STAT24510 (W25): Statistical Theory and Methods IIa

Lecturer: Mei Wang
Notes by: Aden Chen

Thursday 30th January, 2025

Contents

1 Introduction

The goal of statistics is often to estimate a (population) parameter θ. From data, we may obtain point estimates $\hat{\theta}$ that depends on data, and construct confidence intervals to quantify the uncertainty of the estimate, with which we can conduct hypothesis testing.

In this course we will start from confidence intervals and hypothesis testing, and then move on to linear models. This course will be less structured; intend, it will be more like a collection of methods/producers.

1.1 Examples of Construction: Review of Wald and Wilson

1.1.1 Pivotal method

Example 1.1 (Pivotal method, Normal, Known σ^2). Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 5^2)$, $i = 1, \ldots, n$. Our goal is to construct a $1 - \alpha$ CI for μ. We have the MLE estimator $\hat{\mu} = \overline{X} = n^{-1} \sum_{i=1}^{n} X_i$.

Note that

$$\frac{\overline{X} - \mathrm{E}[X]}{\sqrt{\mathrm{Var}(X)}} = \frac{\overline{X} - \mu}{\sqrt{5^2/n}} \sim \mathcal{N}(0, 1).$$

In particular, note that the left side is a function of data and parameters, while the right side is free of parameters.

We thus have

$$\mathbb{P}\left(z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sqrt{5^2/n}} \leq z_{1-\alpha/2} \right) = 0.95,$$

using which we can construct the CI of μ: With $I \coloneqq \left[\overline{X} - z_{1-\alpha/2}\sqrt{5^2/n}, \overline{X} + z_{1-\alpha/2}\sqrt{5^2/n} \right]$, we have $\mathbb{P}\left(\mu \in I\right) = 1 - \alpha$.

Notice that we obtain a probability statement of random interval containing a fixed quantity from a probability statement of a fixed interval containing a random quantity.

Example 1.2 (Pivotal method, Normal, Known μ). Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(4, \sigma^2)$, $i = 1, \ldots, n$. The goal: CI for σ^2, i.e., to find random variables L and U such that $\mathbb{P}\left(L \leq \sigma^2 \leq U\right) = 1 - \alpha$.

Note that

$$Y_i \coloneqq \frac{X_i - 4}{\sigma} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$$

and thus

$$T_n \coloneqq \sum Y_i^2 = \sum \left(\frac{X_i - 4}{\sigma} \right)^2 \sim \chi^2_n.$$

Again, we obtained a function of data and parameters that follows a known distribution. From

$$\mathbb{P}\left(\chi^2_{n,\alpha/2} \leq T_n \leq \chi^2_{n,1-\alpha/2} \right) = 1 - \alpha$$

we may again obtain the CI for σ^2,

$$\left[\frac{\sum (X_i - 4)^2}{\chi^2_{n,1-\alpha/2}}, \frac{\sum (X_i - 4)^2}{\chi^2_{n,\alpha/2}} \right].$$

Example 1.3 (Pivot failing). Let $X_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. The goal: CI for p. The MLE for p is $\hat{p} = \overline{X}$, thus we may be temped to try

$$T_n \coloneqq \frac{\hat{p} - p}{\sqrt{\text{Var}(\hat{p})}},$$

but the distribution of T_n depends on p. The method of pivots fail.

1.1.2 Asymptotic CI

I.e., when we have large sample size n.

Example 1.4 (Wald CI). Let X_i be iid with mean μ and variance σ^2. From the CLT we have

$$\sqrt{n}\left(\overline{X} - \mu\right) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma^2),$$

that is,

$$\lim_{n\to\infty} \mathbb{P}(T_n \le x) = P(Z \le x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, \mathrm{d}t.$$

Thus we have

$$\mathbb{P}\left(z_{\alpha/2} \le \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \le z_{1-\alpha/2}\right) \approx 1 - \alpha.$$

When σ^2 is known, we may derive an approximate CI for μ:

$$\mathbb{P}\left(\overline{X} - z_{1-\alpha/2}\sqrt{\sigma^2/n} \le \mu \le \overline{X} + z_{1-\alpha/2}\sqrt{\sigma^2/n}\right) \approx 1 - \alpha.$$

When σ^2 is unknown: If there exists random variables $U_n \to_{\text{p}} \sigma^2$ (that is, $\lim \mathbb{P}\left(U_n = \sigma^2\right) = 1$), then

$$T_n \coloneqq \frac{\overline{X} - \mu}{\sqrt{U_n/n}} = \frac{\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{U_n/\sigma^2}}$$

where $(\overline{X} - \mu)/\sqrt{\sigma^2/n} \xrightarrow{\mathscr{D}} \mathcal{N}(0,1)$ and $\sqrt{U_n/\sigma^2} \to_{\text{p}} 1$, and thus by Slutsky's theorem we have

$$T_n \xrightarrow{\mathscr{D}} \mathcal{N}(0,1),$$

using which we can again construct an approximate CI. Note that we used asymptotic approximation multiple times. This is called the Wald confidence interval.

Example 1.5 (Wald CI). Let $X_i \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. The goal: asymptotic CI for λ. Note that we have the MLE of λ, $\hat{\lambda} = \overline{X}$, with $\text{E}[\hat{\lambda}] = \lambda$ and $\text{Var}[\hat{\lambda}] = \lambda/n$. We then have

$$\frac{\overline{X} - \lambda}{\sqrt{\lambda/n}} \approx Z \sim \mathcal{N}(0,1).$$

We approximate a second time: $\frac{\overline{X}-\lambda}{\sqrt{\hat{\lambda}/n}} \approx Z$, from which we obtain the Wald CI for λ:

$$\left[\hat{\lambda} - z_{1-\alpha/2}\sqrt{\hat{\lambda}/n}, \hat{\lambda} + z_{1-\alpha/2}\sqrt{\hat{\lambda}/n}\right].$$

Example 1.6 (Wilson's method). Assume the same setup as above. Again, we use

$$\frac{\overline{X} - \lambda}{\sqrt{\lambda/n}} \approx Z \sim \mathcal{N}(0, 1),$$

which gives

$$\mathbb{P}\left(z_{\alpha/2} \leq \frac{\overline{X} - \lambda}{\sqrt{\lambda/n}} \leq z_{1-\alpha/2}\right) = \mathbb{P}\left(\left(\frac{\overline{X} - \lambda}{\sqrt{\lambda/n}}\right)^2 \leq z_{1-\alpha/2}^2\right) \approx 1 - \alpha.$$

Solving for λ in the middle expression gives the Wilson CI. We used one fewer approximation.

