y} \in \mathbb{R}^N$ their corresponding targets.[9] For any subset $P \subseteq \{1 \dots d\}$ we will write $\boldsymbol{X}_P \in \mathbb{R}^{N\times|P|}$ to denote the corresponding columns of $\boldsymbol{X}$. The **least squares estimation (LSE)** of w is

$$\hat{w}_N = \mathop{\arg\min}_{w\in\mathbb{R}^d} ||\boldsymbol{y} - \boldsymbol{X}w||^2 = \boldsymbol{X}^+\boldsymbol{y} \tag{24}$$

where $\boldsymbol{X}^+ \in \mathbb{R}^{d\times N}$ is a pseudo-inverse of $\boldsymbol{X}$. The last term is the unique solution of the optimization problem, so we will equate $\hat{w}_N = \boldsymbol{X}^+\boldsymbol{y}$. Let $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{w}_N$ and $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ denote the predictions and errors (aka. residuals) of LSE on the N inputs. A characterizing property of LSE is that for any $A \in \mathbb{R}^{m\times d}$

$$\begin{aligned}
(\boldsymbol{X}A^\top)^\top\hat{\boldsymbol{\epsilon}} &= A\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{X}^+\boldsymbol{y}) \\
&= A\boldsymbol{X}^\top(I_{N\times N} - \boldsymbol{X}\boldsymbol{X}^+)\boldsymbol{y} \\
&= 0_m
\end{aligned} \tag{25}$$

[9] We use boldface to denote vectors/matrices of sample size, to prevent confusion with random variables and scalars like X and y in other sections.

because $I_{N\times N} - \boldsymbol{X}\boldsymbol{X}^+$ is the orthogonal projection onto null($\boldsymbol{X}^\top$). Thus the LSE residuals are uncorrelated with any linear transformation of the training data (Lemma D.4 gives an infinite-sample version of this statement). In particular,

- For any $P \subseteq \{1 \dots d\}$ with $|P| = p$

$$\boldsymbol{X}_P^\top \hat{\boldsymbol{\epsilon}} = 0_p$$

  Using $P = \{1\}$ gives the well-known property of LSE that the errors are uncorrelated with the input: $\mathbf{1}_N^\top \hat{\boldsymbol{\epsilon}} = 0$.

- If $\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)} \in \mathbb{R}^N$ are any predictions by linear regressors on $\boldsymbol{X}$

$$(\boldsymbol{y}^{(1)} - \boldsymbol{y}^{(2)})^\top \hat{\boldsymbol{\epsilon}} = 0 \tag{26}$$

Let $\hat{\boldsymbol{y}}_P \in \mathbb{R}^N$ denote LSE predictions using *only* $\boldsymbol{X}_P$ (aka. partial LSE). Applying (26) we have

$$\underbrace{||\boldsymbol{y} - \hat{\boldsymbol{y}}_P||^2}_{\text{total sum of squares (TSS)}} = \underbrace{||\hat{\boldsymbol{\epsilon}}||^2}_{\text{residual sum of squares (RSS)}} + \underbrace{||\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_P||^2}_{\text{explained sum of squares (ESS)}} \tag{27}$$

which expresses the variation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_P$ (TSS) as the sum of the variation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ (RSS) and the variation between $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{y}}_P$ (ESS). A well-known special case is given with $P = \{1\}$

$$||\boldsymbol{y} - \bar{\boldsymbol{y}}||^2 = ||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2 + ||\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}||^2 \tag{28}$$

where $\bar{\boldsymbol{y}}_i = (1/N)\sum_{i=1}^N y_i$ is the average target value. RSS, ESS, and TSS play a central role in regression analysis. The following technical lemma will be useful.

Lemma D.8. Let $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ be a full-rank matrix with the first column of $\mathbf{1}_N$. Let $Q \subseteq P \subseteq \{1 \dots d\}$ with $|Q| = q$ and $|P| = p$. Then

$$I_{N\times N} = \underbrace{\left(I_{N\times N} - \boldsymbol{X}\boldsymbol{X}^+\right)}_{A_{\boldsymbol{X}}} + \underbrace{\left(\boldsymbol{X}\boldsymbol{X}^+ - \boldsymbol{X}_P\boldsymbol{X}_P^+\right)}_{B_{\boldsymbol{X}}^P} + \underbrace{\left(\boldsymbol{X}_P\boldsymbol{X}_P^+ - \boldsymbol{X}_Q\boldsymbol{X}_Q^+\right)}_{C_{\boldsymbol{X}}^{P,Q}} + \boldsymbol{X}_Q\boldsymbol{X}_Q^+$$

is a subspace decomposition of $\mathbb{R}^N$ (Lemma J.3) where $A_{\boldsymbol{X}}$, $B_{\boldsymbol{X}}^P$, $C_{\boldsymbol{X}}^{P,Q}$, and $\boldsymbol{X}_Q\boldsymbol{X}_Q^+$ are orthogonal projections with rank $N-d$, $d-p$, $p-q$, and q. Furthermore, given targets $\boldsymbol{y} \in \mathbb{R}^N$, if $\hat{\boldsymbol{y}}$ denotes the LSE predictions using all columns of $\boldsymbol{X}$ with residuals $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{y}}_P$ denotes the LSE predictions using only the columns P of $\boldsymbol{X}$ (similarly for Q), then

$$||\hat{\boldsymbol{\epsilon}}||^2 = \boldsymbol{y}^\top A_{\boldsymbol{X}}\boldsymbol{y} \qquad \text{(RSS)} \tag{29}$$

$$||\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_P||^2 = \boldsymbol{y}^\top B_{\boldsymbol{X}}^P\boldsymbol{y} \qquad \text{(ESS)} \tag{30}$$

$$||\boldsymbol{y} - \hat{\boldsymbol{y}}_P||^2 = \boldsymbol{y}^\top \left(I_{N\times N} - \boldsymbol{X}_P\boldsymbol{X}_P^+\right)\boldsymbol{y} \qquad \text{(TSS)} \tag{31}$$

$$\left|\left|\hat{\boldsymbol{y}}_P - \hat{\boldsymbol{y}}_Q\right|\right|^2 = \boldsymbol{y}^\top C_{\boldsymbol{X}}^{P,Q}\boldsymbol{y} \tag{32}$$

In the lemma we have introduced $Q \subseteq P$ for generality. We may choose $Q = \varnothing$ and have $I_{N\times N} = A_{\boldsymbol{X}} + B_{\boldsymbol{X}}^P + \boldsymbol{X}_P\boldsymbol{X}_P^+$ with ranks $N-d$, $d-p$, and p. Also note that $B_{\boldsymbol{X}}^P = C_{\boldsymbol{X}}^{I_{N\times N},P}$.

D.3.1 Distributional Properties of LSE

Assumption D.1. Let $X \sim \mathbf{pop}_X$ denote a d-dimensional random vector where $d \geq 2$ and $X_1 = 1$. Assume the second moment $\mathbf{E}\left[XX^\top\right] \in \mathbb{R}^{d\times d}$ is invertible.[10] Let $w_{\text{true}} \in \mathbb{R}^d$, $Z_{\sigma^2} \sim \mathcal{N}(0, \sigma^2)$, and

$$\begin{aligned} Y &= w_{\text{true}}^\top X + Z_{\sigma^2} \\ &= [w_{\text{true}}]_1 + [w_{\text{true}}]_2 X_2 + \cdots + [w_{\text{true}}]_d X_d + Z_{\sigma^2} \end{aligned}$$

Let $(x_1, y_1) \ldots (x_N, y_N) \in \mathbb{R}^d \times \mathbb{R}$ denote iid samples of (X, Y) where $N > d$. Let $\boldsymbol{X} = (x_1^\top \ldots x_N^\top) \in \mathbb{R}^{N\times d}$ and $\boldsymbol{y} = (y_1 \ldots y_N) \in \mathbb{R}^N$. For simplicity we assume $\boldsymbol{X}$ is full-rank.[11] We denote the LSE parameter by $\hat{w}_N = \boldsymbol{X}^+\boldsymbol{y}$, the residual by $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{w}_N$, and the error vector by $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X}w_{\text{true}}$.

Lemma D.9. Under Assumption D.1, $\hat{w}_N \in \mathbb{R}^d$ is independent of $\hat{\boldsymbol{\epsilon}} \in \mathbb{R}^N$.

Proof. Since $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}w_{\text{true}}, \sigma^2 I_{N\times N})$, $\hat{w}_N = \boldsymbol{X}^+\boldsymbol{y}$, $\hat{\boldsymbol{\epsilon}} = (I_{N\times N} - \boldsymbol{X}\boldsymbol{X}^+)\boldsymbol{y}$, and

$$(\boldsymbol{X}^+)(\sigma^2 I_{N\times N})(I_{N\times N} - \boldsymbol{X}\boldsymbol{X}^+)^\top = \sigma^2(\boldsymbol{X}^+ - \boldsymbol{X}^+\boldsymbol{X}\boldsymbol{X}^+) = 0_{d\times N}$$

$\hat{w}_N$ and $\hat{\boldsymbol{\epsilon}}$ are independent by Lemma I.3. □

Lemma D.10. Under Assumption D.1,

$$\hat{\boldsymbol{\epsilon}} \sim \mathcal{N}\left(0_N, \sigma^2 A_{\boldsymbol{X}}\right)$$

Proof. Since $A_{\boldsymbol{X}}$ is the orthogonal projection onto range $(\boldsymbol{X})^\perp$,

$$\hat{\boldsymbol{\epsilon}} = A_{\boldsymbol{X}}\boldsymbol{y} \sim \mathcal{N}(A_{\boldsymbol{X}}\boldsymbol{X}w_{\text{true}}, \sigma^2 A_{\boldsymbol{X}}A_{\boldsymbol{X}}^\top) = \mathcal{N}(0_N, \sigma^2 A_{\boldsymbol{X}})$$

□

Lemma D.11. Under Assumption D.1,

$$\frac{||\hat{\boldsymbol{\epsilon}}||^2}{\sigma^2} \sim \chi^2(N - d)$$

Proof. By the quadratic form of the RSS (29),

$$\frac{||\hat{\boldsymbol{\epsilon}}||^2}{\sigma^2} = \frac{\boldsymbol{y}^\top A_{\boldsymbol{X}}\boldsymbol{y}}{\sigma^2} = \frac{(\boldsymbol{X}w_{\text{true}})^\top A_{\boldsymbol{X}}(\boldsymbol{X}w_{\text{true}})}{\sigma^2} + \frac{\boldsymbol{\epsilon}^\top A_{\boldsymbol{X}}\boldsymbol{\epsilon}}{\sigma^2}$$

$A_{\boldsymbol{X}}$ is the orthogonal projection onto range $(\boldsymbol{X})^\perp$ with rank $N - d$. Thus the first term is zero and the second term is distributed as $\chi^2(N - d)$ since $\boldsymbol{\epsilon} \sim \mathcal{N}(0_N, \sigma^2 I_{N\times N})$ (Lemma J.4). □

Corollary D.12. Under Assumption D.1, $\sigma^2 = \mathbf{E}\left[||\hat{\boldsymbol{\epsilon}}||^2\right]/(N - d)$.

Lemma D.13. Under Assumption D.1, if $Q \subset P \subseteq \{1 \ldots d\}$ are any nested subsets with sizes $q < p$,

$$\frac{\left|\left|\hat{\boldsymbol{y}}_P - \hat{\boldsymbol{y}}_Q\right|\right|^2}{\sigma^2} \sim \chi^2\left(p - q, \frac{w_{\text{true}}^\top \boldsymbol{X}^\top C_{\boldsymbol{X}}^{P,Q}\boldsymbol{X}w_{\text{true}}}{\sigma^2}\right)$$

[10] This is equivalent to assuming that the $d \times d$ matrix

$$\mathbf{E}\left[XX^\top\right] = \begin{bmatrix} 1 & \mathbf{E}[X_2] & \ldots & \mathbf{E}[X_d] \\ \mathbf{E}[X_2] & \mathbf{E}[X_2^2] & \ldots & \mathbf{E}[X_2X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{E}[X_d] & \mathbf{E}[X_2X_d] & \ldots & \mathbf{E}[X_d^2] \end{bmatrix}$$

has full rank. This is satisfied if features are not redundant. In particular, this is *not* the same as assuming an invertible covariance matrix $C_{XX} = \mathbf{E}\left[XX^\top\right] - \mathbf{E}[X]\mathbf{E}[X]^\top$ which would be impossible (since the first column/row is zero).

[11] This is always possible for a sufficiently large N, otherwise rank $\left(\mathbf{E}\left[XX^\top\right]\right) < d$.

Proof. By the quadratic form (32),

$$\frac{\left|\left|\hat{\boldsymbol{y}}_P - \hat{\boldsymbol{y}}_Q\right|\right|^2}{\sigma^2} = \frac{\boldsymbol{y}^\top C_{\boldsymbol{X}}^{P,Q} \boldsymbol{y}}{\sigma^2}$$

Since $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X} w_{\text{true}}, \sigma^2 I_{N\times N}\right)$ and $(C_{\boldsymbol{X}}^{P,Q}/\sigma^2)(\sigma^2 I_{N\times N}) = C_{\boldsymbol{X}}^{P,Q}$ is idempotent with rank $p-q$ (Lemma D.8), by Lemma H.1 we have the statement. □

Assumption D.2. Assume the same setting in Assumption D.1 where we have the ground-truth model $Y = w_{\text{true}}^\top X + Z_{\sigma^2}$. Let $Q \subset P \subseteq \{1 \dots d\}$ with sizes $q < p$, and assume that

$$[w_{\text{true}}]_i = 0 \qquad\qquad \forall i \notin Q$$

(In particular, if $Q = \{1\}$ then Y is independent of X.)

