Nov. 25-29

Stat 330 - Tutorial 11

Today, we are talking about statistical inference. Specifically, confidence intervals and hypothesis tests. The ideas in these sections may be a bit different from how you're used to thinking about CIs and tests. ~~Th~~ The main difference is that we will now be thinking ~~ab~~ about these tools as statisticians (i.e. from the perspective of how they were developed), rather than as users of statistics (i.e. only worrying about how to use them). This is a jump in 'statistical maturity' that will help you understand the inner workings of much of statistical inference.

I'm going to follow the book more closely than usual for this tutorial, because ~~many~~ even many ~~examples~~ examples that start simple end up being either very ~~long~~ long or outright impossible.

Confidence Intervals

Let $\{p_\theta : \theta \in \Theta\}$ be a statistical model. A confidence interval for θ is a set of the form $C(\underline{X}) = (l(\underline{X}), u(\underline{X}))$, where $l(\underline{X}), u(\underline{X}) \in \mathbb{R}$, with $p_\theta(\theta \in C(\underline{X})) = p_\theta(l(\underline{X}) < \theta < u(\underline{X})) \geq \gamma$ for ~~all~~ all $\theta \in \Theta$. Note that ~~p~~ we use p_θ to denote probabilities computed with the true parameter value set to θ. We call γ the confidence level.

Before we move on, ~~there~~ let's go over the standard disclaimer for CIs. Once we have calculated an interval, the probability that it contains θ is either 0 or 1. Our probability statement is about all the intervals we could have calculated, since frequentist probabilities are always statements about ~~infi~~ infinite hypothetical resampling of data.

There are a few standard ways to construct CIs. ~~The~~ One of the more natural methods is called the likelihood approach. The idea here is that larger values of the likelihood f'n correspond to parameter values that are better supported by our data. This suggests that we should choose all θ with likelihood values above some threshold, k. The details of the likelihood method then revolve around how we should choose this threshold. It's best to see this with an example.

②

e.g. 1: Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, with σ^2 known. Find an expression for the likelihood method CI for μ.

First, we need the likelihood. An equivalent expression for the likelihood is

$$\mathcal{L}^*(\mu; \underline{X}) = \exp\left[\frac{-n}{2\sigma^2}(\bar{X}-\mu)^2\right]$$

(equivalent in the sense that we have only dropped multiplicative factors that do not depend on μ).
We want to find ~~for~~ which μ make $\mathcal{L}^*$ larger than some threshold, k.

$$\mathcal{L}^*(\mu; \underline{X}) \geq k$$

$$\exp\left[\frac{-n}{2\sigma^2}(\bar{X}-\mu)^2\right] \geq k$$

$$\frac{-n}{2\sigma^2}(\bar{X}-\mu)^2 \geq \log(k)$$

$$(\bar{X}-\mu)^2 \leq \frac{-2\sigma^2 \log(k)}{n}$$

$$-\sqrt{\frac{-2\sigma^2\log(k)}{n}} \leq \bar{X}-\mu \leq \sqrt{\frac{-2\sigma^2\log(k)}{n}}$$

$$-\bar{X} - \frac{\sigma}{\sqrt{n}}k_1 \leq -\mu \leq -\bar{X} + k_1\frac{\sigma}{\sqrt{n}} \quad \text{where } k_1 = \sqrt{-2\log(k)}$$

$$\bar{X} + \frac{k_1\sigma}{\sqrt{n}} \geq \mu \geq \bar{X} - \frac{k_1\sigma}{\sqrt{n}}$$

Therefore to get $\mathcal{L}^*(\mu; \underline{X}) \geq k$, we should choose μ between $l(\underline{X}) := \bar{X} - \frac{k_1\sigma}{\sqrt{n}}$ and $u(\underline{X}) := \bar{X} + \frac{k_1\sigma}{\sqrt{n}}$.
This gives us the form of our interval, but we still need to choose k_1. For this, we will use the significance level. Specifically, ~~we want a na~~ narrower intervals are better, so we will choose the smallest value for k_1 that gives us the required significance level, γ. It's not immediately obvious how to do this, but ~~in ou~~ fortunately, in our model we know the distribution of a quantity that uses all the terms in our interval.