2 Confidence Intervals

2.1 Constructing CI

- Exact (i.e., coverage probability is exactly $1-\alpha$) confidence intervals: **pivot method**:
  1. Find statistic $T_n = T(X, \theta)$ whose distribution is known and independent of θ. Such a statistic is called a pivot.
  2. Using knowledge on the distribution, find c_L and c_U such that $\mathbb{P}(c_L \le T_n \le c_R) = 1-\alpha$.
  3. Convert the probability statement of the pivot to a probability statement of the parameter.
- Approximation methods, or asymptotic (sample size n is large) methods.
  - **Wald confidence interval** (the default CI produced by standard software): more than one approximations.
  - **Wilson's confidence interval /** score method **/** duality method: using CLT once.
    - Wilson's CI is better in the sense that the actual coverage is closer to the desired coverage.
  - **Variance stabilization transformation** (VST) method.

Example 2.1 (Pivot, Normal). Assume $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, $i = 1, \dots, n$. Goal: $(1-\alpha)$ CI for μ.

1. Recall that
$$T_n = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}.^{1}$$
2. We have
$$\mathbb{P}\left(t_{n-1,\alpha/2} \le \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \le t_{n-1,1-\alpha/2} \right) = 1-\alpha.$$
3. Rearrangement gives
$$\mathbb{P}\left(\overline{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} \right) = 1-\alpha.$$

Example 2.2 (Wald, Bernoulli). Let $X_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, so that $\sum X_i \sim \text{Binomial}(n, p)$. Goal: $(1-\alpha)$ CI for p. Note that $\hat{p} = \overline{X}$ is the MLE of p. CLT gives
$$\frac{\hat{p} - p}{\sqrt{\text{Var}(\hat{p})}} \approx \mathcal{N}(0, 1).$$

[1] t distribution has fatter tails than $\mathcal{N}(0, 1)$.

From Slutsky's theorem we have a **second approximation**

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n} \approx \frac{\hat{p}(1-\hat{p})}{n}.$$

This second approximation is characteristic of Wald's method. Using this second approximation we have

$$\frac{\overline{X} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \approx \mathcal{N}(0,1),$$

with which we can construct the desired CI:

$$\left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n}\right].$$

Example 2.3 (Wilson's CI; Bernoulli). We assume the same Bernoulli setup. Note that we obtained using just the CLT that

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

We rewrite the middle expression as

$$n\,(\hat{p} - p)^2 \leq z^2_{1-\frac{\alpha}{2}} p(1-p).$$

Solving a quadratic equation gives the desired Wilson's CI.

Remark 2.4.

- Note that the Wald CI is centered at $\hat{p}$, but Wilson's is not.
- Wilson's CI will always be contained in $[0,1]$; the lower bound of Wald might be negative.

2.2 Variance Stabilization Transformation (VST) Method

Example 2.5 (VST, Poisson). Let $X_i \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. Note that

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\lambda/n}} \approx \mathcal{N}(0,1).$$

and

$$\sqrt{n}(\hat{\lambda} - \lambda) \approx \mathcal{N}(0, \lambda).$$

Goal: find a transformation (usually smooth) g such that

$$\sqrt{n}(g(\hat{\lambda}) - g(\lambda)) \approx \mathcal{N}(0,1).$$[2]

[2] Or just a normal distribution with fixed variance.

Tool: delta method (Taylor expansion for random variables). By Taylor expansion,

$$g(\hat{\lambda}) = g(\lambda) + g'(\lambda)(\hat{\lambda} - \lambda) + \frac{g''(\lambda)}{2}(\hat{\lambda} - \lambda)^2 + \dots.$$

We thus have the approximation

$$g(\hat{\lambda}) \approx g(\lambda) + g'(\lambda)(\hat{\lambda} - \lambda) + O(n^{-1}),$$

where the last term follows from the fact that $(\hat{\lambda} - \lambda)/\sqrt{\lambda/n} \approx \mathcal{N}(0, 1)$. Then,

$$\mathrm{E}[g(\hat{\lambda})] \approx g(\lambda) + g'(\lambda)\,\mathrm{E}[\hat{\lambda} - \lambda].$$

Since $\hat{\lambda}$ is unbiased, we have

$$\sqrt{n}(g(\hat{\lambda}) - g(\lambda)) \approx g'(\lambda)\sqrt{n}(\hat{\lambda} - \lambda),$$

giving

$$\sqrt{n}(g(\hat{\lambda}) - g(\lambda)) \approx \mathcal{N}(0, [g'(\lambda)]^2\lambda),$$

where we used the approximation

$$\mathrm{Var}(g(\hat{\lambda})) = \mathrm{E}[(g(\hat{\lambda}) - \mathrm{E}[g(\hat{\lambda})])^2] \approx [g'(\lambda)]^2\,\mathrm{Var}(\hat{\lambda}).$$

To obtain $\sqrt{n}(g(\hat{\lambda}) - \lambda) \approx \mathcal{N}(0, 1)$, we need only $g'(\lambda)^2 = 1/\lambda$. $g(\lambda) = 2\sqrt{\lambda}$ will do.

That is, we have

$$\sqrt{n}(2\sqrt{\hat{\lambda}} - 2\sqrt{\lambda}) \approx \mathcal{N}(0, 1).$$

Using this we can obtain a CI for λ. Note that the left endpoint of the CI for $\sqrt{\lambda}$ may be negative, so in such cases when obtaining the CI for λ we need to use 0 instead.