Corollary D.14. Under Assumption D.2,

$$\frac{\left|\left|\hat{\boldsymbol{y}}_P - \hat{\boldsymbol{y}}_Q\right|\right|^2}{\sigma^2} \sim \chi^2(p-q)$$

Proof. We only need to show that $w_{\text{true}}^\top \boldsymbol{X}^\top C_{\boldsymbol{X}}^{P,Q} \boldsymbol{X} w_{\text{true}} = 0$ (Lemma D.13). Since $[w_{\text{true}}]_i = 0$ for $i \notin Q$ and $Q \subset P$, we have

$$C_{\boldsymbol{X}}^{P,Q} \boldsymbol{X} w_{\text{true}} = C_{\boldsymbol{X}}^{P,Q} \boldsymbol{X}_Q w_{\text{true}}^Q = \left(\boldsymbol{X}_P \boldsymbol{X}_P^+ - \boldsymbol{X}_Q \boldsymbol{X}_Q^+\right) \boldsymbol{X}_Q w_{\text{true}}^Q = \boldsymbol{X}_Q w_{\text{true}}^Q - \boldsymbol{X}_Q w_{\text{true}}^Q = 0_N$$

where $w_{\text{true}}^Q \in \mathbb{R}^q$ is a subvector of $w_{\text{true}} \in \mathbb{R}^d$ indexed by Q. □

Theorem D.15. Under Assumption D.2,

$$\frac{\left|\left|\hat{\boldsymbol{y}}_P - \hat{\boldsymbol{y}}_Q\right|\right|^2 / (p-q)}{\left|\left|\hat{\boldsymbol{\epsilon}}\right|\right|^2 / (N-d)} \sim F(p-q, N-d)$$

Proof. We have

$$\frac{\left|\left|\hat{\boldsymbol{y}}_P - \hat{\boldsymbol{y}}_Q\right|\right|^2 / (p-q)}{\left|\left|\hat{\boldsymbol{\epsilon}}\right|\right|^2 / (N-d)} = \frac{\frac{\left|\left|\hat{\boldsymbol{y}}_P - \hat{\boldsymbol{y}}_Q\right|\right|^2}{\sigma^2} / (p-q)}{\frac{\left|\left|\hat{\boldsymbol{\epsilon}}\right|\right|^2}{\sigma^2} / (N-d)}$$

where $\frac{\left|\left|\hat{\boldsymbol{y}}_P - \hat{\boldsymbol{y}}_Q\right|\right|^2}{\sigma^2}$ is distributed as $\chi^2(p-q)$ (Corollary D.14) and $\frac{\left|\left|\hat{\boldsymbol{\epsilon}}\right|\right|^2}{\sigma^2}$ is distributed as $\chi^2(N-d)$ (Lemma D.11). It remains to show that they are independent. With the quadratic forms (29) and (32), this is equivalent to showing that $\boldsymbol{y}^\top A_{\boldsymbol{X}} \boldsymbol{y}$ and $\boldsymbol{y}^\top C_{\boldsymbol{X}}^{P,Q} \boldsymbol{y}$ are independent where $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X} w_{\text{true}}, \sigma^2 I_{N\times N}\right)$. This follows from Craig's theorem (Theorem I.2) since $A_{\boldsymbol{X}}\left(\sigma^2 I_{N\times N}\right) C_{\boldsymbol{X}}^{P,Q} = \sigma^2 A_{\boldsymbol{X}} C_{\boldsymbol{X}}^{P,Q} = 0_{N\times N}$ (Lemma D.8). □

Lemma D.16. Under Assumption D.1, conditioning on any $\boldsymbol{X}$,

$$\hat{w}_N \sim \mathcal{N}\left(w_{\text{true}}, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)$$

Proof. Since $\boldsymbol{\epsilon} \sim \mathcal{N}(0_N, \sigma^2 I_{N\times N})$,

$$\hat{w}_N = \boldsymbol{X}^+ \boldsymbol{y} = \boldsymbol{X}^+ \boldsymbol{X} w_{\text{true}} + \boldsymbol{X}^+ \boldsymbol{\epsilon} = w_{\text{true}} + \boldsymbol{X}^+ \boldsymbol{\epsilon} \sim \mathcal{N}(w_{\text{true}}, \sigma^2 \boldsymbol{X}^+ (\boldsymbol{X}^+)^\top) \tag{33}$$

where $\boldsymbol{X}^+ (\boldsymbol{X}^+)^\top = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. □

Lemma D.17. Under Assumption D.1, as $N \to \infty$

$$\hat{w}_N \overset{\text{approx.}}{\sim} \mathcal{N}\left(w_{\text{true}}, \frac{\sigma^2}{N} \mathbf{E}\left[X X^\top\right]^{-1}\right)$$

Corollary D.18. Under Assumption D.1, conditioning on any $\boldsymbol{X}$ we have

$$\frac{[\hat{w}_N]_j - [w_{\text{true}}]_j}{s_j} \sim \tau(N-d) \qquad\qquad s_j := \sqrt{\frac{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{j,j}}{N-d}}\,||\hat{\boldsymbol{\epsilon}}|| \tag{34}$$

for each $j = 1 \dots d$ (s_j is known as the **standard error** of $[\hat{w}_N]_j$).

Proof. Rewrite

$$\frac{[\hat{w}_N]_j - [w_{\text{true}}]_j}{s_j} = \frac{[\hat{w}_N]_j - [w_{\text{true}}]_j}{\sigma\sqrt{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{j,j}}} \times \frac{\sigma\sqrt{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{j,j}}}{s_j}$$

where $\frac{[\hat{w}_N]_j - [w_{\text{true}}]_j}{\sigma\sqrt{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{j,j}}} \sim \mathcal{N}(0,1)$ by Lemma D.17. The second term is rewritten as

$$\frac{\sigma\sqrt{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{j,j}}}{s_j} = \sqrt{\frac{\sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{j,j}}{s_j^2}} = \sqrt{\frac{N-d}{s_j^2 \frac{N-d}{\sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{j,j}}}} = \sqrt{\frac{N-d}{||\hat{\boldsymbol{\epsilon}}||^2 / \sigma^2}}$$

where $||\hat{\boldsymbol{\epsilon}}||^2 / \sigma^2 \sim \chi^2(N-d)$ (Lemma D.11). This is independent of $\frac{[\hat{w}_N]_j - [w_{\text{true}}]_j}{\sigma\sqrt{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{j,j}}}$ conditioning on $\boldsymbol{X}$ since $\hat{\boldsymbol{\epsilon}}$ and $\hat{w}_N$ are independent (Lemma D.9). Thus $\frac{[\hat{w}_N]_j - [w_{\text{true}}]_j}{s_j} \sim \tau(T-d)$ (46). □

One use of Corollary D.18 is establishing confidence intervals. By (10), we have with probability $1-p$ (assuming large enough N)

$$[w_{\text{true}}]_j \in \left[[\hat{w}_N]_j \pm F^{-1}_{\tau(N-d)}\left(1 - \frac{p}{2}\right) s_j \right]$$

E Simple Linear Regression

In **simple linear regression**, we consider Assumption D.1 with $d = 2$ but with an explicit bias parameter rather than assuming a constant feature dimension. Specifically, the model assumes $Z_{\sigma^2} \sim \mathcal{N}(0, \sigma^2)$ for some constant $\sigma^2 > 0$ and defines

$$Y = \alpha + \beta X + Z_{\sigma^2}$$

where $X \sim \mathbf{pop}_X$ is a scalar random variable with $\text{Var}(X) > 0$.[12] Let $(x_1, y_1) \dots (x_N, y_N) \in \mathbb{R} \times \mathbb{R}$ denote iid samples of (X, Y). The least-squares estimator is

$$(\hat{\alpha}_N, \hat{\beta}_N) = \operatorname*{arg\,min}_{(\alpha,\beta)\in\mathbb{R}^2} \sum_{i=1}^N (y_i - \alpha - \beta x_i)^2 \tag{35}$$

Of course this is a special case of (24) so the solution can be obtained by taking the pseudo-inverse of the data matrix augmented with a constant dimension. But the utility of considering simple linear regression is that we can derive a very explicit analytical solution. Denote the sample mean, the sample variance, and the sample correlation coefficient between X and Y by

$$\bar{x}_N = \frac{1}{N}\sum_{i=1}^N x_i \qquad \hat{\sigma}^2_{X,N} = \frac{1}{N-1}\sum_{i=1}^N (x_i - \bar{x}_N)^2 \qquad \hat{r}_{XY,N} = \frac{\sum_{i=1}^N (y_i - \bar{y}_N)(x_i - \bar{x}_N)}{\sqrt{\sum_{i=1}^N (x_i - \bar{x}_N)}\sqrt{\sum_{i=1}^N (y_i - \bar{y}_N)}}$$

$$\bar{y}_N = \frac{1}{N}\sum_{i=1}^N y_i \qquad \hat{\sigma}^2_{Y,N} = \frac{1}{N-1}\sum_{i=1}^N (y_i - \bar{y}_N)^2$$

[12] The condition that $\text{Var}(X) > 0$ follows from the condition that the second moment of $(1, X)$ is invertible in Assumption D.1. Otherwise, $\mathbf{E}[X^2] = \mathbf{E}[X]^2$ and

$$\mathbf{E}\left[\begin{bmatrix}1\\X\end{bmatrix}\begin{bmatrix}1 & X\end{bmatrix}\right] = \begin{bmatrix}1 & \mathbf{E}[X]\\ \mathbf{E}[X] & \mathbf{E}[X^2]\end{bmatrix} = \begin{bmatrix}1 & \mathbf{E}[X]\\ \mathbf{E}[X] & \mathbf{E}[X]^2\end{bmatrix}$$

is not invertible.

Lemma E.1.

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

Note that $\hat{\beta}_N = \hat{r}_{XY,N}(\hat{\sigma}_{Y,N}/\hat{\sigma}_{X,N})$. Thus the least-squares slope is proportional to the correlation coefficient and the ratio of standard deviations. If Y varies more than X, the slope becomes steeper to account for the uncertainty.[13]

Lemma E.2. In simple linear regression, as $N \to \infty$

$$\hat{\beta}_N \overset{\text{approx.}}{\sim} \mathcal{N}\left(\beta, \frac{\sigma^2}{N\text{Var}(X)}\right) \qquad \hat{\alpha}_N \overset{\text{approx.}}{\sim} \mathcal{N}\left(\alpha, \frac{\sigma^2 \mathbf{E}\left[X^2\right]}{N\text{Var}(X)}\right)$$

Proof. By Lemma D.17, as $N \to \infty$

$$\begin{bmatrix}\hat{\alpha}_N \\ \hat{\beta}_N\end{bmatrix} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\begin{bmatrix}\alpha \\ \beta\end{bmatrix}, \frac{\sigma^2}{N}\begin{bmatrix}1 & \mathbf{E}[X] \\ \mathbf{E}[X] & \mathbf{E}\left[X^2\right]\end{bmatrix}^{-1}\right)$$

The matrix inverse is

$$\begin{bmatrix}1 & \mathbf{E}[X] \\ \mathbf{E}[X] & \mathbf{E}\left[X^2\right]\end{bmatrix}^{-1} = \frac{1}{\mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2}\begin{bmatrix}\mathbf{E}\left[X^2\right] & -\mathbf{E}[X] \\ -\mathbf{E}[X] & 1\end{bmatrix} = \begin{bmatrix}\frac{\mathbf{E}\left[X^2\right]}{\text{Var}(X)} & -\frac{\mathbf{E}[X]}{\text{Var}(X)} \\ -\frac{\mathbf{E}[X]}{\text{Var}(X)} & \frac{1}{\text{Var}(X)}\end{bmatrix}$$

□

F Multiple Populations With Equal Variance

Let $\mathbf{pop}_1(\mu_1, \sigma^2) \dots \mathbf{pop}_K(\mu_K, \sigma^2)$ denote distributions with means $\mu_1 \dots \mu_K \in \mathbb{R}$ and equal variance $\sigma^2 \in \mathbb{R}$. For each $k = 1 \dots K$, assuming $N_k \geq 2$ define[14]

$$Y_{k,1} \dots Y_{k,N_k} \overset{\text{iid}}{\sim} \mathbf{pop}_k(\mu_k, \sigma^2) \qquad (N_k \text{ iid samples}) \tag{36}$$

$$\bar{Y}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} Y_{k,i} \qquad \text{(sample mean of the } k\text{-th population)}$$

$$\bar{S}_k^2 = \frac{1}{N_k - 1}\sum_{i=1}^{N_k}(Y_{k,i} - \bar{Y}_k)^2 \qquad \text{(sample variance of the } k\text{-th population)}$$

The mean and variance of each $\bar{Y}_k$ are given by

$$\mathbf{E}\left[\bar{Y}_k\right] = \frac{1}{N_k}\sum_{i=1}^{N_k}\mu_k = \mu_k$$

$$\text{Var}\left(\bar{Y}_k\right) = \text{Var}\left(\frac{1}{N_k}\sum_{i=1}^{N_k} Y_{k,i}\right) = \frac{1}{N_k^2}\text{Var}\left(\sum_{i=1}^{N_k} Y_{k,i}\right) = \frac{1}{N_k^2}\sum_{i=1}^{N_k}\text{Var}\left(Y_{k,i}\right) = \frac{\sigma^2}{N_k}$$

[13] In fact, $\hat{r}_{XY,N}$ is precisely the slope of the line through the origin that relates whitened X to whitened predictions by the LSE. For any $x \in \mathbb{R}$, let $\hat{y} = \hat{\alpha}_N + \hat{\beta}_N x$ denote the LSE prediction. Then

$$\frac{\hat{y} - \bar{y}_N}{\hat{\sigma}_{Y,N}} = \hat{r}_{XY,N} \times \frac{x - \bar{x}_N}{\hat{\sigma}_{X,N}}$$

[14] We name our random variables as Y's here for a connection to regression later.

