6.S951 Modern Mathematical Statistics Fall 2024

Lecture 12 — October 22, 2024

Prof. Stephen Bates

Scribe: Nicolas Emmenegger

1 Outline

Agenda:

1. "Duality": Hypothesis testing $\leftrightarrow$ CIs. Paving the way towards conformal prediction.
2. Example: Binomial CIs with improvements over CLT-CIs.
3. Introduction to p-values.

Recap – Statistical Inference:

1. CIs from CLT or Hoeffding's inequality.
2. Confidence bands for CDFs, DKW Inequality.
3. Bootstrap

2 Confidence Interval and Hypothesis Testing duality

We will connect hypothesis tests & CIs and get more clarity about CIs. As a byproduct, we will get a recipe for generating confidence intervals from a given collection of tests, which will turn out to be useful in our treatment of conformal prediction. The setting is as usual: Θ is our parameter space, and $\mathcal{X}$ is the sample space. We observe data $X \in \mathcal{X}$ for P_θ for some true parameter $\theta \in \Theta$. Recall that a confidence interval is mapping data to a subset of the parameter space, i.e. $C : \mathcal{X} \to 2^\Theta$.

Idea In a specific sense that we make clear below, a confidence interval is equivalent to a *collection* of tests $\{\phi_{\tilde{\theta}} : \mathcal{X} \to \{0,1\} \mid \tilde{\theta} \in \Theta\}$. Such a collection of tests gives rise to a confidence interval, and reversely, any confidence interval gives us a collection of tests.

Recall our definitions of confidence intervals:

Definition 1. *A confidence interval $C : \mathcal{X} \to 2^\Theta$ is level α if*

$$\mathbb{P}_\theta(\theta \in C(X)) \geq 1 - \alpha \quad \text{for all} \quad \theta \in \Theta. \tag{1}$$

That is, the confidence interval includes ("covers") the true parameter with high probability. Next, recall our definition of type-I error in a hypothesis test.

Definition 2. *We call a test* $\phi_{\tilde{\theta}} : \mathcal{X} \to \{0, 1\}$ *level-*α *whenever*

$$\mathbb{P}_{\tilde{\theta}}(\phi_{\tilde{\theta}}(X) = 1) \leq \alpha. \tag{2}$$

This corresponds to α-bounded Type-I error when testing the point null $\Theta_0 = \{\tilde{\theta}\}$.

We can now prove both directions of the equivalence.

Lemma 3. *Suppose* C *is a CI that satisfies* (1)*. Fix* $\tilde{\theta} \in \Theta$ *and define the test*

$$\phi_{\tilde{\theta}}(X) = \begin{cases} 1 & \tilde{\theta} \notin C(X) \\ 0 & \tilde{\theta} \in C(X). \end{cases}$$

Then $\phi_{\tilde{\theta}}$ *is level-*α*.*

Proof.

$$\mathbb{P}_{\tilde{\theta}}(\phi_{\tilde{\theta}}(X) = 1) = \mathbb{P}_{\tilde{\theta}}(\tilde{\theta} \notin C(X)) \overset{(1)}{\leq} \alpha.$$

□

Hence, a CI gives rise to a collection of tests, as claimed. For the other side, we formalize as follows:

Lemma 4. *Take a collection of tests* $\{\phi_{\tilde{\theta}} : \mathcal{X} \to \{0, 1\} \mid \tilde{\theta} \in \Theta\}$ *that each satisfy* (2)*. Construct*

$$C(X) = \{\tilde{\theta} \in \Theta \mid \phi_{\tilde{\theta}}(X) = 0\}.$$

Then, C *is a level-*α *CI.*

Proof. For any fixed $\theta \in \Theta$

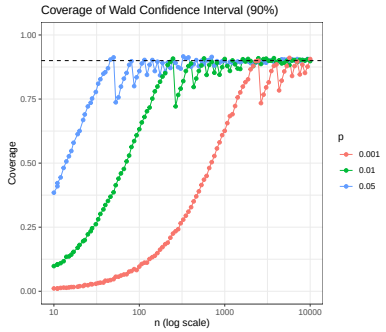
$$\mathbb{P}_{\theta}(\theta \in C(X)) = \mathbb{P}_{\theta}(\phi_{\theta}(X) = 0) \overset{(2)}{\geq} 1 - \alpha.$$

□

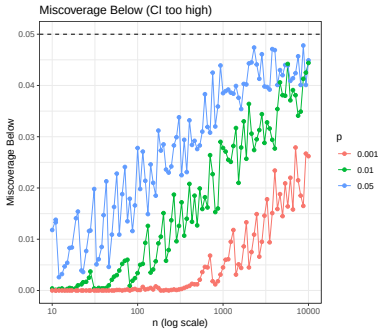
Intuition To build a CI based on a collection of tests, we collect all the parameters that are consistent enough with the data, or more precisely, consistent enough with the data such that the corresponding test does not reject the point null.