2.3 Delta Method

Consider an unbiased estimator $\hat{\theta}$. $\mathrm{E}[\hat{\theta}] = \theta$. Assume that the variance is a function of the mean, $\mathrm{Var}[\hat{\theta}] = v(\theta)$.[3] Goal: find a function $g(t)$ such that $\mathrm{Var}[g(\hat{\theta})]$ is constant. Taylor gives

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta).$$

Note that we have

$$\mathrm{E}[g(\hat{\theta})] \approx g(\theta) + g'(\theta)\,\mathrm{E}[\hat{\theta} - \theta] = g(\theta).$$

and

$$\mathrm{Var}[g(\hat{\theta})] = \mathrm{E}\left[\left(g(\hat{\theta}) - g(\theta)\right)^2\right] \approx \mathrm{E}\left[g'(\theta)^2\left(\hat{\theta} - \theta\right)^2\right] = g'(\theta)^2\,\mathrm{Var}[\hat{\theta}],$$

which we want to be constant, say 1. Thus we seek g such that

$$g'(\theta) = \frac{1}{\sqrt{v(\theta)}}.$$

A choice is of course $g(\theta) = \int 1/\sqrt{v(\theta)}\,\mathrm{d}\theta$.

[3] This is a common case.

Example 2.6 (Delta Method, Exponential, distribution of MLE). Consider $X_i \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$. Recall that the MLE of λ is $\hat{\lambda} = 1/\overline{X}$. By the CLT we have for large n that

$$\frac{\overline{X} - \mathrm{E}\, X}{\sqrt{\mathrm{Var}\, \overline{X}}} \approx \mathcal{N}(0, 1),$$

That is,

$$\frac{\overline{X} - \frac{1}{\lambda}}{\sqrt{\frac{1}{n\lambda^2}}} = \sqrt{n}\left(\overline{X} - \frac{1}{\lambda}\right) \approx \mathcal{N}(0, \lambda^{-2}).$$

Goal: find an approximate distribution of $\hat{\lambda} = 1/\overline{X}$ using Delta Method. Using $g(t) = t^{-1}$ we have

$$\sqrt{n}\left(g(\hat{\lambda}) - g(\lambda)\right) \approx \mathcal{N}(0, [g'(\lambda)]^2 \,\mathrm{Var}(\hat{\lambda})) = \mathcal{N}(0, \lambda^2).$$

(Note that Fisher's theorem gives another way to derive the above.)

Example 2.7 (VST, Exponential, CI). Consider the same setup as above. We can then derive an approximate CI for λ using VST. Recall that from Delta Method, we have

$$\sqrt{n}\left(g(\hat{\lambda}) - g(\lambda)\right) \approx \mathcal{N}(0, [g'(\lambda)]^2 \,\mathrm{Var}(\hat{\lambda})).$$

Thus we seek g such that $g'(\lambda)^2\lambda^2$ is a constant, say 1, or $g'(\lambda) = 1/\lambda$. It is easy to see that $g = \log$ is such an option. Then,

$$\sqrt{n}\left(\log\hat{\lambda} - \log\lambda\right) \approx \mathcal{N}(0, 1),$$

using which we can obtain an approximate CI for $\log\hat{\lambda}$ and then $\hat{\lambda}$:

$$\left[\hat{\lambda}\exp\left(-\frac{z_{1-\alpha/2}}{\sqrt{n}}\right), \hat{\lambda}\exp\left(\frac{z_{1-\alpha/2}}{\sqrt{n}}\right)\right]$$

Theorem 2.8 (Delta Method)**.** *If X_n is such that*

$$\sqrt{n}(X_n - \theta) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma^2)$$

and g is continuously differentiable, then

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma^2[g'(\theta)]^2).$$

Remark 2.9. Intuition: We can write

$$\sqrt{n}(g(X_n) - g(\theta)) = g'(\tilde{\theta}_n)\sqrt{n}(X_n - \theta), \quad \tilde{\theta}_n \in (x_n, \theta).$$

We know that $g'(\tilde{\theta}_n) \to_{\mathrm{p}} g'(\theta)$ and $\sqrt{n}(X_n - \theta) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma^2)$, so Slutsky's gives the desired result.

3 Change of Variable / Variable Transformation

3.1 The Discrete Case

Let X and $Y = g(X)$ be discrete. Then,

$$\mathbb{P}(Y = k) = \sum_{X:g(X)=k} \mathbb{P}(X = k).$$

3.2 The Continuous, 1D Case

Consider a continuous random variable X and transformation $Y = g(X)$ with g smooth with derivative vanishing nowhere, that is, $g' \neq 0$. These conditions guarantee that Y is also continuous. Suppose further that g is smooth and $g' \neq 0$. We have then that

$$f_Y(y) = f_X(x)\frac{1}{|g'(x)|} = f_X\left(g^{-1}(y)\right)\left|\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right|.$$

Example 3.1. Let $g' > 0$. Note that

$$\mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

Thus

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(Y \leq y) = \frac{\mathrm{d}}{\mathrm{d}y}F_X\left(g^{-1}(y)\right) = f_X\left(g^{-1}(y)\right)\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y).$$

Similarly, if $g' < 0$, then

$$\mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)).$$

Then

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(Y \leq y) = -f_X(g^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y) = f_X(g^{-1}(y))\left|\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right|.$$

In general, if g is piecewise monotone, then

$$f_Y(y) = \sum_{\{x:g(x)=y\}} f_X(x)\frac{1}{|g'(x)|}.$$

Example 3.2. $Y = X^2$.

$$f_Y(y) = \sum_{\{x:x^2=y\}} f_X(x)\frac{1}{|2x|} = f_X(\sqrt{y})\frac{1}{2\sqrt{y}} + f_X(-\sqrt{y})\frac{1}{\left|-2\sqrt{y}\right|}.$$

3.3 The Continuous, 2D Case

Suppose X and Y are continuous with densities f_X and f_Y and joint density $f_{X,Y}$. Consider

$$\begin{cases} U = g_1(X, Y) \\ V = g_2(X, Y) \end{cases}$$

Suppose g_i's are smooth. Question: what is the joint density of U and V?

$$f_{U,V}(u, v) = f_{X,Y}(x, y)|J|,$$

where

$$\begin{aligned} J = \det \frac{\partial(x, y)}{\partial(u, v)} &= \det \begin{bmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{bmatrix} \\ &= \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u} \end{aligned}$$

and we assume $J \neq 0$.