F.1 Pooled Variance Estimator

The **pooled variance** is an unbiased estimator of σ^2 defined as

$$\bar{S}^2_{\text{pooled}} = \frac{\sum_{k=1}^K (N_k - 1)\bar{S}_k^2}{N - K} = \frac{\sum_{k=1}^K \sum_{i=1}^{N_k} (Y_{k,i} - \bar{Y}_k)^2}{N - K} \tag{37}$$

where $N = \sum_{k=1}^K N_k$. Check that $\bar{S}^2_{\text{pooled}} = (1/K)\sum_{k=1}^K \bar{S}_k^2$ is just the average if the sample sizes are the same.

Lemma F.1. $\bar{S}^2_{\text{pooled}}$ an (1) unbiased and (2) minimum-variance estimator of σ^2.

F.2 Total, Within-Group, and Between-Group Sum of Squares

We may define the **grand mean** of all $N := \sum_{k=1}^K N_k$ samples:

$$\bar{Y} = \frac{1}{N}\sum_{k=1}^K \sum_{i=1}^{N_k} Y_{k,i} = \sum_{k=1}^K \frac{N_k}{N}\bar{Y}_k$$

The mean and variance of the grand mean are given by

$$\begin{aligned}
\mathbf{E}\left[\bar{Y}\right] &= \sum_{k=1}^K \frac{N_k}{N}\mu_k =: \mu \\
\text{Var}\left(\bar{Y}\right) &= \text{Var}\left(\sum_{k=1}^K \frac{N_k}{N}\bar{Y}_k\right) = \sum_{k=1}^K \frac{N_k^2}{N^2}\frac{\sigma^2}{N_k} = \sum_{k=1}^K \frac{N_k\sigma^2}{N^2} = \frac{\sigma^2}{N}
\end{aligned}$$

Lemma F.2.

$$\underbrace{\sum_{k=1}^K \sum_{i=1}^{N_k} (Y_{k,i} - \bar{Y})^2}_{Q_{\text{total}}} = \underbrace{\sum_{k=1}^K \sum_{i=1}^{N_k} (Y_{k,i} - \bar{Y}_k)^2}_{Q_{\text{within}}} + \underbrace{\sum_{k=1}^K N_k(\bar{Y}_k - \bar{Y})^2}_{Q_{\text{between}}}$$

Lemma F.3.

$$\mathbf{E}\left[Q_{\text{total}}\right] = (N-1)\sigma^2 + \sum_{k=1}^K (\mu_k - \mu)^2 \tag{38}$$

$$\mathbf{E}\left[Q_{\text{within}}\right] = (N-K)\sigma^2 \tag{39}$$

$$\mathbf{E}\left[Q_{\text{between}}\right] = (K-1)\sigma^2 + \sum_{k=1}^K (\mu_k - \mu)^2 \tag{40}$$

This yields an interesting variance estimator that we call **between-group variance**, defined as

$$\bar{S}^2_{\text{between}} = \frac{Q_{\text{between}}}{K-1} = \frac{1}{K-1}\sum_{k=1}^K N_k(\bar{Y}_k - \bar{Y})^2 \tag{41}$$

Note that $\mathbf{E}\left[\bar{S}^2_{\text{between}}\right] = \sigma^2 + \frac{1}{K-1}\sum_{k=1}^K (\mu_k - \mu)^2 \geq \sigma^2$ by (40), thus $\bar{S}^2_{\text{between}}$ is an unbiased estimator of σ^2 only if $\mu_1 = \cdots = \mu_K$.

F.3 Emergence of the F-Statistic

The current setting, multiple populations $\mathbf{pop}_k(\mu_k, \sigma^2)$ with equal variance, induces random variables suitable for hypothesis testing under additional assumptions.

Assumption F.1 (Normality)**.** $\mathbf{pop}_k(\mu_k, \sigma^2) = \mathcal{N}(\mu_k, \sigma^2)$

Assumption F.2 (Equal mean)**.** $\mu_1 = \cdots = \mu_K = \mu$

Lemma F.4. Under Assumption F.1

$$\frac{N-K}{\sigma^2}\bar{S}^2_{\text{pooled}} = \frac{1}{\sigma^2}\sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i} - \bar{Y}_k)^2 \sim \chi^2(N-K)$$

Proof. Under normality, this is a sum of K independent $W_k = \frac{1}{\sigma^2}\sum_{i=1}^{N_k}(Y_{k,i} - \bar{Y}_k)^2 \sim \chi^2(N_k - 1)$ (45). By the additivity of chi-square (43), $\sum_{k=1}^{K} W_k \sim \chi^2(N-K)$. □

Lemma F.5. Under Assumptions F.1 and F.2,

$$\frac{K-1}{\sigma^2}\bar{S}^2_{\text{between}} = \frac{1}{\sigma^2}\sum_{k=1}^{K} N_k(\bar{Y}_k - \bar{Y})^2 \sim \chi^2(K-1)$$

Proof. From Lemma F.2

$$\frac{1}{\sigma^2}\sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i} - \bar{Y})^2 = \frac{1}{\sigma^2}\sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i} - \bar{Y}_k)^2 + \frac{1}{\sigma^2}\sum_{k=1}^{K} N_k(\bar{Y}_k - \bar{Y})^2$$

Since the means are the same, the LHS is distributed as $\chi^2(N-1)$. By Lemma F.4 the first term on the RHS is distributed as $\chi^2(N-K)$. Finally, we claim that the two terms on the RHS are independent as follows. The first term consists of the (scaled) sample variance $\bar{S}^2_k$ of each group $k = 1 \dots K$. $\bar{S}^2_k$ is independent of $\bar{Y}_l$ for $l \neq k$ by premise, and also of $\bar{Y}_k$ under normality (Theorem I.1). It follows that $\bar{S}^2_k$ is independent of $\bar{Y}$ which is a linear combination of $\bar{Y}_1 \dots \bar{Y}_K$. By the subtractivity of the chi-square distribution (44), we conclude that the second term on the RHS is distributed as $\chi^2(K-1)$. □

Corollary F.6. Under Assumptions F.1 and F.2,

$$\frac{\bar{S}^2_{\text{between}}}{\bar{S}^2_{\text{pooled}}} \sim F(K-1, N-K)$$

Proof.

$$\frac{\bar{S}^2_{\text{between}}}{\bar{S}^2_{\text{pooled}}} = \frac{\frac{K-1}{\sigma^2}\bar{S}^2_{\text{between}}/(K-1)}{\frac{N-K}{\sigma^2}\bar{S}^2_{\text{pooled}}/(N-K)}$$

By Lemma F.5 and F.4, we have $\frac{K-1}{\sigma^2}\bar{S}^2_{\text{between}} \sim \chi^2(K-1)$ and $\frac{N-K}{\sigma^2}\bar{S}^2_{\text{pooled}} \sim \chi^2(N-K)$. The proof of Lemma F.5 shows that $\bar{S}^2_k$ is independent of $\bar{Y}_l$ for all l and $\bar{Y}$, thus independent of $\bar{S}^2_{\text{between}} = (1/(K-1))\sum_{l=1}^{l} N_l(\bar{Y}_l - \bar{Y})^2$. Since this holds for all k, $\bar{S}^2_{\text{pooled}} = (1/(N-K))\sum_{k=1}^{K}(N_k - 1)\bar{S}^2_k$ is independent of $\bar{S}^2_{\text{between}}$. The statement now follows from the definition of the F-statistic (48). □

F.3.1 As Regression

Let $\mathbf{pop}_{\{1\dots K\}}$ denote any full-support distribution over "levels" $\{1 \dots K\}$ and $A \sim \mathbf{pop}_{\{1\dots K\}}$ be a "factor". Let

$$Y = \alpha + \beta^\top C^d_A(A) + Z_{\sigma^2} \qquad\qquad Z_{\sigma^2} \sim \mathcal{N}(0, \sigma^2)$$

where $C^d_A(A) \in \{0,1\}^{K-1}$ denotes the dummy coding of $A \in \{1 \dots K\}$ with $A = 1$ as the reference level (Appendix G). Let $Y_1 \dots Y_N$ denote iid samples of Y associated with $A_1 \dots A_N \sim \mathbf{pop}_{\{1\dots K\}}$. We may equivalently express these samples as $Y_{k,i}$ which indicates the i-th sample in group k. Then

$$\begin{aligned} Y_{1,1} \dots Y_{1,N_1} &\overset{\text{iid}}{\sim} \mathcal{N}(\alpha, \sigma^2) \\ Y_{k,1} \dots Y_{k,N_k} &\overset{\text{iid}}{\sim} \mathcal{N}(\alpha + \beta_k, \sigma^2) \qquad\qquad \forall k \in \{2 \dots K\} \end{aligned}$$

where we use the indexing $\beta = (\beta_2 \dots \beta_K)$ for convenience. Thus the generative story coincides with multiple populations with equal variance in (36) under normality (Assumption F.1). Furthermore, the equal-mean assumption (Assumption F.2) corresponds to assuming $\beta_k = 0$ for all $k = 2 \dots K$ (i.e., Y is independent of A). Then we may apply Theorem D.15 with $P = \{1\}$ and have that

$$\frac{\left|\left|\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\right|\right|^2 / (K-1)}{\left|\left|\boldsymbol{y} - \hat{\boldsymbol{y}}\right|\right|^2 / (N-K)} \sim F(K-1, N-K)$$

where $\boldsymbol{y}_j = Y_j$ and $\hat{\boldsymbol{y}}, \bar{\boldsymbol{y}} \in \mathbb{R}^N$ are least-squares predictions using $[1] \oplus C_A^d(A) \in \mathbb{R}^K$ and $[1] \in \mathbb{R}$ as independent variables. This means $\hat{\boldsymbol{y}}_j = \bar{Y}_{A_j}$ by the property of dummy coding (Lemma G.1) and $\bar{\boldsymbol{y}}_j = \bar{Y}$. So the above term can be written as

$$\frac{\left|\left|\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\right|\right|^2 / (K-1)}{\left|\left|\boldsymbol{y} - \hat{\boldsymbol{y}}\right|\right|^2 / (N-K)} = \frac{\sum_{j=1}^N (\hat{\boldsymbol{y}}_j - \bar{\boldsymbol{y}}_j)^2 / (K-1)}{\sum_{j=1}^N (\boldsymbol{y}_j - \hat{\boldsymbol{y}}_j)^2 / (N-K)} = \frac{\sum_{k=1}^K \sum_{i=1}^{N_k} (\bar{Y}_k - \bar{Y})^2 / (K-1)}{\sum_{k=1}^K \sum_{i=1}^{N_k} (Y_{i,k} - \bar{Y}_k)^2 / (N-K)} = \frac{\bar{S}^2_{\text{between}}}{\bar{S}^2_{\text{pooled}}}$$

In summary,

- Ratio of between-group and pooled variance in N samples from K populations over $Y \in \mathbb{R}$ that are: (1) equal-variance, (2) normal, and (3) equal-mean
- Ratio of ESS and RSS (scaled by degrees of freedom) when regressing to $Y \in \mathbb{R}$ from the dummy coding of $A \in \{1 \dots K\}$ with N samples where A and Y are independent

are exactly the same F-statistic distributed as $F(K-1, N-K)$.

G Regression with Discrete Input

G.1 Statistics Jargon

In regression, each discrete input variable is called a **factor** and the different values it takes are called **levels**. The combinations of levels of different factors are called **treatments**. For example, we may regress to $Y \in \mathbb{R}$ using two factors $A \in \{1, 2\}$ (2 levels) and $B \in \{1, 2, 3\}$ (3 levels), associated with 6 treatments. If the number of samples is the same for each treatment, the data is called **balanced**. Otherwise, the data is called **unbalanced**. Below, the first table is balanced data since each of 6 treatments has 2 samples. The second table is unbalanced.

(balanced data)

A	B	Y
1	1	3.2
1	1	3.5
1	2	5.7
1	2	4.9
1	3	-1.7
1	3	-0.1
2	1	10.7
2	1	10.8
2	2	1.3
2	2	1.3
2	3	4.4
2	3	6.1

(unbalanced data)

A	B	Y
1	1	3.2
1	2	5.6
1	2	5.7
1	3	-1.7
2	1	10.9
2	1	10.7
2	1	10.8
2	2	1.3
2	2	5.1
2	3	4.4

A desirable property of balanced data is the lack of any linear association between factors (called **collinearity**). Knowing the level of one factor doesn't tell us anything about the level of another factor because all levels are still equally likely. Intuitively, zero collinearity makes it easier to assess the impact of one factor on Y while controlling for the others.

G.2 Coding Schemes

A naive way to encode $A \in \{1 \dots K\}$ is to use the one-hot encoding $\textbf{one-hot}(A) \in \{0, 1\}^K$ where $\textbf{one-hot}_k(A) = 1$ iff $A = k$. This has an unfortunate consequence that the data matrix $\boldsymbol{X} \in \mathbb{R}^{N \times (K+1)}$ where the first column

is $\mathbf{1}_N$ (bias dimension) is never full-rank. This is because the encoding implies the constraint $\textbf{one-hot}_k(A) = 1 - \sum_{k' \neq k} \textbf{one-hot}_{k'}(A)$, so that the sum of columns $2 \dots K+1$ always equals $\mathbf{1}_N$ (the first column). For instance, if we have $N = 4$ samples of $A \in \{1, 2\}$ taking values $1, 1, 2, 2$, the data matrix is given by

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

which is not full-rank even though all values of A have been observed. While we don't need the full-rank condition for least-squares linear regression, it's inconvenient for analysis and can be easily fixed.