$$p\left(\bar{X} - k_1\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + k_1\frac{\sigma}{\sqrt{n}}\right)$$

$$= p\left(-k_1\frac{\sigma}{\sqrt{n}} \leq \mu - \bar{X} \leq k_1\frac{\sigma}{\sqrt{n}}\right)$$

$$= p\left(-k_1 \leq \frac{\mu-\bar{X}}{\sigma/\sqrt{n}} \leq k_1\right) = p\left(k_1 \geq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \geq -k_1\right)$$

$$= p(k_1 \geq Z \geq -k_1)$$

where $Z \sim N(0,1)$. The smallest value of k_1 that makes this prob. at least γ is
$k_1 = \Phi^{-1}\left(1 - \frac{1-\gamma}{2}\right) = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$, where Φ^{-1} is the standard normal quantile f'n.

Therefore, our likelihood based CI for μ is $C(\underline{X}) = \left(\bar{X} - \Phi^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma}{\sqrt{n}},\ \bar{X} + \Phi^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma}{\sqrt{n}}\right)$. $\diamond$

Hypothesis Tests

The core idea ~~behi~~ behind testing a hypothesis, H_0, based on a sample of data, $\underline{X}$, is to assume that H_0 is true and calculate the prob. of getting a sample similar to $\underline{X}$, called a p-value. If this prob. is small, we reject H_0. If this prob. is large, we do not reject H_0.

Here is the standard disclaimer for hypothesis tests: not rejecting H_0 is different from accepting H_0. We cannot accept H_0 based on hypothesis testing. For example, consider the normal mean model from e.g. 1. Imagine testing $H_0^{(1)}: \mu = 1$ and $H_0^{(2)}: \mu = 1.00001$ with $\sigma^2 = 10$. If we get a large p-value assuming $H_0^{(1)}$, we should also get a large p-value assuming $H_0^{(2)}$. We can't accept both hypotheses because they can't both be true. However, it does make sense to "not reject" both hypotheses.

Now for some details. Let $\alpha \in (0,1)$ be a significance level. Let $H_0: \theta \in \Theta_0$, $H_A: \theta \in \Theta \setminus \Theta_0$. The p-value based on a sample, $\underline{X}$, is ~~defined~~ calculated in two steps. First, ~~we find~~ for some $\theta \in \Theta_0$, we find the probability of getting evidence against H_0 that is at least as extreme as the evidence provided by $\underline{X}$. We then take the largest such prob. over all $\theta \in \Theta_0$. This ~~largest~~ largest value is called our p-value. Note: in practice, it's often obvious which $\theta \in \Theta_0$ will give us the largest prob., so we can just do calculations with this optimal θ. To check whether we should reject H_0, we compare our p-value against α. If our p-value is smaller, we reject H_0. If our p-value is larger, we do not reject H_0. Let's apply these ideas in our example from last section.

e.g. 1 cont.: Consider testing $H_0: \mu \le 3$ vs $H_A: \mu > 3$ at $\alpha = 0.02$, based on a sample $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with σ^2 known. One way to have evidence against H_0 is with large values of $\bar{X}$. Let w be the value of $\bar{X}$ calculated from our sample. We ~~start by ass~~ must calculate $P_\mu(\bar{X} > w)$ for all $\mu \le 3$ and take the largest

$$P_\mu(\bar{X} > w) = P_\mu\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{w-\mu}{\sigma/\sqrt{n}}\right) = P_\mu\left(Z > \frac{w-\mu}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{w-\mu}{\sigma/\sqrt{n}}\right)$$

You can check by plugging in numbers that this prob. gets larger as μ gets larger, so our p-value is $\rho := 1 - \Phi\left(\frac{w-3}{\sigma/\sqrt{n}}\right)$, since 3 is the largest μ under H_0. We then check if $\rho < \alpha = 0.02$ to decide if we should reject H_0 or not.