3 Binomial CIs

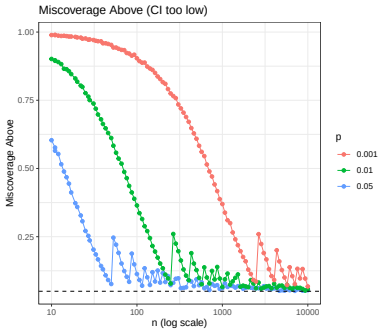
We investigate the specific class of models $X \sim Binom(n, \theta)$, where $\theta \in [0, 1] = \Theta$. We also define the sample average as $\bar{X} = X/n$.



(a) Observe that for small θ we need more samples to converge to the desired coverage. Also note that the sawtooth pattern is not due to simulation error, but due to the discrete nature of the Binomial Distribution.



(b) Miscoverage below happens comparatively rarely especially when n is small. Only asymptotically do we incur an $\alpha/2 = 0.05$ portion of miscoverage from cases where the true θ is below the CI. For very small $\theta = 0.001$ we see that it takes a lot of samples until we start making any errors.



(c) On the other hand, if we plot the miscoverage above, we see that especially in small sample regimes, we incur a lot of error (this makes sense, since the coverage in (a) is poor, which is almost entirely due to cases where the CI is too low.)

Figure 1: Coverage of CLT Intervals with $\theta = p$.

3.1 Asymptotic CIs

We know how to build asymptotically valid CIs by way of the CLT. The CI

$$C^{Wald}(X) = \left\{\theta \in [0,1] \mid \bar{X} - 1.64\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq \theta \leq \bar{X} + 1.64\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right\}$$

is asympotically of level-α for $\alpha = 0.1$. Such intervals are called Wald intervals. These intervals can be poor for some choices of n and θ, as we observe in the miscoverage plots in Figure 1.

Since the plotted values of θ are small, most of the miscoverage happens above, i.e. the true θ lies to the right of our confidence interval. This is exacerbated when n isn't too large, and when θ is small, because then the sample mean can be zero, and then $C^{Wald}(X) = \{0\}$.

Remark 5. *The contrast between Figure 1 (b) and (c) suggests that most miscoverage is incurred because the CIs are too low. Only asymptotically, the errors are balanced out between miscoverage below and above. The reason for this is that $C^{\text{Wald}}(X)$ gets narrower when $\bar{X}$ is smaller. In particular when $\bar{X}$ is smaller than $\mathbb{E}[\bar{X}] = \theta$, in which case the CI is centered below θ. This is something that is likely to happen when θ is small, and even more likely when additionally n is small too. In such cases, $\theta \notin C^{\text{Wald}}(X)$ lies to the right of the CI. Note that this is not specific to small values of θ. What really matters is whether $\bar{X}$ is close to zero or with similar effects, whether $1 - \bar{X}$ is close to 1. In other words, θ and $1 - \theta$ are symmetric problems.*

3.2 Exact Confidence Intervals

In contrast, the exact confidence intervals C^{exact} we build from (one-sided) hypothesis tests will have exact coverage (up to discretization errors), along with balanced coverage (by design). Recall

that we wish to include all the $\tilde{\theta} \in \Theta$ that can plausibly generate the data. For any value $\tilde{\theta} \in [0,1]$, We will test the null hypothesis $\Theta_0 = \{\tilde{\theta}\}$ against $\Theta_1 = [0,1] \setminus \Theta_0$. We want balanced coverage, so we will conduct two one-sided tests and allocate the level-α Type-I error budget between them evenly. For any $\tilde{\theta} \in \Theta$

1. We are going to test $\Theta_0 = \{\tilde{\theta}\}$ against $\Theta_1 = \{\theta > \tilde{\theta}\}$ with the test $\phi^{\uparrow}_{\tilde{\theta}}$
2. We are going to test $\Theta_0 = \{\tilde{\theta}\}$ against $\Theta_1 = \{\theta < \tilde{\theta}\}$ with the test $\phi^{\downarrow}_{\tilde{\theta}}$