G.2.1 Dummy coding

A simplest fix is to designate one level as a **reference level** and eliminate it in the one-hot encoding. This still indicates when the factor takes the reference level because only that level has all zeros in non-bias dimensions. This is known as **dummy coding**. WLOG we always assume $k = 1$ is the reference level and write $C_A^d(a) \in \mathbb{R}^{K-1}$ to indicate the dummy coding of $A = a \in \{1 \dots K\}$. The dummy coding of $A \in \{1, 2\}$ and $B \in \{1, 2, 3\}$ is

$$\begin{aligned} C_A^d(1) &= 0 & C_B^d(1) &= (0, 0) \\ C_A^d(2) &= 1 & C_B^d(2) &= (1, 0) \\ & & C_B^d(3) &= (0, 1) \end{aligned}$$

Note that the data matrix associated with the example above (samples $1, 1, 2, 2$ of $A \in \{1, 2\}$) under dummy coding is now full-rank:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Lemma G.1. Let $(a_1, y_1) \dots (a_N, y_N)$ denote N samples of $(A, Y) \in \{1 \dots K\} \times \mathbb{R}$. Denote the least-squares regressor under dummy coding $C_A^d(a) \in \mathbb{R}^{K-1}$ by

$$(\alpha^*, \beta^*) = \underset{\alpha \in \mathbb{R},\ \beta \in \mathbb{R}^{K-1}}{\arg\min} \sum_{i=1}^{N} \left(y_i - \alpha - \beta^\top C_A^d(a_i)\right)^2$$

Let $\bar{y}_k = (1/\textbf{count}(k)) \sum_{i:a_i=k} y_i$ denote the mean response in $A = k$. Using the indexing $\beta^* = (\beta_2^* \dots \beta_K^*)$, we have

$$\begin{aligned} \alpha^* &= \bar{y}_1 \\ \beta_k^* &= \bar{y}_k - \bar{y}_1 & \forall k \in \{2 \dots K\} \end{aligned}$$

Lemma G.1 shows that for $k \geq 2$, the LSE parameter β_k^* is positive iff $\bar{y}_k > \bar{y}_1$. We can add additional factors under dummy coding and similar properties hold, as shown in the following example with 2 factors and their interaction.

Lemma G.2. Let $(a_1, b_1, y_1) \dots (a_N, b_N, y_N)$ denote N samples of $(A, B, Y) \in \{1 \dots K\} \times \{1 \dots L\} \times \mathbb{R}$. Denote the least-squares regressor under dummy coding $C_A^d(a) \in \mathbb{R}^{K-1}$ and $C_B^d(b) \in \mathbb{R}^{L-1}$ by

$$(\alpha^*, \beta^*, \gamma^*, \kappa^*) = \underset{\alpha \in \mathbb{R},\ \beta \in \mathbb{R}^{K-1},\ \gamma \in \mathbb{R}^{L-1},\ \kappa \in \mathbb{R}^{(K-1)(L-1)}}{\arg\min} \sum_{i=1}^{N} \left(y_i - \alpha - \beta^\top C_A^d(a_i) - \gamma^\top C_B^d(b_i) - \kappa^\top C_A^d(a_i) \otimes C_B^d(b_i)\right)^2$$

where $u \otimes v \in \mathbb{R}^{dd'}$ denotes the kronecker product of $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^{d'}$. Let $\bar{y}_{k,l} = (1/\textbf{count}(k, l)) \sum_{i:a_i=k, b_i=l} y_i$ denote the mean response in $(A = k) \wedge (B = l)$. Using the indexing $\beta^* = (\beta_2^* \dots \beta_K^*)$, $\gamma^* = (\gamma_2^* \dots \gamma_L^*)$, and

$\kappa^* = (\kappa^*_{2,2} \dots \kappa^*_{K,L})$, we have

$$\begin{aligned}
\alpha^* &= \bar{y}_{1,1} && \\
\beta^*_k &= \bar{y}_{k,1} - \bar{y}_{1,1} && \forall k \in \{2 \dots K\} \\
\gamma^*_l &= \bar{y}_{1,l} - \bar{y}_{1,1} && \forall l \in \{2 \dots L\} \\
\kappa^*_{k,l} &= \bar{y}_{k,l} - (\bar{y}_{k,1} + \bar{y}_{1,l} - \bar{y}_{1,1}) && \forall k \in \{2 \dots K\},\ l \in \{2 \dots L\}
\end{aligned}$$

G.2.2 Sum coding

The LSE weights under dummy coding change depending on the choice of a specific reference level. In **sum coding**, we represent the reference level as a vector of $K-1$ negative ones instead of zeros. WLOG we always assume $k = K$ (which is consistent with the implementation in R) is the reference level and write $C^s_A(a) \in \mathbb{R}^{K-1}$ to indicate the sum coding of $A \in \{1 \dots K\}$. The sum coding of $A \in \{1, 2\}$ and $B \in \{1, 2, 3\}$ is

$$\begin{aligned}
C^s_A(1) &= 1 \\
C^s_A(2) &= -1
\end{aligned}
\qquad\qquad
\begin{aligned}
C^s_B(1) &= (1, 0) \\
C^s_B(2) &= (0, 1) \\
C^s_B(3) &= (-1, -1)
\end{aligned}$$

This has the effect that a balanced data matrix is automatically centered. For example, the data matrix associated with samples $1, 1, 2, 2, 3, 3$ of $B \in \{1, 2, 3\}$ under sum coding is given by

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix}$$

which is full-rank and centered.

Lemma G.3. Let $(a_1, y_1) \dots (a_N, y_N)$ denote N samples of $(A, Y) \in \{1 \dots K\} \times \mathbb{R}$. Assume that the data is balanced so that each level has M samples (so $N = MK$). Denote the least-squares regressor under sum coding $C^s_A(a) \in \mathbb{R}^{K-1}$ by

$$(\alpha^*, \beta^*) = \underset{\alpha \in \mathbb{R},\ \beta \in \mathbb{R}^{K-1}}{\arg\min} \sum_{i=1}^{N} \left(y_i - \alpha - \beta^\top C^s_A(a_i)\right)^2$$

Let $\bar{y} = (1/N) \sum_{i=1}^N y_i$ denote the grand mean, and $\bar{y}_k = (1/M) \sum_{i:a_i=k} y_i$ denote the mean when $A = k$. Then

$$\begin{aligned}
\alpha^* &= \bar{y} && \\
\beta^*_k &= \bar{y}_k - \bar{y} && \forall k \in \{1 \dots K-1\}
\end{aligned}$$

Lemma G.3 shows that for $k = 1 \dots K-1$, the LSE parameter β^*_k is positive iff $\bar{y}_k > \bar{y}$ (assuming balanced data). Unlike in dummy coding, the parameters are not affected by the choice of reference level since we always compare the level mean with the grand mean (although we don't have a parameter associated with the reference level). We can add additional factors under sum coding and similar properties hold. We give the following lemma and omit the proof.

Lemma G.4. Let $(a_1, b_1, y_1) \dots (a_N, b_N, y_N)$ denote N samples of $(A, B, Y) \in \{1 \dots K\} \times \{1 \dots L\} \times \mathbb{R}$. Assume that the data is balanced so that each treatment has M samples (so $N = MKL$). Denote the least-squares regressor under sum coding $C^s_A(a) \in \mathbb{R}^{K-1}$ and $C^d_B(b) \in \mathbb{R}^{L-1}$ by

$$(\alpha^*, \beta^*, \gamma^*, \kappa^*) = \underset{\alpha \in \mathbb{R},\ \beta \in \mathbb{R}^{K-1},\ \gamma \in \mathbb{R}^{L-1},\ \kappa \in \mathbb{R}^{(K-1)(L-1)}}{\arg\min} \sum_{i=1}^{N} \left(y_i - \alpha - \beta^\top C^s_A(a_i) - \gamma^\top C^s_B(b_i) - \kappa^\top C^s_A(a_i) \otimes C^s_B(b_i)\right)^2$$

where $u \otimes v \in \mathbb{R}^{dd'}$ denotes the kronecker product of $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^{d'}$. Let $\bar{y} = (1/N)\sum_{i=1}^N y_i$ denote the grand mean, $\bar{y}_{k,\cdot} = (1/ML)\sum_{i:a_i=k} y_i$ denote the mean when $A = k$, $\bar{y}_{\cdot,l} = (1/MK)\sum_{i:b_i=l} y_i$ denote the mean when $B = l$, and $\bar{y}_{k,l} = (1/M)\sum_{i:a_i=k,b_i=l} y_i$ denote the mean when $(A = k) \wedge (B = l)$. Then

$$\begin{aligned}
\alpha^* &= \bar{y} && \\
\beta_k^* &= \bar{y}_{k,\cdot} - \bar{y} && \forall k \in \{1 \dots K-1\} \\
\gamma_l^* &= \bar{y}_{\cdot,l} - \bar{y} && \forall l \in \{1 \dots L-1\} \\
\kappa_{k,l}^* &= \bar{y}_{k,l} - (\bar{y}_{k,\cdot} + \bar{y}_{\cdot,l} - \bar{y}) && \forall k \in \{1 \dots K-1\},\ l \in \{1 \dots L-1\}
\end{aligned}$$

H Satellite Distributions Derived for Normal Variables

These are distributions that emerge when estimating certain quantities under normal distributions. Specifically,

- Sample variance, scaled by sample size and true variance, follows a chi-square distribution (45).
- Whitened average using sample variance is distributed as a t-distribution (47).
- Ratio of sample variances, scaled by true variances, is distributed as an F-distribution (49).

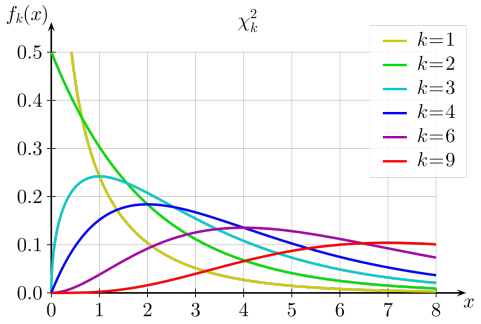
They have complicated PDFs and CDFs, but we don't care as long as we can calculate probabilities under these distributions (e.g., see this calculator). Image credit: Wikipedia.

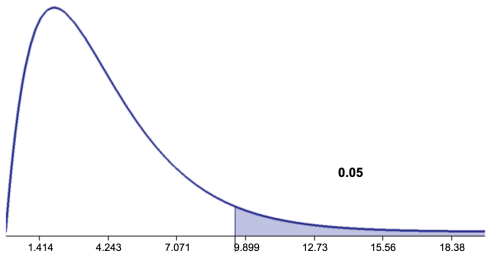
H.1 Chi-Square Distribution

The **chi-square distribution with k degrees of freedom**, denoted by $\chi^2(k)$, is the distribution of $Q_k \in \mathbb{R}_{\geq 0}$ defined as

$$Q_k = \sum_{i=1}^k Z_i^2 \qquad\qquad Z_1 \dots Z_k \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \tag{42}$$

Equivalently $Q_k = ||Z||^2$ where $Z \sim \mathcal{N}(0_k, I_{k\times k})$. There are many chi-square distributions (one for each $k \in \mathbb{N}$). They are assymetric and look like these.





The right figure shows the right-tail critical value for $\chi^2(4)$ at significance level 0.05 (9.488).

Additivity and subtractivity.

$$Q_{k_1} \sim \chi^2(k_1),\ Q_{k_2} \sim \chi^2(k_2) \quad \Rightarrow \quad Q_{k_1} + Q_{k_2} \sim \chi^2(k_1 + k_2) \tag{43}$$

$$Z = X + Y,\ X \perp Y,\ Z \sim \chi^2(k_1 + k_2),\ X \sim \chi^2(k_1) \quad \Rightarrow \quad Y \sim \chi^2(k_2) \tag{44}$$

The additivity follows immediatly from the definition (42). The subtractivity can be verified by deriving the MGF of Y from the MGF of Z, which is the product of the MGFs of X and Y (since $X \perp Y$), the MGF of X (see p. 35 of here).

For general normal distributions. Let $X_1 \dots X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. From (43) we have $\frac{1}{\sigma^2}\sum_{i=1}^N (X_i - \mu)^2 \sim \chi^2(N)$, but this requires the knowledge of μ. A more useful characterization is the following. Let $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$ denote the sample mean and assume $N \geq 2$. Then

$$\frac{1}{\sigma^2}\sum_{i=1}^N \left(X_i - \bar{X}_N\right)^2 \sim \chi^2(N-1) \tag{45}$$

H.1.1 Noncentral Chi-Square Distribution

Let $Z \sim \mathcal{N}(\mu, I_{k\times k})$ for some $\mu \in \mathbb{R}^k$. Then $||Z||^2$ distributed as $\chi^2(k, ||\mu||^2)$ which is called the **noncentral chi-square distribution with k degrees of freedom and the noncentrality parameter $||\mu||^2$**. If $\mu = 0_k$ then $\chi^2(k, 0) = \chi^2(k)$. A result we use is the following.

Lemma H.1. Let $y \sim \mathcal{N}(\mu, \Sigma)$ for some $\mu \in \mathbb{R}^N$ and positive definite $\Sigma \in \mathbb{R}^{N\times N}$. If $A \in \mathbb{R}^{N\times N}$ is symmetric with rank k, and if $A\Sigma$ is idempotent (i.e., $(A\Sigma)(A\Sigma) = A\Sigma$), then $y^\top Ay \sim \chi^2(k, \mu^\top A\mu)$.

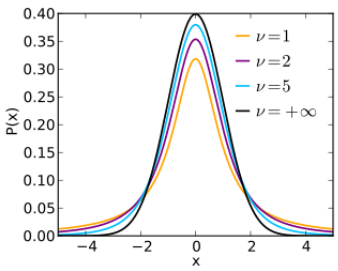
While we omit the proof (see this), we make a quick note that we can get rid of Σ in the lemma since $(A\Sigma)(A\Sigma) = A\Sigma$ and $\Sigma \succ 0$ together imply $A\Sigma A = A$. The rank of A becomes the degree of freedom of the noncentral chi-square variable $y^\top Ay$.

