CSE446/546 Notes on Gaussian Confidence Intervals

Kevin Jamieson, University of Washington

1 Gaussian random variables

A Gaussian random vector X with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$ has probability density function

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-(x-\mu)\Sigma^{-1}(x-\mu)/2}$$

where for any square matrix A, $|A|$ denotes the determinant of A. For convenience, we will denote a draw from this distribution as $X \sim \mathcal{N}(\mu, \Sigma)$.

Proposition 1. *Fix $A \in \mathbb{R}^{p\times d}$ and $b \in \mathbb{R}^p$. If $X \sim \mathcal{N}(\mu, \Sigma)$ then $AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top)$.*

Proposition 1 allows us to transform a standard Gaussian vector, i.e. $Z \sim \mathcal{N}(0, I_d)$, into an arbitrary p-dimensional Gaussian vector with arbitrary mean and covariance for any $p \leq d$. It also illuminates some properties of empirical means of Gaussian random variables.

Example 1. *Let $X_1, \dots, X_n \in \mathbb{R}$ be IID with $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Note that if $\mathbf{1}$ is a vector of all ones, then the vector $X \sim \mathcal{N}(\mu\mathbf{1}, \sigma^2 I_n)$. Thus, we can choose $A = (\frac{1}{n}, \dots, \frac{1}{n}) \in \mathbb{R}^{1\times n}$ and $b = 0$ to conclude that $\frac{1}{n}\sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \sigma^2/n)$.*

Because Proposition 1 gives an explicit transformation of arbitrary Gaussian random variables, many of the properties of a Gaussian are understood through a standard Gaussian $Z \sim \mathcal{N}(0, 1)$ with PDF $p(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. Indeed, some elemantary calculations gives us the following result:

Proposition 2. *For any $t \geq 0$ we have*

$$\frac{1}{\sqrt{2\pi}}\frac{t}{t^2+1}e^{-t^2/2} \leq \int_{x=t}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \leq \min\{1, \frac{1}{\sqrt{2\pi}}\frac{1}{t}\}e^{-t^2/2}. \tag{1}$$

In what follows we're only going to use the upper bound on the right hand side, but the lower bound on the left hand side provides us with confidence that this upper bound is indeed tight, and much easier to use than the integral expression itself.

We can leverage the above proposition to build a confidence interval on a Gaussian random variables. Indeed, suppose that $X \sim \mathcal{N}(\mu, \sigma^2)$. By an appropriate application of Proposition 1 we have that $(X-\mu)/\sigma \sim \mathcal{N}(0, 1)$. This means that

$$\begin{aligned} \mathbb{P}(X \geq \mu + \sigma t) &= \mathbb{P}((X-\mu)/\sigma \geq t) \\ &= \int_{x=t}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx. \end{aligned}$$

For some $\delta \in (0, 1)$, if we set $t = \sqrt{2\log(2/\delta)}$ then the right hand side of Proposition 2 is upper bounded by $\delta/2$. We conclude that $\mathbb{P}(X \geq \mu + \sqrt{2\sigma^2\log(2/\delta)}) \leq \delta/2$. By noticing that $-(X-\mu)/\sigma \sim \mathcal{N}(0, 1)$ as well, we can run the same calculations to conclude that $\mathbb{P}(X \leq \mu - \sqrt{2\sigma^2\log(2/\delta)}) \leq \delta/2$. From this we can construct a confidence interval:

$$\begin{aligned} \mathbb{P}(|X-\mu| > \sqrt{2\sigma^2\log(2/\delta)}) &\leq \mathbb{P}(X \geq \mu + \sqrt{2\sigma^2\log(2/\delta)}) + \mathbb{P}(X \leq \mu - \sqrt{2\sigma^2\log(2/\delta)}) \\ &\leq \delta/2 + \delta/2 \leq \delta. \end{aligned}$$

Stated another way, with probability at least $1-\delta$ we have that X will be contained within the confidence interval $[\mu - \sqrt{2\sigma^2\log(2/\delta)}, \mu + \sqrt{2\sigma^2\log(2/\delta)}]$. The next example puts this to work for empirical means.