$\diamond$

So far, everything has worked out for us because the normal distribution is well-suited for doing inference. What if things aren't as nice?

e.g. 2: Let $X_1, \dots, X_n \sim Exp(\lambda)$. Let's see what happens when we try to construct a likelihood based CI for λ. The likelihood f'n is

$$\mathcal{L}(\lambda;\underline{X}) = \lambda^n \exp\left(\frac{-1}{\lambda}\sum_{i=1}^{n} X_i\right) =: \lambda^n \exp\left(\frac{-1}{\lambda} S_n\right) \quad \text{where } S_n = \sum_{i=1}^{n} X_i$$

Our interval should contain all λ values with likelihood values above some threshold, k.

$$\mathcal{L}(\lambda;\underline{X}) > k$$

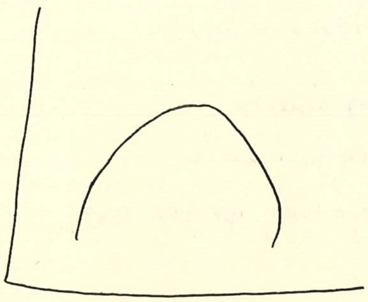
$$\lambda^n \exp\left(\frac{-1}{\lambda} S_n\right) > k$$

This inequality cannot be ~~sol~~ solved analytically for λ. Even if we do get the intervals numerically, it's not clear how we would calculate the probability ~~a~~ that one will cover λ. ~~Th (especially since the distribution of S_n depends on λ~~ We need a way to get CIs (and tests) for λ that doesn't involve solving for our intervals numerically. This is where asymptotic theory comes in ~~to~~ very useful.

Asymptotic Inference

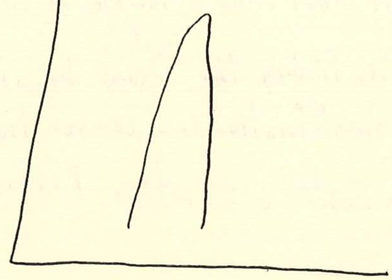
The idea behind asymptotic theory is that as $n \to \infty$, distributions tend to get simpler. We already know ~~about~~ some asymptotic theorems, e.g. the CLT. This th'm says that, regardless of how complicated the distribution of X_i is, $\frac{\bar{X}-\mu}{\sqrt{V(\bar{X})}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$. Another very useful asymptotic theorem (which is actually based on the CLT), is that MLEs also ~~satisfy this~~ converge to normals. Specifically, (assuming some regularity conditions) if $\hat{\theta}$ is an MLE, then $\frac{\hat{\theta}-\theta}{\sqrt{V(\hat{\theta})}} \xrightarrow{D} N(0,1)$. If we can get ~~$E(\hat{\theta})$ and~~ $V(\hat{\theta})$ then we can start to do inference. Fortunately, there is already general theory for ~~this~~ calculating $V(\hat{\theta})$. ~~The It can be shown that~~ Before getting into any formulas, let's think about $V(\hat{\theta})$. We calculate $\hat{\theta}$ by maximizing the log-likelihood f'n, but some log-likelihoods are easier to maximize than others.

A typical likelihood



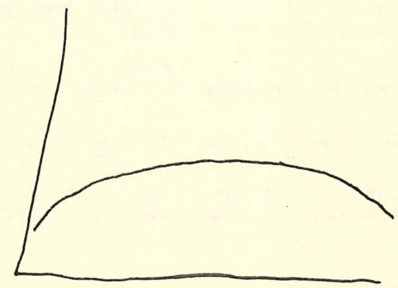
A

Easier to find maximizer



B

Harder to find maximizer



C

The important difference between these ~~3~~ three likelihoods is their ~~sec~~ ~~second derivative curvit~~ second derivative (i.e. curviture) at the maximizer. Specifically, the negative second ~~derivative~~ derivative at $\hat{\theta}$ is much larger in B than in C. This makes $\hat{\theta}$ easier to estimate in B than in C and thus, $V(\hat{\theta})$ should be smaller in B than in C. The important quantity here, the negative second derivative of the log-likelihood at $\hat{\theta}$, is called the observed Fisher Information. The population Fisher Information is just the expected value of this quantity. ~~evaluated at the true θ.~~

It is ~~someti~~ so often useful to consider the Fisher Information at different θ, so we treat it as a f'n of θ.

Observed Fisher Information: $\hat{I}(\underline{X}) := -\frac{\partial^2 \ell(\theta;\underline{X})}{\partial \theta^2}\Big|_{\theta=\hat{\theta}}$

Fisher Information: $I(\theta;\underline{X}) := E_\theta\left[-\frac{\partial^2 \ell(\theta;\underline{X})}{\partial \theta^2}\right]$

Note: Don't forget the ~~m~~ minus sign!