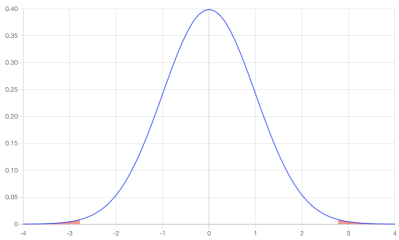
H.2 T-Distribution

The **t-distribution with $\nu \in \mathbb{N}$ degrees of freedom**, denoted by $\tau(\nu)$, is the distribution of $T_\nu \in \mathbb{R}$ defined as

$$T_\nu = Z\sqrt{\frac{\nu}{V}} \qquad\qquad Z \sim \mathcal{N}(0,1) \perp V \sim \chi^2(\nu) \tag{46}$$

t-distributions are symmetric and look like standard normal with fatter tails (in fact $T_\nu \xrightarrow{d} Z$).





The right figure shows the two-tail critical values for $\tau(4)$ at significance level 0.05 (-2.776 and 2.776). Note that the critical region is smaller than $\mathcal{N}(0,1)$ at the same significance level (-1.96 and 1.96). Intuitively, this means it is harder to reject the null hypothesis (of zero mean) due to uncertainty of a small sample size.

Lemma H.2. Let $X_1 \dots X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $N \geq 2$. Let $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and $\bar{S}_N^2 = \frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X}_N)^2$ denote unbiased estimators of the mean and variance. Then

$$\frac{\bar{X}_N - \mu}{\bar{S}_N/\sqrt{N}} \sim \tau(N-1) \tag{47}$$

Proof. Define $\bar{Z}_N = \frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)$ using the knowledge of σ^2. We may rewrite

$$\frac{\bar{X}_N - \mu}{\bar{S}_N/\sqrt{N}} = \bar{Z}_N\sqrt{\frac{\sigma^2}{\bar{S}_N^2}} = \bar{Z}_N\sqrt{\frac{N-1}{(N-1)\bar{S}_N^2/\sigma^2}}$$

It remains to show that $(N-1)\bar{S}_N^2/\sigma^2 \sim \chi^2(N-1)$ and it is independent of $\bar{Z}_N$. The first follows from the definition of $\bar{S}_N^2$ and (45). The second follows from the fact that $\bar{X}_N$ and $\bar{S}_N^2$ are independent under normality (Theorem I.1). □

From Lemma H.2, we have

$$\frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1) \qquad \text{v.s.} \qquad \frac{\bar{X}_N - \mu}{\bar{S}_N/\sqrt{N}} \sim \tau(N-1)$$

Given this fact, the shape of the t-distribution makes sense. It is "standard normal" but has fatter tails to account for the uncertainty of the sample variance when the sample size N is finite.

We give another useful lemma.

Lemma H.3. Let $X_1 \dots X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $Y_1 \dots Y_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $N \geq 2$. Let $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and $\bar{S}_X^2 = \frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X}_N)^2$. Let $\bar{Y}_N = \frac{1}{N}\sum_{i=1}^N Y_i$ and $\bar{S}_Y^2 = \frac{1}{N-1}\sum_{i=1}^N (Y_i - \bar{Y}_N)^2$. Let $\bar{S}^2_{\text{pooled}} = (\bar{S}_X^2 + \bar{S}_Y^2)/2$ denote the pooled estimator of σ^2 (37). Then

$$\frac{\bar{X}_N - \bar{Y}_N}{\bar{S}_{\text{pooled}}\sqrt{\frac{2}{N}}} \sim \tau(2N-2)$$

Proof. We have $\bar{X}_N - \bar{Y}_N \sim \mathcal{N}\left(0, 2\sigma^2/N\right)$ by the independence of $\bar{X}_N$ and $\bar{Y}_N$. Define $\bar{Z}_N = \frac{\bar{X}_N - \bar{Y}_N}{\sigma\sqrt{2/N}} \sim \mathcal{N}(0,1)$ using the knowledge of σ^2. We may rewrite

$$\frac{\bar{X}_N - \bar{Y}_N}{\bar{S}_{\text{pooled}}\sqrt{\frac{2}{N}}} = \bar{Z}_N \sqrt{\frac{\sigma^2}{\bar{S}^2_{\text{pooled}}}} = \bar{Z}_N \sqrt{\frac{2N-2}{(2N-2)\bar{S}^2_{\text{pooled}}/\sigma^2}}$$

It remains to show that $(2N-2)\bar{S}^2_{\text{pooled}}/\sigma^2 \sim \chi^2(2N-2)$ and it is independent of $\bar{Z}_N$. As in the proof of Lemma H.2 $(N-1)\bar{S}_X^2/\sigma^2$ and $(N-1)\bar{S}_X^2/\sigma^2$ are distributed as $\chi^2(N-1)$ and are independent, thus by the additivity of the chi-square distribution (43)

$$\frac{(2N-2)\bar{S}^2_{\text{pooled}}}{\sigma^2} = \frac{(N-1)\bar{S}_X^2}{\sigma^2} + \frac{(N-1)\bar{S}_Y^2}{\sigma^2} \sim \chi^2(2N-2)$$

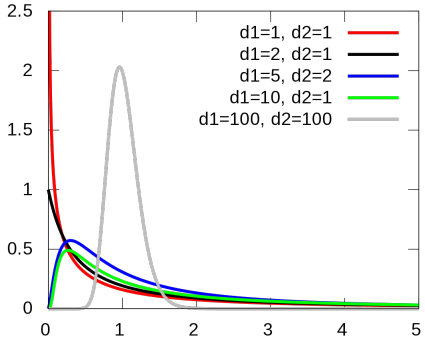
The second follows since $\bar{X}_N \perp \bar{S}_X^2$ and $\bar{Y}_N \perp \bar{S}_Y^2$ under normality (Theorem I.1) so that $\bar{X}_N - \bar{Y}_N \perp \bar{S}_X^2 + \bar{S}_Y^2$. □

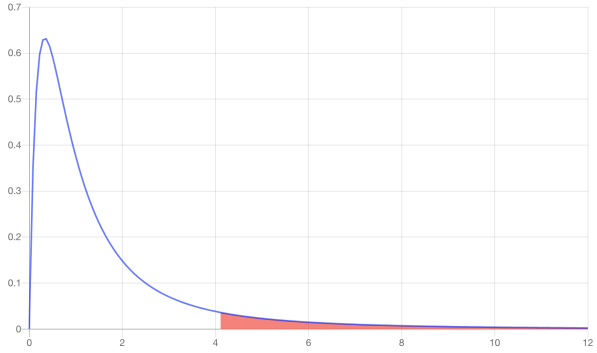
H.3 F-Distribution

The **F-distribution with $(d_1, d_2) \in \mathbb{N}^2$ degrees of freedom**, denoted by $F(d_1, d_2)$, is the distribution of $F \in \mathbb{R}_{\geq 0}$ defined as

$$F_{d_1,d_2} = \frac{U_1/d_1}{U_2/d_2} \qquad U_1 \sim \chi^2(d_1) \;\perp\; U_2 \sim \chi^2(d_2) \tag{48}$$

F-distributions are assymetric and look like these[15]





The right figure shows the upper-tail critical value for $F(4,4)$ at significance level 0.1 (4.107).

[15] The mean $\mathbf{E}[F_{d_1,d_2}] = d_2/(d_2-2)$ exists only if $d_2 > 2$ and is roughly 1 for large d_2.

Lemma H.4. Let $X_1 \dots X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1 \dots Y_M \overset{\text{iid}}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$ where $N, M \geq 2$. Let $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and $\bar{Y}_M = \frac{1}{M}\sum_{i=1}^M Y_i$ denote sample means. Let $\bar{S}^2_{X,N} = \frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X}_N)^2$ and $\bar{S}^2_{Y,M} = \frac{1}{M-1}\sum_{i=1}^M (Y_i - \bar{Y}_M)^2$ denote sample variances. Then

$$\frac{\bar{S}^2_{X,N}/\sigma_X^2}{\bar{S}^2_{Y,M}/\sigma_Y^2} \sim F(N-1, M-1) \tag{49}$$

Proof. Let $Q_{X,N} = \frac{1}{\sigma_X^2}\sum_{i=1}^N (X_i - \bar{X}_N)^2$ and $Q_{Y,M} = \frac{1}{\sigma_Y^2}\sum_{i=1}^M (Y_i - \bar{Y}_M)^2$. They are independent and distributed as $\chi^2(N-1)$ and $\chi^2(M-1)$ by (45). The claim then follows since

$$\frac{\bar{S}^2_{X,N}/\sigma_X^2}{\bar{S}^2_{Y,M}/\sigma_Y^2} = \frac{\frac{1}{\sigma_X^2}\sum_{i=1}^N (X_i - \bar{X}_N)^2/(N-1)}{\frac{1}{\sigma_Y^2}\sum_{i=1}^M (Y_i - \bar{Y}_M)^2/(M-1)} = \frac{Q_{X,N}/(N-1)}{Q_{Y,M}/(M-1)}$$

□

Corollary H.5. Let $X_1 \dots X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu_X, \sigma^2)$ and $Y_1 \dots Y_M \overset{\text{iid}}{\sim} \mathcal{N}(\mu_Y, \sigma^2)$ where $N, M \geq 2$. Let $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and $\bar{Y}_M = \frac{1}{M}\sum_{i=1}^M Y_i$ denote sample means. Let $\bar{S}^2_{X,N} = \frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X}_N)^2$ and $\bar{S}^2_{Y,M} = \frac{1}{M-1}\sum_{i=1}^M (Y_i - \bar{Y}_M)^2$ denote sample variances. Then

$$\frac{\bar{S}^2_{X,N}}{\bar{S}^2_{Y,M}} \sim F(N-1, M-1) \tag{50}$$

H.4 Studentized Range Distribution

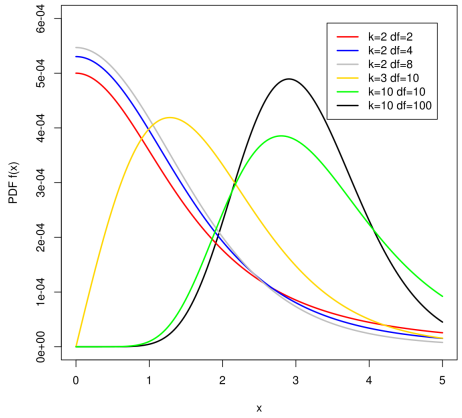
For $k = 1 \dots K$, let $X_{k,1} \dots X_{k,N_k} \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $\bar{X}_k = (1/N_k)\sum_{i=1}^{N_k} X_{k,i}$. The **studentized range distribution with K groups and $N-K$ degrees of freedom** denoted by **srange**$(K, N-K)$, is the distribution of $q_{K,N-K} > 0$ defined as

$$q_{K,N-K} = \frac{\max_{k=1}^K \bar{X}_k - \min_{k=1}^K \bar{X}_k}{\bar{S}_{\text{pooled}}\sqrt{\frac{2}{N}}}$$

where $\bar{S}^2_{\text{pooled}}$ is pooled variance (37). Note that for $K = 2$ and $N_1 = N_2 = N$ we can write

$$q_{2,2N-2} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\bar{S}_{\text{pooled}}\sqrt{\frac{2}{N}}}$$

which follows the t-distribution with $2N-2$ degrees of freedom (Lemma H.3). Thus the studentized range distribution is equivalent (up to normalization) to the upper half of the t-distribution with 2 groups of the same sample size. Studentized range distirbutions are assymetric and look like these (using a definition that differs by a factor of $\sqrt{2}$):



To get a sense of critical values, the right-tail critical value for **srange**$(3, 10)$ (the yellow one) at significance level 0.05 is 3.879.

I Facts About Independence

Theorem I.1. Let $X_1 \dots X_N \overset{\text{iid}}{\sim} \textbf{pop}$ be random vectors with sample mean $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and sample covariance matrix $\bar{S}_N^2 = \frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X}_N)(X_i - \bar{X}_N)^\top$. Then

$$\bar{X}_N \text{ and } \bar{S}_N^2 \text{ are independent} \quad \Leftrightarrow \quad \textbf{pop} \text{ is normal}$$

A proof can be found here. The fact that this property characterizes the normal distribution was first shown by Geary (1936).

Theorem I.2 (Craig's theorem)**.** Let $X \sim \mathcal{N}(\mu, \Sigma)$ be a normal vector in $\mathbb{R}^d$. For any $A, B \in \mathbb{R}^{d\times d}$,

$$A\Sigma B = 0_{d\times d} \quad \Leftrightarrow \quad X^\top AX \text{ and } X^\top BX \text{ are independent}$$

Theorem I.2 is attributed to Allen T. Craig. Despite its simple form, the theorem is difficult to prove and has a long and complicated history (Driscoll and Gundberg Jr, 1986).

Lemma I.3. Let $X \sim \mathcal{N}(\mu, \Sigma)$, $A \in \mathbb{R}^{n\times d}$, and $B \in \mathbb{R}^{m\times d}$. Then $AX \in \mathbb{R}^n$ and $BX \in \mathbb{R}^m$ are independent iff $A\Sigma B^\top = 0_{n\times m}$.

Lemma I.3 is a well-known property of the normal distribution exploiting the equivalence between uncorrelatedness and independence for jointly normal variables (see here).

J Projection Matrices

A square matrix $P \in \mathbb{R}^{N\times N}$, not necessarily symmetric, is called a **projection** (aka. **idempotent**) matrix if $P^2 = P$, because in that case $P^n x \in \text{range}\,(P)$ for all $n \in \mathbb{N}$. Properties of a projection P include:

- The eigenvalues must be either zero or one. If $Pu = \lambda u$ for some nonzero vector u, then $\lambda u = Pu = PPu = \lambda Pu = \lambda^2 u$ so $\lambda(\lambda - 1)u = 0_N$, meaning $\lambda \in \{0, 1\}$.
- $\text{tr}\,(P) = \text{rank}\,(P)$ (see here).
- It is diagonalizable. So it admits an eigendecomposition of the form $P = V\,\text{diag}\,(1 \dots 1, 0 \dots 0)\,V^{-1}$.