We have also

$$f_{XY}(x, y) = f_{UV}(u, v)|J|^{-1},$$

where

$$J^{-1} = \det \frac{\partial(u, v)}{\partial(x, y)} = \frac{1}{J}.$$

Example 3.3 (2D Change of Variables, Quotient). Let $U = X/Y$ with $Y \neq 0$ and $V = Y$. We have then that

$$\begin{cases} X = UV \\ Y = V \end{cases}.$$

Then

$$\frac{\partial(x, y)}{\partial(u, v)} = \begin{bmatrix} v & u \\ 0 & 1 \end{bmatrix}$$

so $J = v$. Alternatively,

$$\det \frac{\partial(u, v)}{\partial(x, y)} = \det \begin{bmatrix} \frac{1}{y} & -\frac{x}{y^2} \\ 0 & 1 \end{bmatrix} = \frac{1}{y} = J^{-1}.$$

This gives

$$f_{UV}(u, v) = f_{XY}(uv, v)|v|.$$

If we assume further that $X \perp\!\!\!\perp Y$, then

$$f_{UV}(u, v) = f_X(uv) f_Y(v)|v|.$$

4 Binomial and Poisson Distributions

Consider Binomial(n, p_n) with $np_n \to \lambda$ (thus $p_n \to 0$). We have then that

$$\begin{aligned}\mathbb{P}(X_n = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n!}{k!(n-k)!} \left(\frac{\lambda_n}{n}\right)^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-k+1}{n} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}.\end{aligned}$$

Since k is fixed, as $n \to \infty$, and recalling that $(1 + X/n)^n \to e^X$, we have

$$\mathbb{P}(X = k) \to \frac{\lambda^k}{k!} e^{-\lambda}.$$

That is, if $np \to \lambda$, we have Binomial$(n, p) \to$ Poisson(λ). In this sense, the Poisson distribution is the limit of the binomial distribution, and we can thus use it as an approximation for the binomial distribution.

But note that for large n and np, since a Binomial is the sum of Bernoulli's, the CLT can be used to approximate the binomial distribution as $\mathcal{N}(np, np(1-p))$. Similarly, the Poisson distribution can be approximated as $\mathcal{N}(\lambda, \lambda)$.

Example 4.1. Let $X \sim$ Binomial(n, p). The CLT gives

$$\mathbb{P}\left(\frac{X - np}{\sqrt{np(1-p)}} \le x\right) \approx \Phi(x)$$

and thus

$$\begin{aligned}\mathbb{P}(X = k) = \mathbb{P}(k-1 < X \le k) &= \mathbb{P}\left(\frac{k-1-np}{\sqrt{np(1-p)}} < \frac{X-np}{\sqrt{np(1-p)}} \le \frac{k-np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-1-np}{\sqrt{np(1-p)}}\right).\end{aligned}$$

It turns out that the **continuity correction** almost always gives a better approximation:

$$\begin{aligned}\mathbb{P}(X = k) = \mathbb{P}(k-0.5 < X \le k+0.5) &= \mathbb{P}\left(\frac{k-0.5-np}{\sqrt{np(1-p)}} < \frac{X-np}{\sqrt{np(1-p)}} \le \frac{k+0.5-np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{k+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-0.5-np}{\sqrt{np(1-p)}}\right).\end{aligned}$$

5 Correlation

5.1 Bivariate Normal

Let (X, Y) be bivariate normal. That is, X and Y have joint density

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right).$$

If $X \perp\!\!\!\perp Y$, then $\rho = 0$ and the joint density simplifies to $f_X(x) f_Y(y)$.

5.2 Bivariate Normal, the Standard Case

The standard bivariate normal (U, V) is

$$f_{UV}(u,v) = \frac{1}{2\pi(1-\rho^2)} \exp\left(-\frac{1}{2(1-\rho^2)}(u^2+v^2-2\rho uv)\right).$$

Note that

$$u^2 + v^2 - 2\pi uv = (v-\rho u)^2 + (1-\rho^2)u^2.$$

Thus we have

$$\begin{aligned} f_V(v) = \int_{-\infty}^{\infty} f_{UV}(u,v)\,\mathrm{d}v &= \frac{1}{2\pi(1-\rho^2)} \int_{-\infty}^{\infty} \exp\left(-\frac{(1-\rho^2)u^2}{2(1-\rho^2)}\right) \exp\left(-\frac{(v-\rho v)^2}{2(1-\rho^2)}\right) \mathrm{d}v \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(v-\rho u)^2}{2(1-\rho^2)}\right) \mathrm{d}v. \end{aligned}$$

So the marginal distributions are also normal. One can verity similarly that $\mathrm{E}[UV] = \rho$, giving

$$\mathrm{Cov}(U,V) = \mathrm{E}(UV) - \mathrm{E}(U)\,\mathrm{E}(V) = \rho$$

and

$$\mathrm{Corr}(U,V) = \frac{\mathrm{Cov}(U,V)}{\mathrm{Var}(U)\,\mathrm{Var}(V)} = \rho.$$

Note that

$$\begin{aligned} f_{U|V}(u|v) = \frac{f_{UV}(u,v)}{f_V(v)} &= \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(u^2+v^2-2\rho uv)\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{v^2}{2}\right)} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(u^2+v^2-2\pi uv-(1-\rho^2)v^2)\right) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{(u-\rho v)^2}{2(1-\rho^2)}\right). \end{aligned}$$

Thus,

$$U|V = v \sim \mathcal{N}(\rho v, (1-\rho^2))$$

and similarly

$$V|U = u \sim \mathcal{N}(\rho u, (1 - \rho^2)).$$

Thus the conditional expectation functions are

$$U = \rho V$$

and

$$V = \rho U.$$

Remark 5.1. This demonstrates regression toward the mean, since $\rho \leq 1$.