It turns out that $V(\hat{\theta}) \to \frac{1}{\sqrt{I(\theta_0;\underline{X})}}$ as $n \to \infty$ where θ_0 is the true θ_0. This makes sense, since larger Fisher Information gives smaller ~~va~~ asymptotic variance. In practice, we don't know θ_0, but we still want to know the limit for $V(\hat{\theta})$, so we need to estimate $I(\theta;\underline{X})$. One way to do this is with ~~a~~ the observed Fisher Information, $\hat{I}(\underline{X})$. Another option is to calculate the true Fisher Information at our ~~es~~ MLE, $I(\hat{\theta};\underline{X})$. This is called the plug-in estimator because we take the thing we want to know, $I(\theta_0;\underline{X})$, and plug-in our estimate for the thing we don't know, $\hat{\theta}$ for θ_0.

It turns out that asymptotic normality still holds for $\hat{\theta}$ if we replace the true Fisher Information with its plug-in estimator. I.e. $\frac{\hat{\theta}-\theta}{1/\sqrt{I(\hat{\theta};\underline{X})}} \xrightarrow{D} N$.

⑥

One more observation about Fisher Information, then we'll do an example. So far, we have only discussed the F.I. based on a sample. It turns out that, for a sample of size n, $I(\theta;\underline{X}) = n\,I(\theta;X_i)$. This is nice, because $I(\theta;X_i)$ only requires finding the negative second derivative of the log-density, which is often easier to work with. This simplification is so common that people often just write $I(\theta) := I(\theta;X_i)$ for the F.I. based on a single observation. Putting this all together, we get that

$$\sqrt{n I(\hat{\theta})}\,(\hat{\theta}-\theta) \xrightarrow{D} N$$

Now let's return to our exponential CI example.

e.g. 2 cont.: We saw previously that a likelihood based CI in this model seems hopeless. We know from previous work that the MLE of λ is $\hat{\lambda} = 1/\bar{X}$. Next, we need the F.I. Let's find $I(\theta)$ (i.e. based on a single obs.)

$\boxed{\text{Oops! Let } \theta = \lambda.}$

$$\begin{aligned} \ell(\theta;x) &= \log[f(x;\theta)] \\ &= \log(\theta e^{-\theta x}) \\ &= \log(\theta) - \theta x \end{aligned}$$

The first and second derivatives of this ~~is~~ are

$$\frac{\partial \ell(\theta;x)}{\partial\theta} = \frac{1}{\theta} - x \qquad \frac{\partial^2 \ell(\theta;x)}{\partial\theta^2} = \frac{-1}{\theta^2}$$

Therefore, the F.I. is

$$I(\theta) = E\left[-\left(-\frac{1}{\theta^2}\right)\right] = \frac{1}{\theta^2} = \frac{1}{\lambda^2}$$

The plug-in estimate of the F.I. is

$$I(\hat{\lambda}) = \frac{1}{\hat{\lambda}^2} = \bar{X}^2$$

The plug-in estimate of the F.I. based on a sample of size n is $nI(\hat{\lambda}) = n\bar{X}^2$. Therefore, the asymptotic variance of $\hat{\lambda}$ is $\frac{1}{nI(\hat{\lambda})} = \frac{1}{n\bar{X}^2}$. We can construct an asymptotic CI for λ by just treating $\hat{\theta}$ as $N(\theta, 1/n\bar{X}^2)$. In e.g. 1, we saw how to get a CI based on a normally distributed statistic: $\hat{\theta} \pm \Phi^{-1}\left(\frac{1+\gamma}{2}\right)\sqrt{V(\hat{\theta})}$. In this case, our interval is

$$\hat{\lambda} \pm \Phi^{-1}\left(\tfrac{1+\gamma}{2}\right)\Big/\sqrt{nI(\hat{\lambda})} = \frac{1}{\bar{X}} \pm \Phi\left(\tfrac{1+\gamma}{2}\right)\Big/\sqrt{n}\,\bar{X}.$$

◇

Note: If the observed F.I., $\hat{I}(\underline{X})$, is easier to work with, you can instead do everything we just covered, replacing $nI(\hat{\theta}) = I(\hat{\theta};\underline{X})$ with $\hat{I}(\underline{X})$ (this is true in general, not just for our example).