The second and third properties are nontrivial because P is generally nonsymmetric.[16] For instance, the following matrix is a projection onto $\text{span}\,(\{(1, 1)\})$ and has the eigendecomposition even though it is not symmetric (note that V is invertible but not orthonormal):

$$P = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} = V \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} V^{-1} \qquad V = \begin{bmatrix} 0.707 & 0 \\ 0.707 & 1 \end{bmatrix}$$

If P is any projection matrix, $R = I_{N\times N} - P$ is a projection matrix onto $\text{null}(P)$ since $R^2 = R$ (easy to check) and $\text{range}\,(R) = \text{null}(P)$.[17] Thus we can decompose any point $x \in \mathbb{R}^N$ into $\text{range}\,(P)$ and $\text{null}(P)$ by

$$x = \underbrace{Px}_{\text{proj. onto range}\,(P)} + \underbrace{(I_{N\times N} - P)x}_{\text{proj. onto null}(P)} \tag{51}$$

[16] Recall that any symmetric matrix is diagonalizable and its rank is equal to the number of nonzero eigenvalues.

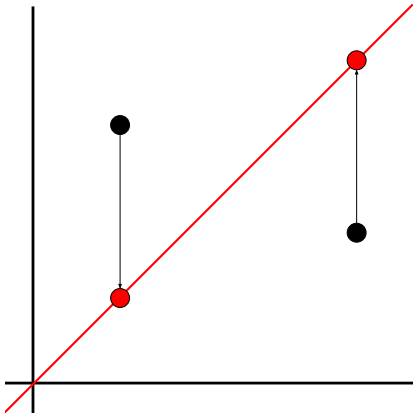
[17] If $x \in \text{range}\,(I_{N\times N} - P)$, then $x = y - Py$ for some $y \in \mathbb{R}^N$, so $Px = Py - PPy = 0$ and hence $x \in \text{null}(P)$. If $x \in \text{null}(P)$, then $Px = 0_N$ and thus $(I_{N\times N} - P)x = x - Px = x$ which shows that $x \in \text{range}\,(I_{N\times N} - P)$.

Now given a subspace $S \subset \mathbb{R}^N$, there are generally many possible projections onto S. For instance, two projections onto $S = \text{span}(\{(1,1)\}) \subset \mathbb{R}^2$ are (with corresponding null spaces):

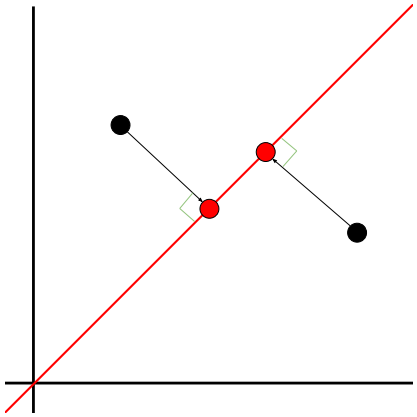
$$P_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \qquad P_2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

$$\text{null}(P_1) = \left\{ \begin{bmatrix} 0 \\ c \end{bmatrix} : \ c \in \mathbb{R} \right\} \qquad \text{null}(P_2) = \left\{ \begin{bmatrix} c \\ -c \end{bmatrix} : \ c \in \mathbb{R} \right\}$$

Note how the null spaces are different even though the ranges are the same. They are clearly very different transformations. If we project two points $(1,5)$ and $(5,3)$ onto S, we will have $(1,1)$ and $(5,5)$ if we use P_1 and $(3,3)$ and $(4,4)$ if we use P_2.



Projection onto S using P_1



Projection onto S using P_2

While both are valid projections and the range-null decomposition (51) holds in either case (exercise: check this), the second is more "efficient" in the sense that it's the *closest* projection. This property is formalized as follows.

Lemma J.1. Let $P \in \mathbb{R}^{N \times N}$ be a projection onto subspace $S = \text{range}(P) \subset \mathbb{R}^N$. The following statements are equivalent:

1. For any $x \in \mathbb{R}^N$, let $y = Px \in S$. Then $||x - y|| \leq ||x - z||$ for all $z \in S$ with equality iff $z = y$.
2. $\text{range}(P) \perp \text{null}(P)$[18]
3. $P = P^\top$

If any holds, we say P is an **orthogonal projection**.

Corollary J.2. For any subspace $S \subset \mathbb{R}^N$, there exists an orthogonal projection onto S and it is unique.

Proof. An orthogonal projection onto S is given by $P = UU^\top$ where $U \in \mathbb{R}^{N \times \dim(S)}$ is an orthonormal basis of S. If P' is any orthogonal projection onto S, $||x - Px|| = ||x - P'x||$ for all $x \in \mathbb{R}^N$ by 1. This implies $P = P'$. □

Construction from a matrix. For any nonzero matrix $A \in \mathbb{R}^{m \times n}$, the orthogonal projection onto $\text{range}(A) \subset \mathbb{R}^m$ is given by $P = AA^+ \in \mathbb{R}^{m \times m}$. Then $R = I_{m \times m} - P \in \mathbb{R}^{m \times m}$ is the orthogonal projection onto $\text{null}(P) = \text{range}(A)^\perp = \text{null}(A^\top) \subset \mathbb{R}^m$.

J.1 Subspace Decomposition Lemma

We state the lemma below without proof (see here).

Lemma J.3. Let $A_1 \dots A_K \in \mathbb{R}^{N \times N}$ denote square matrices such that $I_{N \times N} = \sum_{k=1}^K A_k$. The following statements are equivalent:

1. $A_1 \dots A_K$ are projections.

[18] Note that we are only able to compare the range and the null space because the matrix is square (otherwise the dimensions do not match), and even for square matrices the range and the null space are not necessarily orthogonal. See the figure using P_1 where the null space is the y-axis.

2. $A_k A_l = 0_{N\times N}$ for any $k \neq l$.

3. $\sum_{k=1}^{K} \operatorname{rank}(A_k) = N$

If any holds, we call $\sum_{k=1}^{K} A_k$ a **subspace decomposition** of $\mathbb{R}^N$.

The lemma generalizes the above discussion involving a single projection P. There we have $K = 2$ and consider $A_1 = P$ and $A_2 = I_{N\times N} - P$. Both are projections (1). We have $P(I_{N\times N} - P) = 0_{N\times N}$ (2). We also have $\operatorname{rank}(P) + \operatorname{rank}(I_{N\times N} - P) = \operatorname{rank}(P) + \operatorname{nullity}(P) = N$ by the rank-nullity theorem (3).

J.2 Chi-Square Quadratic Form

Lemma J.4. Let $X \sim \mathcal{N}(0_N, \sigma^2 I_{N\times N})$. For any orthogonal projection $P \in \mathbb{R}^{N\times N}$,

$$\frac{X^\top P X}{\sigma^2} \sim \chi^2(\operatorname{rank}(P))$$

Proof. P is a projection so it is diagonalizable with zero-one eigenvalues. P is furthermore symmetric so we can find orthonormal eigenvectors $V \in \mathbb{R}^{N\times N}$ such that $P = VJV^\top$ where $J = \operatorname{diag}(1\dots1, 0\dots0)$ has $\operatorname{rank}(P)$ 1's. Note that $Y := V^\top X \sim \mathcal{N}(0_{N\times N}, \sigma^2 I_{N\times N})$ by the orthonormality of V, and

$$\frac{X^\top P X}{\sigma^2} = \frac{Y^\top J Y}{\sigma^2} = \sum_{i=1}^{\operatorname{rank}(P)} \left(\frac{Y_i}{\sigma}\right)^2$$

is the sum of squares of $\operatorname{rank}(P)$ iid standard normal variables $Y_i/\sigma \sim \mathcal{N}(0,1)$, thus distributed as $\chi^2(\operatorname{rank}(P))$. □

K Proofs and Lemmas

Lemma K.1. For any $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} [(y - f^*(x))(f^*(x) - f(x))] = 0$$

Proof. For any $x \in \mathbb{R}^d$,

$$\begin{aligned}
\mathop{\mathbf{E}}_{y\sim \mathbf{pop}_{Y|X}(\cdot|x)} [(y - f^*(x))(f^*(x) - f(x))] &= \mathop{\mathbf{E}}_{y\sim \mathbf{pop}_{Y|X}(\cdot|x)} \left[y f^*(x) - y f(x) - f^*(x)^2 + f^*(x) f(x)\right] \\
&= \mathop{\mathbf{E}}_{y\sim \mathbf{pop}_{Y|X}(\cdot|x)} [y]\, f^*(x) - \mathop{\mathbf{E}}_{y\sim \mathbf{pop}_{Y|X}(\cdot|x)} [y]\, f(x) - f^*(x)^2 + f^*(x) f(x) \\
&= f^*(x)^2 - f^*(x) f(x) - f^*(x)^2 + f^*(x) f(x) \\
&= 0
\end{aligned}$$

Therefore

$$\mathop{\mathbf{E}}_{(x,y)\sim \mathbf{pop}_{XY}} [(y - f^*(x))(f^*(x) - f(x))] = \mathop{\mathbf{E}}_{x\sim \mathbf{pop}_X} \left[\mathop{\mathbf{E}}_{y\sim \mathbf{pop}_{Y|X}(\cdot|x)} [(y - f^*(x))(f^*(x) - f(x))] \right] = 0$$

□

Proof of Lemma D.8.

Proof. $A_{\boldsymbol{X}}$ is the orthogonal projection onto $\operatorname{null}(\boldsymbol{X}^\top) \subset \mathbb{R}^N$ and thus has rank $N - d$ by the rank-nullity theorem. $\boldsymbol{X}\boldsymbol{X}^+$ is the orthogonal projection onto $\operatorname{range}(\boldsymbol{X})$ while $\boldsymbol{X}_P \boldsymbol{X}_P^+$ is the orthogonal projection onto $\operatorname{range}(\boldsymbol{X}_P) \subset \operatorname{range}(\boldsymbol{X})$, thus $\boldsymbol{X}_P \boldsymbol{X}_P^+ = (\boldsymbol{X}\boldsymbol{X}^+)(\boldsymbol{X}_P \boldsymbol{X}_P^+) = (\boldsymbol{X}_P \boldsymbol{X}_P^+)(\boldsymbol{X}\boldsymbol{X}^+)$ and

$$\begin{aligned}
B^P_{\boldsymbol{X}} B^P_{\boldsymbol{X}} &= \left(\boldsymbol{X}\boldsymbol{X}^+ - \boldsymbol{X}_P \boldsymbol{X}_P^+\right)\left(\boldsymbol{X}\boldsymbol{X}^+ - \boldsymbol{X}_P \boldsymbol{X}_P^+\right) \\
&= \boldsymbol{X}\boldsymbol{X}^+ + \boldsymbol{X}_P \boldsymbol{X}_P^+ - (\boldsymbol{X}\boldsymbol{X}^+)(\boldsymbol{X}_P \boldsymbol{X}_P^+) - (\boldsymbol{X}_P \boldsymbol{X}_P^+)(\boldsymbol{X}\boldsymbol{X}^+) \\
&= \boldsymbol{X}\boldsymbol{X}^+ - \boldsymbol{X}_P \boldsymbol{X}_P^+ \\
&= B^P_{\boldsymbol{X}}
\end{aligned}$$

So $B_{\boldsymbol{X}}^P$ is a projection. Since it's a projection, its rank is given by the trace:

$$\operatorname{rank}\left(B_{\boldsymbol{X}}^P\right) = \operatorname{tr}\left(\boldsymbol{X}\boldsymbol{X}^+ - \boldsymbol{X}_P\boldsymbol{X}_P^+\right) = \operatorname{tr}\left(\boldsymbol{X}\boldsymbol{X}^+\right) - \operatorname{tr}\left(\boldsymbol{X}_P\boldsymbol{X}_P^+\right) = d - p$$

Finally, $B_{\boldsymbol{X}}^P$ is symmetric, so it's an orthogonal projection. Since $Q \subseteq P$, the same argument can be used to show that $C_{\boldsymbol{X}}^{P,Q} = \boldsymbol{X}_P\boldsymbol{X}_P^+ - \boldsymbol{X}_Q\boldsymbol{X}_Q^+$ is an orthogonal projection with rank $p - q$. Now we show the quadratic forms:

$$\begin{aligned}
||\hat{\boldsymbol{\epsilon}}||^2 &= \left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{X}^+\boldsymbol{y}\right|\right|^2 = ||A_{\boldsymbol{X}}\boldsymbol{y}||^2 = \boldsymbol{y}^\top A_{\boldsymbol{X}}\boldsymbol{y} \\
||\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_P||^2 &= \left|\left|\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{y} - \boldsymbol{X}_P\boldsymbol{X}_P^+\boldsymbol{y}\right|\right|^2 = \left|\left|B_{\boldsymbol{X}}^P\boldsymbol{y}\right|\right|^2 = \boldsymbol{y}^\top B_{\boldsymbol{X}}^P\boldsymbol{y} \\
||\boldsymbol{y} - \hat{\boldsymbol{y}}_P||^2 &= ||\hat{\boldsymbol{\epsilon}}||^2 + ||\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_P||^2 = \boldsymbol{y}^\top(A_{\boldsymbol{X}} + B_{\boldsymbol{X}}^P)\boldsymbol{y} = \boldsymbol{y}^\top(I_{N\times N} - \boldsymbol{X}_P\boldsymbol{X}_P^+)\boldsymbol{y} \\
\left|\left|\hat{\boldsymbol{y}}_P - \hat{\boldsymbol{y}}_Q\right|\right|^2 &= \left|\left|\boldsymbol{X}_P\boldsymbol{X}_P^+\boldsymbol{y} - \boldsymbol{X}_Q\boldsymbol{X}_Q^+\boldsymbol{y}\right|\right|^2 = \left|\left|C_{\boldsymbol{X}}^{P,Q}\boldsymbol{y}\right|\right|^2 = \boldsymbol{y}^\top C_{\boldsymbol{X}}^{P,Q}\boldsymbol{y}
\end{aligned} \tag{52}$$

where we use the fact that $A_{\boldsymbol{X}}$, $B_{\boldsymbol{X}}^P$, and $C_{\boldsymbol{X}}^{P,Q}$ are projections and (52) follows from the decomposition (27). □

Proof of Lemma E.1.