5.3 Bivariate Normal, the General Case

We may write

$$\begin{cases} X = \mu_X + \sigma_X U \sim \mathcal{N}(\mu_x, \sigma_X^2) \\ Y = \mu_Y + \sigma_Y V \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \end{cases}.$$

We have

$$\text{Cov}(X, Y) = \text{Cov}(\mu_X + \sigma_X U, \mu_Y + \sigma_Y V) = \sigma_X \sigma_Y \text{Cov}(U, V)$$

and thus

$$\text{Corr}(X, Y) = \rho.$$

From

$$\frac{X - \mu_X}{\sigma_X} | \frac{Y - \mu_Y}{\sigma_Y} \sim \mathcal{N}\left(\rho \cdot \frac{y - \mu_Y}{\sigma_Y}, 1 - \rho^2\right)$$

we know

$$X|Y = y \sim \mathcal{N}\left(\mu_X + \rho \frac{\sigma_X}{\sigma_Y}(y - \mu_Y), \sigma_X^2(1 - \rho^2)\right).$$

Similarly,

$$Y|X = x \sim \mathcal{N}\left(\mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X), \sigma_Y^2(1 - \rho^2)\right).$$

Note that the variance is smaller.

5.4 Bivariate Normal Data

Suppose (x_i, y_i) are iid bivariate normal with

$$(x_i, y_i) \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{\sigma_1 \sigma_2} \\ \rho_{\sigma_1 \sigma_2} & \sigma_2^2 \end{bmatrix}\right).$$

Then $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and $\text{Cov}(X_i, Y_i) = \rho$.

5.5 MLE

We have

$$\begin{aligned}
&L(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) \\
&= \prod\prod f(x_i, y_i) \\
&= \frac{1}{(2\pi)^2(1-\rho^2)^{n/2}} \\
&\quad \exp\left(\frac{1}{2(1-\rho^2)} \sum \left(\left(\frac{x_i-\mu_X}{\sigma_X}\right)^2 + \left(\frac{x_i-\mu_X}{\sigma_X}\right)^2 - 2\pi\left(\frac{x_i-\mu_X}{\sigma_X}\right)\left(\frac{y_i-\mu_Y}{\sigma_Y}\right)\right)\right).
\end{aligned}$$

We have[4]

$$\hat{\mu}_X = \overline{X}, \quad \hat{\mu}_X = \overline{Y}, \quad \hat{\sigma}_X = n^{-1}\sum (X_i - \overline{X})^2, \quad \hat{\sigma}_Y = n^{-1}\sum (Y_i - \overline{Y})^2$$

and

$$R \coloneqq \hat{\rho} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}.$$

Definition 5.2. R is called the **Pearson correlation coefficient**.

5.6 Distribution of the Pearson Correlation Coefficient

It is easy to see that $\operatorname{supp} R = [-1, 1]$, but its distribution is quite complicated. We know however that

$$\mathrm{E}[\hat{\rho}] = \rho - \frac{\rho(1-\rho^2)}{2n} + \dots$$

and

$$\operatorname{Var}[\hat{\rho}] = \frac{(1-\rho^2)^2}{n} + \dots.$$

We have approximately that

$$\operatorname{Var}[\hat{\rho}] \propto (1-\rho^2)^2 = v(\rho).$$

By Fisher we know that for large n we have

$$\hat{\rho} \sim \mathcal{N}\left(\rho, \frac{(1-\rho^2)^2}{n}\right).$$

To implement VST, we seek g such that $g'(\rho) = 1/(1-\rho^2)$. A choice is

$$g(\rho) = \int \frac{1}{(1+\rho)(1-\rho)}\, \mathrm{d}\rho = \frac{1}{2}\int \frac{1}{1+\rho} + \frac{1}{1-\rho}\, \mathrm{d}\rho = \frac{1}{2}\log\frac{1+\rho}{1-\rho}.$$

[4] These are *not* unbiased estimators. For the unbiased estimators, we need to divide by $n-1$ instead of n.

This is called the **Fisher transformation**. Note that it is monotone (and thus the construction of CIs is somewhat easy).

We have that

$$\frac{1}{2}\log\frac{1+\hat{\rho}}{1-\hat{\rho}} \sim \mathcal{N}\left(\frac{1}{2}\log\frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right).$$

It turns out that $1/(n-3)$ is better than $1/n$ for $n > 5$.

Using this we can construct a CI for $g(\rho)$

$$\left[\frac{1}{2}\log\frac{1+\hat{\rho}}{1-\hat{\rho}} - z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}, \frac{1}{2}\log\frac{1+\hat{\rho}}{1-\hat{\rho}} + z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}\right] = [L, U]$$

and for then ρ: Note that

$$T = \frac{1}{2}\log\frac{1+R}{1-R} \iff e^{2T} = \frac{1+R}{1-R} \iff R = \frac{e^{2T}-1}{e^{2T}+1} = \tanh T.$$

So a CI for ρ is

$$\left[\frac{e^{2L}-1}{e^{2L}+1}, \frac{e^{2U}-1}{e^{2U}+1}\right].$$

Remark 5.3. This is the standard way to construct a CI for the correlation coefficient in software packages, regardless of the distribution of the original data. In our derivation, however, we assumed bivariate normality. For small n, be mindful of the many layers of approximation!

5.6.1 The Matrix Notation

For the standard bivariate normal, with

$$\begin{bmatrix} U \\ V \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right),$$

we may write $[U, V]^\intercal \sim \mathcal{N}([0, 0]^\intercal, \Sigma)$, and then

$$f(u, v) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}}\exp\left(\frac{1}{2}[uv]^\intercal\Sigma^{-1}[uv]\right),$$

where

$$\Sigma^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}.$$

For the more general case, with $X \sim \mathcal{N}(\mu, \Sigma)$, we have

$$f(x) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}}\exp\left(\frac{1}{2}[x-\mu]^\intercal\Sigma^{-1}[x-\mu]\right),$$

5.6.2 Properties

Proposition 5.4. *For bivariate normally distributed variables, we have the variables are uncorrelated if and only if they are independence.*

Proposition 5.5. *If $X \sim \mathcal{N}_2(\mu, \Sigma)$ if and only if $aX + bY$ is of (univariate) normal distribution for any $a, b \in \mathbb{R}$.*

Example 5.6 (Non-example). Let $X \sim \mathcal{N}(0,1)$ and

$$Y = \begin{cases} X & \text{with probability } 1/2 \\ -X & \text{with probability } 1/2 \end{cases}.$$

Then $Y \sim \mathcal{N}(0,1)$, but $[XY]^\intercal$ is not of bivariate normal distribution. A way to see this is that the linear combination $X + Y$ has point mass at 0 with probability $1/2$.