Example 2. *Let $X_1, \dots, X_n \in \mathbb{R}$ be IID with $X_i \sim \mathcal{N}(\mu, \sigma^2)$. By example 1, if $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ then $\widehat{\mu}_n \sim \mathcal{N}(\mu, \sigma^2/n)$. By an appropriate application of Proposition 1 we have that $(\widehat{\mu}_n - \mu)\sqrt{n}/\sigma \sim \mathcal{N}(0, 1)$. Thus, with probability at least $1-\delta$ we have that $\widehat{\mu}_n$ will be contained within the confidence interval $[\mu - \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}}, \mu + \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}}]$.*

2 Linear regression with Gaussian noise

We now consider linear regression. Suppose we observed $\{(x_i, y_i)\}_{i=1}^n$ where each pair is drawn IID from an unknown distribution, but it is know that there exists some $\theta_* \in \mathbb{R}^d$ such that $y_i = \langle x_i, \theta_* \rangle + \eta_i$ where $\eta_i \sim \mathcal{N}(0, \sigma^2)$. As discussed in class, consider the least squares estimator (equivalent to the maximum likelihood estimator)

$$\begin{aligned}
\widehat{\theta} &= \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n (y_i - \langle x_i, \theta \rangle)^2 \\
&= \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1} \sum_{i=1}^n x_i y_i \\
&= \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1} \sum_{i=1}^n x_i (x_i^\top \theta_* + \eta_i) \\
&= \theta_* + \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1} \sum_{i=1}^n x_i \eta_i
\end{aligned}$$

where the second inequality has assumed that $\left(\sum_{i=1}^n x_i x_i^\top \right)^{-1}$ exists, and the third inequality plugs in our assumed model $y_i = \langle x_i, \theta_* \rangle + \eta_i$. If we let $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times d}$ and $\eta = (\eta_1, \dots, \eta_n)^\top$ then we may write $\widehat{\theta} = \theta_* + (X^\top X)^{-1} X^\top \eta$. Applying Proposition 1 with $A = (X^\top X)^{-1} X^\top$ and $b = \theta_*$ we observe that $\widehat{\theta} \sim \mathcal{N}(\theta_*, \sigma^2 (X^\top X)^{-1})$ using the fact that $(X^\top X)^{-1} X^\top \cdot \sigma^2 I_n \cdot X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1}$. For any $z \in \mathbb{R}^d$ we may apply Proposition 1 yet again to observe that $z^\top (\widehat{\theta} - \theta_*) \sim \mathcal{N}(0, \sigma^2 z^\top (X^\top X)^{-1} z)$ and Proposition 2 to conclude that

$$\mathbb{P}\left(|z^\top (\hat{\theta} - \theta_*)| \geq \sqrt{2\sigma^2 z^\top (X^\top X)^{-1} z \log(2/\delta)} \right) \leq \delta. \tag{2}$$

If we fix $i \in \{1, \dots, d\}$ and set $z = \mathbf{e}_i$ where $\mathbf{e}_i$ denotes a vector of all zeros except for a 1 in the ith position, then Equation 2 says $\mathbb{P}(|\widehat{\theta}_i - \theta_{*,i}| > \sqrt{2\sigma^2 [(X^\top X)^{-1}]_{i,i} \log(2/\delta)}) \leq \delta$. However, this only holds for a *single* component $i \in \{1, \dots, d\}$. To account for spurious noise, we may take a union bound over all $i \in \{1, \dots, d\}$:

$$\begin{aligned}
\mathbb{P}(\bigcup_{i=1}^d \{|\widehat{\theta}_i - \theta_{*,i}| > \sqrt{2\sigma^2 [(X^\top X)^{-1}]_{i,i} \log(2d/\delta)}\}) &\leq \sum_{i=1}^d \mathbb{P}(|\widehat{\theta}_i - \theta_{*,i}| > \sqrt{2\sigma^2 [(X^\top X)^{-1}]_{i,i} \log(2d/\delta)}) \\
&\leq \sum_{i=1}^d \frac{\delta}{d} = \delta.
\end{aligned}$$

We conclude with the observation that with probability at least $1 - \delta$, $\widehat{\theta}_i > \sqrt{2\sigma^2 [(X^\top X)^{-1}]_{i,i} \log(2d/\delta)}$ implies that $\theta_{*,i} > 0$ for all i (and an analogous conclusion for $\theta_{*,i} < 0$). This provides a rigorous way of justifying whether the coefficients of $\widehat{\theta}$ imply "real" correlation with the observed phenomenon.