Proof. Let $J(\alpha,\beta) = \sum_{i=1}^N (y_i - \alpha - \beta x_i)^2$ which is strongly convex in either parameter. Setting the partial derivatives to zero yields two equations:

$$\begin{aligned}
\frac{\partial J(\alpha,\beta)}{\partial \alpha} &= -2\sum_{i=1}^N (y_i - \alpha - \beta x_i) = 0 \qquad \Leftrightarrow \qquad \alpha = \bar{y}_N - \beta\bar{x}_N \\
\frac{\partial J(\alpha,\beta)}{\partial \beta} &= -2\sum_{i=1}^N (y_i - \alpha - \beta x_i)x_i = 0 \qquad \Leftrightarrow \qquad \sum_{i=1}^N y_i x_i - \alpha x_i - \beta x_i^2 = 0
\end{aligned}$$

The expression for α is true no matter what β is. Plugging it in in the second equation we have

$$\sum_{i=1}^N y_i x_i - (\bar{y}_N - \beta\bar{x}_N)x_i - \beta x_i^2 = \sum_{i=1}^N y_i x_i - \bar{y}_N x_i + \beta\bar{x}_N x_i - \beta x_i^2 = 0 \qquad \Leftrightarrow \qquad \beta = \frac{\sum_{i=1}^N y_i x_i - \bar{y}_N x_i}{\sum_{i=1}^N x_i^2 - \bar{x}_N x_i}$$

The statement follows from the observation

$$\begin{aligned}
\sum_{i=1}^N (y_i - \bar{y}_N)(x_i - \bar{x}_N) &= \sum_{i=1}^N y_i x_i - y_i\bar{x}_N - \bar{y}_N x_i + \bar{y}_N\bar{x}_N = \sum_{i=1}^N y_i x_i - \bar{y}_N x_i \\
\sum_{i=1}^N (x_i - \bar{x}_N)^2 &= \sum_{i=1}^N x_i^2 - 2\bar{x}_N x_i + \bar{x}_N^2 = \sum_{i=1}^N x_i^2 - 2N\bar{x}_N^2 + N\bar{x}_N^2 = \sum_{i=1}^N x_i^2 - N\bar{x}_N^2 = \sum_{i=1}^N x_i^2 - \bar{x}_N x_i
\end{aligned}$$

□

Proof of Lemma D.17.

Proof. From (33),

$$\hat{w}_N = w_{\text{true}} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\epsilon} = w_{\text{true}} + \left(\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{\epsilon}$$

Here, $\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}$ converges to $\mathbf{E}\left[XX^\top\right]$ in probability by the law of large numbers. Since matrix inversion is continuous, $(\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ converges to $\mathbf{E}\left[XX^\top\right]^{-1}$ in probability by the continuous mapping theorem (13). As for $\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{\epsilon} = \frac{1}{N}\sum_{i=1}^N \epsilon_i x_i$, by the CLT as $N \to \infty$ we have

$$\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{\epsilon} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mathbf{E}\left[Z_{\sigma^2}X\right], \frac{1}{N}\operatorname{Cov}\left(Z_{\sigma^2}X, Z_{\sigma^2}X\right)\right)$$

where

$$\mathbf{E}\left[Z_{\sigma^2}X\right] = \mathbf{E}\left[Z_{\sigma^2}\right]\mathbf{E}\left[X\right] = 0_d$$
$$\mathrm{Cov}\left(Z_{\sigma^2}X, Z_{\sigma^2}X\right) = \mathbf{E}\left[Z_{\sigma^2}^2 XX^\top\right] = \mathbf{E}\left[Z_{\sigma^2}^2\right]\mathbf{E}\left[XX^\top\right] = \sigma^2\mathbf{E}\left[XX^\top\right]$$

Let $Q \sim \mathcal{N}\left(0_d, \frac{\sigma^2}{N}\mathbf{E}\left[XX^\top\right]\right)$. By Slutsky's theorem C.1,

$$\left(\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{\epsilon} \xrightarrow{d} \mathbf{E}\left[XX^\top\right]^{-1}Q$$

Thus $\hat{w}_N \xrightarrow{d} w_{\text{true}} + \mathbf{E}\left[XX^\top\right]^{-1}Q$ again by Slutsky's theorem C.1. Finally we observe that

$$w_{\text{true}} + \mathbf{E}\left[XX^\top\right]^{-1}Q \sim \mathcal{N}\left(w_{\text{true}}, \frac{\sigma^2}{N}\mathbf{E}\left[XX^\top\right]^{-1}\right)$$

□

Proof of Lemma F.1.

Proof sketch. Consistency easily follows from the independence of $\bar{S}_k^2$ which equals σ^2 in expectation. For the second claim, assume $K = 2$ and $\mathbf{pop}_k$ is normal for simplicity. Let $R_\alpha = \alpha\bar{S}_1^2 + (1-\alpha)\bar{S}_2^2$ and note

$$\mathrm{Var}\left(R_\alpha\right) = \alpha^2\mathrm{Var}\left(\bar{S}_1^2\right) + (1-\alpha)^2\mathrm{Var}\left(\bar{S}_2^2\right) \quad \Rightarrow \quad \frac{\alpha^*}{1-\alpha^*} = \frac{\mathrm{Var}\left(\bar{S}_2^2\right)}{\mathrm{Var}\left(\bar{S}_1^2\right)}$$

where $\alpha^* = \arg\min_{\alpha\in\mathbb{R}}\mathrm{Var}\left(R_\alpha\right)$. The variance of $\bar{S}_k^2$ can be easily derived by using the normality assumption:

$$\mathrm{Var}\left(\frac{(N_k-1)\bar{S}_k^2}{\sigma^2}\right) = 2(N_k-1) \quad \Leftrightarrow \quad \mathrm{Var}\left(\bar{S}_k^2\right) = \frac{2\sigma^4}{N_k-1} \tag{53}$$

where the first equality comes from (45) and the fact that the variance of $\chi^2(k)$ is $2k$. Then

$$\frac{\alpha^*}{1-\alpha^*} = \frac{N_1-1}{N_2-1} \quad \Leftrightarrow \quad \alpha^* = \frac{N_1-1}{N_1+N_2-2}$$

□

Proof of Lemma F.2.

Proof.

$$\begin{aligned}\sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i}-\bar{Y})^2 &= \sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i}-\bar{Y}_k+\bar{Y}_k-\bar{Y})^2 \\ &= \sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i}-\bar{Y}_k)^2 + \sum_{k=1}^{K}N_k(\bar{Y}_k-\bar{Y})^2 + \sum_{k=1}^{K}(\bar{Y}_k-\bar{Y})\left(\sum_{i=1}^{N_k}(Y_{k,i}-\bar{Y}_k)\right)\end{aligned}$$

The last term is zero because

$$\sum_{i=1}^{N_k}(Y_{k,i}-\bar{Y}_k) = \sum_{i=1}^{N_k}Y_{k,i} - N_k\bar{Y}_k = N_k\bar{Y}_k - N_k\bar{Y}_k = 0$$

□

Proof of Lemma F.3.

Proof. The expected value of Q_{within} is simple,

$$\mathbf{E}\left[Q_{\text{within}}\right] = \mathbf{E}\left[\sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i}-\bar{Y}_k)^2\right] = \sum_{k=1}^{K}(N_k-1)\mathbf{E}\left[\bar{S}_k^2\right] = \sum_{k=1}^{K}(N_k-1)\sigma^2 = (N-K)\sigma^2$$

We now show that $\mathbf{E}\left[Q_{\text{total}}\right] = (N-1)\sigma^2 + \sum_{k=1}^{K}(\mu_k-\mu)^2$. Once this is shown, it follows that $\mathbf{E}\left[Q_{\text{between}}\right] = (K-1)\sigma^2 + \sum_{k=1}^{K}(\mu_k-\mu)^2$. We start by writing

$$\begin{aligned} Q_{\text{total}} = \sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i}-\bar{Y})^2 &= \sum_{k=1}^{K}\sum_{i=1}^{N_k}((Y_{k,i}-\mu_k)-(\bar{Y}-\mu_k))^2 \\ &= \sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i}-\mu_k)^2 + \sum_{k=1}^{K}N_k(\bar{Y}-\mu_k)^2 - 2\sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i}-\mu_k)(\bar{Y}-\mu_k) \end{aligned} \tag{54}$$

We focus on the expected value of each of the three terms. The first is

$$\mathbf{E}\left[\sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i}-\mu_k)^2\right] = \sum_{k=1}^{K}N_k\sigma^2 = N\sigma^2$$

using the fact that $(1/N_k)\sum_{i=1}^{N_k}(Y_{k,i}-\mu_k)^2$ (without Bessel's correction since we're using the true mean) is an unbiased estimator of σ^2. The second is

$$\begin{aligned} \mathbf{E}\left[\sum_{k=1}^{K}N_k(\bar{Y}-\mu_k)^2\right] &= \sum_{k=1}^{K}N_k\mathbf{E}\left[(\bar{Y}-\mu+\mu-\mu_k)^2\right] \\ &= \sum_{k=1}^{K}N_k\mathbf{E}\left[(\bar{Y}-\mu)^2\right] + \sum_{k=1}^{K}N_k(\mu-\mu_k)^2 + 2\sum_{k=1}^{K}N_k\mathbf{E}\left[(\bar{Y}-\mu)(\mu-\mu_k)\right] \\ &= N\underbrace{\mathbf{E}\left[(\bar{Y}-\mu)^2\right]}_{\text{Var}(\bar{Y})=\frac{\sigma^2}{N}} + \sum_{k=1}^{K}N_k(\mu-\mu_k)^2 + 2\sum_{k=1}^{K}N_k(\mu-\mu_k)\underbrace{\mathbf{E}\left[(\bar{Y}-\mu)\right]}_{0} \\ &= \sigma^2 + \sum_{k=1}^{K}N_k(\mu-\mu_k)^2 \end{aligned}$$

Finally, the third is

$$\mathbf{E}\left[\sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i}-\mu_k)(\bar{Y}-\mu_k)\right] = \sum_{k=1}^{K}N_k\mathbf{E}\left[(\bar{Y}_k-\mu_k)(\bar{Y}-\mu_k)\right] \tag{55}$$

We have

$$\begin{aligned} \mathbf{E}\left[(\bar{Y}_k-\mu_k)(\bar{Y}-\mu_k)\right] &= \mathbf{E}\left[\bar{Y}_k\bar{Y}\right] - \mathbf{E}\left[\bar{Y}_k\right]\mu_k - \mathbf{E}\left[\bar{Y}\right]\mu_k + \mu_k^2 \\ &= \mathbf{E}\left[\bar{Y}_k\bar{Y}\right] - \mu\mu_k \end{aligned}$$

where

$$\begin{aligned}
\mathbf{E}\left[\bar{Y}_k\bar{Y}\right] &= \mathbf{E}\left[\bar{Y}_k\left(\sum_{l=1}^{K}\frac{N_l}{N}\bar{Y}_l\right)\right] \\
&= \sum_{l=1}^{K}\frac{N_l}{N}\mathbf{E}\left[\bar{Y}_k\bar{Y}_l\right] \\
&= \frac{N_k}{N}\mathbf{E}\left[\bar{Y}_k^2\right] + \sum_{l\neq k}\frac{N_l}{N}\mathbf{E}\left[\bar{Y}_k\right]\mathbf{E}\left[\bar{Y}_l\right] && \text{(independence of } \bar{Y}_1 \dots \bar{Y}_K\text{)} \\
&= \frac{N_k}{N}\left(\frac{\sigma^2}{N_k} + \mu_k^2\right) + \sum_{l\neq k}\frac{N_l}{N}\mu_k\mu_l && \left(\text{Var}\left(\bar{Y}_k\right) = \tfrac{\sigma^2}{N_k} = \mathbf{E}\left[\bar{Y}_k^2\right] - \mu_k^2\right) \\
&= \frac{\sigma^2}{N} + \frac{N_k}{N}\mu_k^2 + \sum_{l\neq k}\frac{N_l}{N}\mu_k\mu_l \\
&= \frac{\sigma^2}{N} + \sum_{l=1}^{K}\frac{N_l}{N}\mu_k\mu_l
\end{aligned}$$

Going back to (55),

$$\begin{aligned}
\mathbf{E}\left[\sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i}-\mu_k)(\bar{Y}-\mu_k)\right] &= \sum_{k=1}^{K}N_k\mathbf{E}\left[(\bar{Y}_k-\mu_k)(\bar{Y}-\mu_k)\right] \\
&= \sum_{k=1}^{K}N_k\left(\mathbf{E}\left[\bar{Y}_k\bar{Y}\right] - \mu\mu_k\right) \\
&= \sum_{k=1}^{K}N_k\left(\frac{\sigma^2}{N} + \sum_{l=1}^{K}\frac{N_l}{N}\mu_k\mu_l - \mu\mu_k\right) \\
&= \sigma^2 + \sum_{k,l}\frac{N_kN_l}{N}\mu_k\mu_l - \mu\sum_{k=1}^{K}N_k\mu_k \\
&= \sigma^2
\end{aligned}$$

The last equality holds because

$$\mu\sum_{k=1}^{K}N_k\mu_k = \left(\sum_{l=1}^{K}\frac{N_l}{N}\mu_l\right)\left(\sum_{k=1}^{K}N_k\mu_k\right) = \sum_{k,l}\frac{N_kN_l}{N}\mu_k\mu_l$$

Finally, putting everything together into (54)

$$\begin{aligned}
\mathbf{E}\left[\sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i}-\bar{Y})^2\right] &= \mathbf{E}\left[\sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i}-\mu_k)^2\right] + \mathbf{E}\left[\sum_{k=1}^{K}N_k(\bar{Y}-\mu_k)^2\right] - 2\mathbf{E}\left[\sum_{k=1}^{K}\sum_{i=1}^{N_k}(Y_{k,i}-\mu_k)(\bar{Y}-\mu_k)\right] \\
&= N\sigma^2 + \sigma^2 + \sum_{k=1}^{K}N_k(\mu-\mu_k)^2 - 2\sigma^2 \\
&= (N-1)\sigma^2 + \sum_{k=1}^{K}N_k(\mu-\mu_k)^2
\end{aligned}$$

□

Proof of Lemma J.1.