Proposition 5.7. *If $Z_1, Z_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, then*

$$\begin{bmatrix} X \\ Y \end{bmatrix} = A \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} a_{11}Z_1 + a_{12}Z_2 \\ a_{21}Z_1 + a_{22}Z_2 \end{bmatrix} \sim \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, AA^\intercal \right)$$

is bivariate normal.

Proposition 5.8. *If $\left[X_1 X_2\right]^\intercal \sim \mathcal{N}_2(\mu, \Sigma)$ and $[Y_1 Y_2]^\intercal = A[X_1 X_2]^\intercal$ is a linear transformation of $[X_1 X_2]^\intercal$ such that A^{-1} exists, then $[Y_1 Y_2]^\intercal \sim \mathcal{N}(A\mu, A\Sigma A')$. Similarly, $A[X_1 X_2]^\intercal + [b_1 b_2]^\intercal \sim \mathcal{N}_2([b_1 b_2]^\intercal + A\mu, A\Sigma A')$.*

5.7 Theorem of Sampling Distributions

5.7.1 The χ and t Distributions

χ^2 distribution. Let $Z_1, \ldots, Z_k \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Then

$$Z_1^2 + \cdots + Z_n^2 \sim \chi_k^2.$$

It turns out that $\chi_k^2 \sim \text{Gamma}(k/2, 1/2)$. If $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi_n^2$ with Z and W independent, then

$$\frac{Z}{\sqrt{W/k}} \sim t_k.$$

We have

$$f_{t_k}(t) = C_k \frac{1}{(1 + t^2/k)^{\frac{k+1}{2}}},$$

where C_k is a constant.

Remark 5.9.

- Note that as $k \to \infty$,
$$(1 + t^2/k)^{\frac{k+1}{2}} \longrightarrow e^{t^2/2}.$$

- $t_1 = Z_1/Z_2$ with $Z_1, Z_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ is also called the **Cauchy distribution**. It has density

$$f_1(t) = \frac{1}{\pi} \cdot \frac{1}{1+t^2}.$$

Note however that it does not have an expected value, since

$$\frac{1}{\pi} \int_0^\infty \frac{t}{1+t^2} \, \mathrm{d}t \longrightarrow \infty,$$

and similarly the left side also diverges. The variance also does not exist. The law of large number does not apply!

Theorem 5.10. *Suppose $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma)$. Then we have*

(i) $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.

(ii) $\frac{1}{\sigma^2} \sum (X_i - \overline{X}_n)^2 \sim \chi^2_{n-1}$.

(iii) $\sum (X_i - \overline{X}_n)^2 \perp\!\!\!\perp \overline{X}_n$.

(iv) $\frac{\overline{X}_n - \mu}{\sqrt{S^2/n}} \sim t_{n-1}$, *where* $S^2 = \frac{1}{n-1} \sum X_i$ *is the sample variance.*

These are exact distributions, not approximations.

Proof.

(i) Follows from linearity of the normal distribution.

(ii) & (iii) We prove by induction. Let $Z_i \coloneqq (X_i - \mu)/\sigma$. Then $\overline{Z}_n = n^{-1} \sum Z_i = (\overline{X} - \mu)/\sigma$. Thus we have

$$\frac{1}{\sigma^2} \sum (X_i - \overline{X}_n)^2 = \sum \left(\frac{X_i - \overline{X}_n}{\sigma} \right)^2 = \sum (Z_i - \overline{Z})^2.$$

We need thus only show that $\sum (Z_i - \overline{Z})^2 \sim \chi^2_{n-1}$ and is independent of $\overline{Z}_n$.

For $n = 2$, we have

$$(Z_1 - \overline{Z}_2)^2 + (Z_2 - \overline{Z}_2)^2 = \left(Z_1 - \frac{Z_1 + Z_2}{2} \right)^2 + \left(Z_2 - \frac{Z_1 + Z_2}{2} \right)^2 = \frac{(Z_1 - Z_2)^2}{2}.$$

Recalling that $(Z_1 - Z_2)/\sqrt{2} \sim \mathcal{N}(0,1)$, we have $(Z_1 - Z_2)^2/2 \sim \chi^2_1$.

Next, recall that for bivariate normal (X, Y) we have that $X \perp\!\!\!\perp Y$ if and only if $\rho = 0$. Thus, noting that $[Z_1 - Z_2 \; Z_1 + Z_2]^\intercal$ is bivariate normal and

$$\mathrm{Cov}(Z_1 - Z_2, Z_1 + Z_2) = \mathrm{Var}(Z_1) - \mathrm{Var}(Z_2) = 0,$$

we know that $Z_1 - Z_2 \perp\!\!\!\perp Z_1 + Z_2$ and thus $(Z_1 + Z_2)^2/2 \perp\!\!\!\perp (Z_1 + Z_2)/2 = \overline{Z}_2$.

For the general case, suppose that $\sum_{i=1}^{n}(Z_i - \overline{Z})n)^2$ is of χ^2_{n-1} and is independent of $\overline{Z}_n$. Note that

$$\begin{aligned}\sum_{i=1}^{n}(Z_i - \overline{Z}_{n+1})^2 &= \sum_{i=1}^{n}(Z_i - \overline{Z}_n + \overline{Z}_n - \overline{Z}_{n+1})^2 + (Z_{n+1} - \overline{Z}_{n+1}) \\ &= \sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2 + 2\sum_{i=1}^{n}(Z_1 - \overline{Z}_n)(\overline{Z}_n - \overline{Z}_{n+1}) \\ &\quad + \sum_{i=1}^{n}(\overline{Z}_n - \overline{Z}_{n+1})^2 + (Z_{n+1} - \overline{Z}_{n+1})^2.\end{aligned}$$

Note that the first term is χ^2_{n-1} and the second term is 0. Thus we need only show that

$$J \coloneqq n(\overline{Z}_n - \overline{Z}_{n-1})^2 + (Z_{n+1} - \overline{Z}_{n+1})^2$$

is of χ^2_1 distribution and is independent of the first term.