Our CI will include all parameters that neither test rejects:

$$C^{\text{exact}}(X) = \{\tilde{\theta} \in \Theta \mid \phi^{\uparrow}_{\tilde{\theta}}(X) = \phi^{\downarrow}_{\tilde{\theta}}(X) = 0\}.$$

For a binomial, we have the monotone likelihood ratio property, so a Neyman-Pearson test reduces to a thresholding rule on $\bar{X}$:

$$\phi^{\uparrow}_{\tilde{\theta}}(X) = \begin{cases} 1 & \bar{X} > \tau^{\uparrow} \\ 0 & \text{otherwise} \end{cases}$$

where $\tau^{\uparrow}$ is the $\left(1 - \frac{\alpha}{2}\right)$-quantile of $\frac{Binom(n,\tilde{\theta})}{n}$. Similarly,

$$\phi^{\downarrow}_{\tilde{\theta}}(X) = \begin{cases} 1 & \bar{X} < \tau \\ 0 & \text{otherwise} \end{cases}$$

where $\tau^{\downarrow}$ is the $\frac{\alpha}{2}$-quantile of $\frac{Binom(n,\tilde{\theta})}{n}$.

Example 6. *We can illustrate the difference with the example $n = 100, \bar{X} = 0$. We will have $C^{\text{exact}}(X) = [0, 0.03]$, but $C^{\text{Wald}}(X) = \{0\}$. Similarly, if $n = 1000$ and $\bar{X} = 0$, we will have $C^{\text{exact}}(X) = [0, 0.003]$, but again $C^{\text{Wald}}(X) = \{0\}$. We can see that C^{exact} scales appropriately with n, while $C^{\text{Wald}}(X)$ is very dependent on the realization of $\bar{X}$.*

We conclude with a series of remarks

Remark 7. *In finite samples*

$$\mathbb{P}_{\theta}(\theta \in C^{\text{exact}}(X)) \geq 1 - \alpha,$$

while in contrast, the Wald intervals only give asymptotic coverage

$$\mathbb{P}_{\theta}(\theta \in C^{\text{Wald}}(X)) \to 1 - \alpha \quad \text{as } n \to \infty.$$

Remark 8. *Instead of splitting the test into two parts, we can also formulate our procedure using the two-sided tests*

$$\phi^{combined}_{\tilde{\theta}}(X) = \begin{cases} 1 & \phi^{\uparrow}_{\tilde{\theta}}(X) = 1 \text{ or } \phi^{\downarrow}_{\tilde{\theta}}(X) = 1 \\ 0. & \end{cases}$$

This even more clearly illustrates that this construction was following the general "duality" recipe.

4 Introduction to p-values

In a nutshell, a p-value is a measurement of disagreement of X with $X \sim P_{\tilde{\theta}}$.

Definition 9. *Let $f : \mathcal{X} \to [0,1]$. Then $f(X)$ is a p-value for $\Theta_0 \subset \Theta$ if*

$$\mathbb{P}_{\tilde{\theta}}(f(X) \leq t) \leq t \quad \text{for all } t \in [0,1] \text{ and } \tilde{\theta} \in \Theta_0.$$

We also call such a random variable $f(X)$ a super-uniform random variable

Note that when the inequality is exact, $f(X)$ is just a uniform random variable when i.e. $\mathbb{P}_{\tilde{\theta}}(f(X) \leq t) = t$ or in other words $f(X) \sim Unif(0,1)$ under $\tilde{\theta} \in \Theta_0$.

Example 10. *Let $X \sim N(\theta, 1)$ and $\Theta_0 = \{0\}$. Then $f(X) = 1 - \Phi(X)$ is a p-value. It's uniformly distributed under the null because*

$$\mathbb{P}_0(f(X) \leq t) = \mathbb{P}(1 - \Phi(X) \leq t) = \mathbb{P}(\Phi(X) \geq 1 - t) = t.$$

Remark 11. *Here, the p-value corresponds to the amount of mass that $N(\theta_0, 1)$ assigns to observations that are more extreme than what we have observed. Indeed, generally, p-values are tail probabilities. Assume now that we observe $n = 10000$ i.i.d. Normal random variables $X_i \sim N(\theta, 1)$. Suppose $\bar{X} = 0.1$ is observed. A corresponding $p-value$ for the null hypothesis $\Theta_0 = \{0\}$ would be 0.0008, suggesting that this event is very unlikely to happen under the null Θ_0. Finally note that in this case, a valid CI could include values between $1 - 0.05$, and $1 + 0.05$, yielding different types of information.*