Proof. We will establish the equivalence between (1, 2) and the equivalence between (2, 3).

- (1 ⇒ 2) Given any $z \in S$, define $f_z : \mathbb{R} \to \mathbb{R}$ by $f_z(t) = ||x - y - tz||^2$. By premise the minimum of f_z is achieved at $t = 0$. Since f_z is strongly convex we may phrase the optimality of $t = 0$ as the stationary condition $f_z'(0) = 0$. Then we have

$$\begin{aligned} & f_z'(t) = -2z^\top(x - y - tz) \\ \Rightarrow\ & f_z'(0) = -2z^\top(x - y) = -2z^\top(x - Px) = -2z^\top(I_{N\times N} - P)x = 0 \end{aligned} \tag{56}$$

  Recall that $I_{N\times N} - P$ is a projection onto null(P), so null$(P) = \left\{(I_{N\times N} - P)x : x \in \mathbb{R}^N\right\}$. Since (56) holds for all $x \in \mathbb{R}^N$ and $z \in S$, it implies range $(P) \perp$ null(P).

- (2 ⇒ 1) For any $x \in \mathbb{R}^N$ and $z \in S$, define $y = Px \in$ range (P) and note that $x - y = (I_{N\times N} - P)x \in$ null(P) and $y - z \in$ range (P) (definition of subspace). Thus we have

$$\begin{aligned} ||x - z||^2 &= ||x - y + y - z||^2 \\ &= ||x - y||^2 + ||y - z||^2 + 2(x - y)^\top(y - z) \\ &= ||x - y||^2 + ||y - z||^2 \\ &\geq ||x - y||^2 \end{aligned}$$

  where the equality holds iff $z = y$.

- (2 ⇒ 3) If dim (range (P)) $= 0$ or dim (null(P)) $= 0$ then either $P = I_{N\times N}$ or $P = 0_{N\times N}$ so trivially $P = P^\top$. Now assume that dim (range (P)) and dim (null(P)) are both positive. We may select nonzero $u \in$ null(P) and $x \in \mathbb{R}^N$ such that $Px \in$ range (P) is nonzero. If $P \neq P^\top$,

$$x^\top P^\top u \neq x^\top P u = 0$$

  This contradicts the premise that $v^\top u = 0$ for all $v \in$ range (P) and $u \in$ null(P). Thus $P = P^\top$.

- (3 ⇒ 2) If $u = Px$ for some $x \in \mathbb{R}^N$ and $Pv = 0_N$, then $u^\top v = x^\top P^\top v = x^\top P v = 0$. Thus range $(P) \perp$ null(P).

□

Proof of Lemma G.1.

Proof. The loss $J : \mathbb{R} \times \mathbb{R}^{K-1} \to \mathbb{R}$ is

$$J(\alpha, \beta) = \sum_{i=1}^{N} \left(y_i - \alpha - \beta^\top C_A^d(a_i)\right)^2 = \sum_{i=1}^{N} \left(y_i - \alpha - \beta_{a_i}\left[[a_i > 1]\right]\right)^2$$

where we use the indexing $\beta = (\beta_2 \dots \beta_K)$. J is strongly convex and its partial derivatives are

$$\begin{aligned} \frac{\partial J(\beta)}{\partial \alpha} &= -2\sum_{i=1}^{N}\left(y_i - \alpha - \beta_{a_i}\left[[a_i > 1]\right]\right) \\ \frac{\partial J(\beta)}{\partial \beta_k} &= -2\sum_{i:a_i=k}\left(y_i - \alpha - \beta_k\right) \qquad\qquad k \in \{2 \dots K\} \end{aligned}$$

Setting $\frac{\partial J(\beta)}{\partial \beta_k} = 0$, we have

$$\sum_{i:a_i=k}\left(y_i - \alpha - \beta_k\right) = \sum_{i:a_i=k} y_i - \mathbf{count}(k)\alpha - \mathbf{count}(k)\beta_k = 0 \qquad \Leftrightarrow \qquad \beta_k = \bar{y}_k - \alpha \tag{57}$$

Setting $\frac{\partial J(\beta)}{\partial \alpha} = 0$, we have

$$\sum_{i=1}^{N}\left(y_i - \alpha - \beta_{a_i}\left[[a_i > 1]\right]\right) = \sum_{i:a_i=1}\left(y_i - \alpha\right) + \sum_{i:a_i\geq 2}\left(y_i - \alpha - \beta_{a_i}\right) = 0 \tag{58}$$

Using condition (57) on the second term, we have

$$\sum_{i:a_i\geq 2}(y_i-\alpha-\beta_{a_i})=\sum_{k=2}^{K}\sum_{i:a_i=k}(y_i-\alpha-\beta_k)=\sum_{k=2}^{K}\sum_{i:a_i=k}(y_i-\bar{y}_k)=0$$

Thus condition (58) implies $\alpha=\bar{y}_1$ □

Proof of Lemma G.2.

Proof. The loss $J:\mathbb{R}\times\mathbb{R}^{K-1}\times\mathbb{R}^{L-1}\times\mathbb{R}^{(K-1)(L-1)}\to\mathbb{R}$ is

$$\begin{aligned}J(\alpha,\beta,\gamma,\kappa)&=\sum_{i=1}^{N}\left(y_i-\alpha-\beta^\top C_A^d(a_i)-\gamma^\top C_B^d(b_i)-\kappa^\top C_A^d(a_i)\otimes C_B^d(b_i)\right)^2\\&=\sum_{i=1}^{N}\left(y_i-\alpha-\beta_{a_i}\left[[a_i>1]\right]-\gamma_{b_i}\left[[b_i>1]\right]-\kappa_{a_i,b_i}\left[[a_i>1,b_i>1]\right]\right)^2\end{aligned}$$

where we use the indexing $\beta=(\beta_2\dots\beta_K)$, $\gamma=(\gamma_2\dots\gamma_L)$, and $\kappa=(\kappa_{2,2}\dots\kappa_{K,L})$. The partial derivative wrt $\kappa_{k,l}$ is

$$\frac{\partial J(\beta)}{\partial\kappa_{k,l}}=-2\sum_{i:a_i=k,b_i=l}(y_i-\alpha-\beta_k-\gamma_l-\kappa_{k,l})$$

Setting this to zero we have

$$\kappa_{k,l}=\bar{y}_{k,l}-\alpha-\beta_k-\gamma_l \tag{59}$$

The partial derivative wrt γ_l is

$$\frac{\partial J(\beta)}{\partial\gamma_l}=-2\underbrace{\sum_{i:b_i=l}\left(y_i-\alpha-\beta_{a_i}\left[[a_i>1]\right]-\gamma_l-\kappa_{a_i,l}\left[[a_i>1]\right]\right)}_{①}$$

where we can write

$$①=\sum_{i:a_i=1,\ b_i=l}(y_i-\alpha-\gamma_l)+\underbrace{\sum_{i:a_i>1,\ b_i=l}(y_i-\alpha-\beta_{a_i}-\gamma_l-\kappa_{a_i,l})}_{②} \tag{60}$$

Using (59), we have

$$\begin{aligned}②&=\sum_{k\geq 2}\sum_{i:a_i=k,\ b_i=l}(y_i-\alpha-\beta_k-\gamma_l-\kappa_{k,l})\\&=\sum_{k\geq 2}\sum_{i:a_i=k,\ b_i=l}(y_i-\bar{y}_{k,l})=0\end{aligned}$$

Thus ① $=0$ implies $\gamma_l=\bar{y}_{1,l}-\alpha$. Similarly, it can be verified that the other stationary points are $\beta_k=\bar{y}_{k,1}-\alpha$ and $\alpha=\bar{y}_{1,1}$. □

Proof of Lemma G.3.

Proof. Recall that the sum coding $C_A^s(k)\in\mathbb{R}^{K-1}$ is defined for $k=1\dots K$ as $(0,\dots,1,\dots,0)$ if $k<K$ and $(-1,-1,\dots,-1)$ if $k=K$. Thus we can write the loss $J:\mathbb{R}\times\mathbb{R}^{K-1}\to\mathbb{R}$ as

$$\begin{aligned}J(\alpha,\beta)&=\sum_{i=1}^{N}\left(y_i-\alpha-\beta^\top C_A^s(a_i)\right)^2\\&=\sum_{i:a_i<K}(y_i-\alpha-\beta_{a_i})^2+\sum_{i:a_i=K}\left(y_i-\alpha+\sum_{k=1}^{K-1}\beta_k\right)^2\end{aligned}$$

The partial derivative wrt α is

$$\frac{\partial J(\alpha,\beta)}{\partial \alpha} = -2 \underbrace{\sum_{i:a_i<K} (y_i - \alpha - \beta_{a_i})}_{①} -2 \underbrace{\sum_{i:a_i=K} \left(y_i - \alpha + \sum_{k=1}^{K-1} \beta_k \right)}_{②}$$

so setting it to zero is equivalent to setting ① + ② = 0. Using the fact that the data is balanced so that we have M samples for each level $1 \dots K$ (in particular $N = MK$), we have

$$\begin{aligned} ① &= \sum_{i:a_i<K} y_i - M(K-1)\alpha - M \sum_{k=1}^{K-1} \beta_k \\ ② &= \sum_{i:a_i=K} y_i - M\alpha + M \sum_{k=1}^{K-1} \beta_k \\ ① + ② &= \sum_{i=1}^{N} y_i - N\alpha = 0 \end{aligned}$$

Thus $\alpha = \bar{y}$. For any $l \in \{1 \dots K-1\}$ the partial derivative wrt β_l is

$$\frac{\partial J(\alpha,\beta)}{\partial \beta_l} = -2 \underbrace{\left(\sum_{i:a_i=l} y_i - \alpha - \beta_l \right)}_{③_l} +2 \left(\sum_{i:a_i=K} y_i - \alpha + \sum_{k=1}^{K-1} \beta_k \right)$$

so setting it to zero is equivalent to setting ③ − ② = 0. We have

$$\begin{aligned} ③_l &= \sum_{i:a_i=l} y_i - M\alpha - M\beta_l \\ ③_l - ② &= \sum_{i:a_i=l} y_i - \sum_{i:a_i=K} y_i - M\beta_l - M \sum_{k=1}^{K-1} \beta_k = 0 \end{aligned}$$

Summing both sides over $l = 1 \dots K-1$, we have

$$\begin{aligned} \sum_{l=1}^{K-1} \left(③_l - ② \right) &= \sum_{i:a_i<K} y_i - (K-1) \sum_{i:a_i=K} y_i - M \sum_{l=1}^{K-1} \beta_l - M(K-1) \sum_{k=1}^{K-1} \beta_k \\ &= \sum_{i=1}^{N} y_i - K \sum_{i:a_i=K} y_i - MK \sum_{k=1}^{K-1} \beta_k = 0 \end{aligned}$$

This implies

$$M \sum_{k=1}^{K-1} \beta_k = \frac{1}{K} \sum_{i=1}^{N} y_i - \sum_{i:a_i=K} y_i$$

Therefore

$$\begin{aligned} ③_l - ② &= \sum_{i:a_i=l} y_i - \sum_{i:a_i=K} y_i - M\beta_l - M \sum_{k=1}^{K-1} \beta_k \\ &= \sum_{i:a_i=l} y_i - \sum_{i:a_i=K} y_i - M\beta_l - \frac{1}{K} \sum_{i=1}^{N} y_i + \sum_{i:a_i=K} y_i \\ &= \sum_{i:a_i=l} y_i - M\beta_l - \frac{1}{K} \sum_{i=1}^{N} y_i = 0 \end{aligned}$$

which implies $\beta_l = \bar{y}_l - \bar{y}$. □

Lemma K.2. If $K = 2$ and $M = N_1 = N_2$, then $\bar{S}^2_{\text{between}} = (M/2)(\bar{Y}_1 - \bar{Y}_2)^2$

Proof. The statement is equivalent to

$$\begin{aligned}
&2M((\bar{Y}_1 - \bar{Y})^2 + (\bar{Y}_2 - \bar{Y})^2) = M(\bar{Y}_1 - \bar{Y}_2)^2 \\
&\Leftrightarrow (\bar{Y}_1 - \bar{Y})^2 + (\bar{Y}_2 - \bar{Y})^2 + 2(\bar{Y}_1 - \bar{Y})(\bar{Y}_2 - \bar{Y}) = 0 \\
&\Leftrightarrow ((\bar{Y}_1 - \bar{Y}) + (\bar{Y}_2 - \bar{Y}))^2 = 0
\end{aligned}$$

This holds since

$$\bar{Y}_1 + \bar{Y}_2 - 2\bar{Y} = \frac{1}{M}\sum_{i=1}^{M} Y_{1,i} + \frac{1}{M}\sum_{i=1}^{M} Y_{2,i} - \frac{1}{M}\left(\sum_{i=1}^{M} Y_{1,i} + \sum_{i=1}^{M} Y_{2,i}\right) = 0$$

□