Observe now that

$$\overline{Z}_{n+1} = \frac{n}{n+1}\overline{Z}_n + \frac{Z_{n+1}}{n+1},$$

and thus

$$\begin{aligned}J &= n\left(\overline{Z}_n - \frac{n}{n+1}\overline{Z}_n - \frac{1}{n+1}Z_{n+1}\right)^2 + \left(Z_{n+1} - \frac{n}{n+1}\overline{Z}_n - \frac{1}{n+1}Z_{n+1}\right) \\ &= \frac{n}{(n+1)^2}(\overline{Z}_n - Z_{n+1})^2 + \frac{n^2}{(n+1)^2}(Z_{n+1} - \overline{Z}_n)^2 \\ &= \frac{n+n^2}{(n+1)^2}(\overline{Z}_n - Z_{n-1})^2 = \frac{n}{n+1}(\overline{Z}_n - Z_{n+1})^2.\end{aligned}$$

Since $\overline{Z} - Z_{n+1} \sim \mathcal{N}(0, 1/n+1)$, we know that J is of χ^2_1 distribution. For claim (ii), it remains to show that $J = n/(n+1)\cdot(\overline{Z}_n - Z_{n+1})^2$ is independent to $\sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2$. Since $Z_{n+1}, \overline{Z}_n$, and $\sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2$ are pairwise independent, we obtain (ii). To complete the proof, note that a similar argument as before shows (iii).

(iv) We have

$$\frac{\overline{X}_n - \mu}{\sqrt{S^2/n}} = \frac{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\sum(X_i - \overline{X})^2}{\sigma^2}/(n-1)}} = \frac{Z}{\sqrt{W/n-1}},$$

where $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_n$.

□

5.8 Two-Sample

Suppose $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$ and we want to estimate $\mu_1 - \mu_2$. A choice of an estimator is $\hat{\mu}_1 - \hat{\mu}_2 = \overline{X} - \overline{Y}$, but how can we obtain an CI?

Case 1: A "paired" sample. Suppose $\text{Cov}(X_1, Y_1) > 0$. We can use

$$\text{Var}(\overline{X} - \overline{Y}) = \text{Var}(\overline{X}) + \text{Var}(\overline{Y}) - 2\,\text{Cov}(\overline{X}, \overline{Y}) < \text{Var}(\overline{X}) + \text{Var}(\overline{Y})$$

to obtain a narrower CI.

Define $D_i \coloneqq X_i - Y_i$ to reduce the problem to that of a one sample problem. We have in particular that

$$D_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right),$$

where $\rho \coloneqq \text{Corr}(X_i, Y_i)$.

If σ is unknown, we may use $\hat{\text{Var}}[\overline{D}] = S_D^2/n$ to obtain

$$\frac{\overline{D}_i - (\mu_1 - \mu_2)}{\sqrt{S_D^2/n}} \sim t_{n-1}.$$

Case 2: $\text{Cov}(X_i, Y_i) = 0$ and $\sigma_1 = \sigma_2$. Note that the normality assumption gives $X_i \perp\!\!\!\perp Y_i$. Then, $\text{Var}[\overline{X} - \overline{Y}] = \text{Var}[\overline{X}] + \text{Var}[\overline{Y}] = \sigma^2/n + \sigma^2/m$. We may use the sample variances to estimate σ^2:

$$S_X^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2, \quad S_Y^2 = \frac{1}{m-1}\sum (Y_i - \overline{Y})^2,$$
$$S_{\text{pool}}^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

We use the last term for higher power and narrower CI. We use

$$\hat{\text{Var}}[\overline{X} - \overline{Y}] = S_{\text{pool}}^2\left(\frac{1}{n} + \frac{1}{m}\right)$$

to obtain

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_{\text{pool}}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

and

$$\frac{S_{\text{pool}^2}}{\sigma^2} \sim \chi^2_{n+m-2}.$$

Case 2: $X_i \perp\!\!\!\perp Y_i$ and $\sigma_1 \neq \sigma_2$. We have $\text{Var}[\overline{X} - \overline{Y}] = \sigma_1^2/n + \sigma_2^2/m$. Then, approximately, we have

$$T \coloneqq \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim \mathcal{N}(0, 1).$$

It turns out that for moderate sized n and m, a better approximation, called the **Welch–Satterthwaite** approximation is

$$T \sim t_\nu,$$

where ν is a function of sample variances and sample sizes. This is what is implemented in software packages.

5.8.1 Hypothesis Testing for Two-Sample

We define the likelihood ratio as

$$LR := \frac{\max_{H_0} \text{likelihood function}}{\max_{H_0 \cup H_a} \text{likelihood function}} \in (0, 1].$$

Note that is H_0 is true, LR is likely to be large (close to 1).

Proposition 5.11. *Under H_0, we have the following approximate asymptotic distribution of LR:*

$$-2 \log LR \approx \chi^2_d,$$

where $d = \dim(H_a \cup H_0) - \dim H_0$.

Example 5.12. Suppose $X_1, \dots, X_n \sim \mathcal{N}(\mu_1, \sigma^2)$ and $Y_1, \dots, Y_m \sim \mathcal{N}(\mu_2, \sigma^2)$. Consider the hypothesis $H_0 : \mu_1 = \mu_2$ and $H_a : \mu_1 \neq \mu_2$. Under H_0 we have $\mu_1 = \mu_2 = \mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}$, so $\dim H_0 = 2$. Similarly we have $\dim(H_a \cup H_0) = 3$. Thus $d = 1$.

We have

$$\begin{aligned} L(\mu_1, \mu_2, \sigma^2) &= \prod f_X(x_i) f_Y(y_i) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n+m}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - \mu_1)^2 - \frac{1}{2\sigma^2} \sum (y_i - \mu_2)^2 \right). \end{aligned}$$

The log-likelihood function is

$$\ell(\mu_1, \mu_2, \sigma^2) = -\frac{n+m}{2} \log(2\pi) - \frac{n+m}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum (x_i - \mu_1)^2 - \frac{1}{2\sigma^2} \sum (y_i - \mu_2)^2.$$

Under H_0 we have $\mu_1 = \mu_2 = \mu_0$, and for large n,

$$-2 \log LR \approx \chi^2_1.$$

Further, under H_0,

$$\frac{\partial \log L}{\partial \mu_0} = -\frac{1}{\sigma^2} \sum (x_i - \mu_0) - \frac{1}{\sigma^2} \sum (y_i - \mu_0).$$

Setting $\partial \log L / \partial \mu_0 = 0$, we get

$$\hat{\mu}_0 = \frac{\sum x_i + \sum y_i}{n + m}$$

for any σ^2. Thus

$$\max_{H_0} L(\mu_1, \mu_2, \sigma^2) = L(\hat{\mu}_0).$$

Next, we have

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n+m}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum (x_i - \mu_1)^2 + \frac{1}{2(\sigma^2)^2}\sum (y_i - \mu_2)^2.$$

Setting $\partial \log L/\partial \sigma^2 = 0$ at $\mu_1 = \mu_2 = \hat{\mu}_0$, we get

$$\hat{\sigma}_0^2 = \frac{1}{n+m}\left(\sum (x_i - \hat{\mu}_0)^2 + \sum (y_i - \hat{\mu}_0)^2\right).$$

We have then that

$$\begin{aligned}
\max_{H_0} L(\mu_1, \mu_2, \sigma^2) &= L(\hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0^2) \\
&= \left(\frac{1}{\sqrt{\pi\hat{\sigma}_0^2}}\right)^{n+m} \exp\left(-\frac{\sum (x_i - \hat{\mu}_0)^2 + \sum (y_j - \hat{\mu}_0)^2}{2\hat{\sigma}_0^2}\right) \\
&= \left(\frac{1}{\sqrt{\pi\hat{\sigma}_0^2}}\right)^{n+m} \exp\left(-\frac{n+m}{2}\right).
\end{aligned}$$

Under $H_0 \cup H_a$: we have

$$\frac{\partial \log L}{\partial \mu_1} = \frac{1}{\sigma^2}\sum (x_i - \mu_1), \quad \frac{\partial \log L}{\partial \mu_2} = \frac{1}{\sigma^2}\sum (y_i - \mu_2).$$

Setting the above terms to zero gives

$$\hat{\mu}_1 = \overline{X}, \quad \hat{\mu}_2 = \overline{Y}.$$

Again, setting $\partial \log L/\partial \sigma^2 = 0$ at $\mu_1 = \hat{\mu}_1$ and $\mu_2 = \hat{\mu}_2$, we get

$$\hat{\sigma}^2 = \frac{1}{n+m}\left(\sum (x_i - \hat{\mu}_1)^2 + \sum (y_i - \hat{\mu}_2)^2\right).$$

We have then that

$$\begin{aligned}
\max_{H_0 \cup H_a} L(\mu_1, \mu_2, \sigma^2) &= L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2) \\
&= \left(\frac{1}{\sqrt{\pi\hat{\sigma}^2}}\right)^{n+m} \exp\left(-\frac{\sum (x_i - \hat{\mu}_1)^2 + \sum (y_j - \hat{\mu}_2)^2}{2\hat{\sigma}^2}\right) \\
&= \left(\frac{1}{\sqrt{\pi\hat{\sigma}^2}}\right)^{n+m} \exp\left(-\frac{n+m}{2}\right).
\end{aligned}$$

Then,

$$LR = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{\frac{n+m}{2}} = \left(\frac{\sum (x_i - \overline{x})^2 + \sum (y_i - \overline{y})^2}{\sum (x_i - \hat{\mu}_0)^2 + \sum (y_i - \hat{\mu}_0)^2}\right)^{\frac{n+m}{2}}.$$

5.9 ANOVA (Analysis of Variance)

For $i = 1, \ldots, g$ with $g > 2$ suppose $Y_{i1}, \ldots, Y_{in} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$. Consider $H_0 : \mu_1 = \cdots = \mu_g$ and H_a: $\mu_i \neq \mu_j$ for some $i \neq j$. (Consider for example the outcomes after different treatment levels i.)

Note that there are two types of variation: within group and across group. Denote the mean of the whole sample as $\overline{y} = y_{..}$. We may decompose the variance accordingly:

$$\begin{aligned}\sum_{i=1}^{g}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}\right)^2 &= \sum_{i=1}^{g}\sum_{j=1}^{n}\left[\left(y_{ij}-\overline{y}_i\right)-\left(\overline{y}_i-\overline{y}\right)\right]^2 \\ &= \sum_{i=1}^{g}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}_i\right)^2+\sum_{i=1}^{g} n_i\left(\overline{y}_i-\overline{y}\right)^2,\end{aligned}$$

where the first term is within group variation (the noise), and the second term is the across group variation (the treatment effect, the signal). Equivalently, we write this decomposition as

$$SS_{\text{total}} = SS_{\text{residual}} + SS_{\text{treatment}},$$

where SS stands for sum of squares.

5.9.1 ANOVA Table

Source	SS	df	MS	Var Ratio (F)
Treatment	SS_{trt}	$g-1$	$SS_{\text{trt}}/(g-1)$	$MS_{\text{trt}}/MS_{\text{e}}$
Residual (error)	SS_{e}	$n-g$	$SS_{\text{e}}/(n-g)$	
Total	SS_{total}	$n-1$		

We have under H_0 that

$$\frac{MS_{\text{trt}}}{MS_e} \sim F_{g-1,n-g}.$$

5.9.2 F distribution

If $W_1 \sim \chi^2_{k_1}$ and $W_2 \sim \chi^2_{k_2}$ with $W_1 \perp\!\!\!\perp W_2$, then

$$\frac{W_1/k_1}{W_2/k_2} \sim F_{k_1,k_2}.$$

6 Linear Models

We may rewrite the assumption as: $Y_{ij} = \mu_i + \epsilon_{ij}$ with $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. (Alternatively, this may be viewed as an assumption of the data generation process.)

Proposition 6.1. *The MLE of μ_i is $\overline{y}_i$.*

Another way to parameterize the model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, g$ and $j = 1, \ldots, n_i$. We impose constraint on the parameters of the form $\sum \alpha_i = 0$, or $\alpha_0 = 0$, or $\alpha_g = 0$. These are constraints that we can freely impose without loss of generality; they may be viewed as a way to define μ.

Example 6.2. Use $\partial \log L / \partial \alpha_i = 0$ for $i = 0, \ldots, g-1$ and use constraint $\sum \alpha_i = 0$ to derive MLE for α_i and μ. It turns out that $\hat{\mu} = \sum \overline{y}_i / g$.