Lecture 26

Our attention now turns to *statistical inference* where we try to understand poplns based on sample data. We first study *confidence intervals.*

The Problem: Given a statistical model (eg. $X \sim \text{Normal}(\mu, \sigma^2)$, $Y \sim \text{Bin}(n, p)$, $W \sim \text{Poisson}(\theta)$**), the estimation problem is to learn about unknown parameters (eg.** μ, σ, p, θ**) given observed data (eg.** X**'s,** Y**'s,** W**'s).**

Idea 1: We might estimate the population mean μ **with the *point estimate*** $\bar{X}$**. Point estimation is barely mentioned in the text. Although seemingly sensible, the problem is that we do not know about the closeness of the estimate** $\bar{X}$ **to the unknown parameter** μ**.**

Idea 2: Interval estimation involves constructing an interval (eg. (7.3,12.6)) in which we are confident that μ **resides.**

We begin with confidence interval construction in the simplest context. Consider $X_1, \dots, X_n$ **iid** $\text{Normal}(\mu, \sigma^2)$ **where** μ **is unknown,** σ **is known and the observed value of** $\bar{X}$ **is** $\bar{X}_{\text{obs}}$**.**

Note that this is an unrealistic scenario. When is it ever the case that the mean parameter is unknown but the variance parameter is known? Ignoring the criticism, $\bar{X} \sim \text{Normal}(\mu, \sigma^2/n)$**. A 95% confidence interval for** μ **is obtained via:**

$$\begin{aligned}
& \text{P}\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95 \\
\Leftrightarrow\ & \text{P}\left(-1.96\frac{\sigma}{\sqrt{n}} - \bar{X} < -\mu < 1.96\frac{\sigma}{\sqrt{n}} - \bar{X}\right) = 0.95 \\
\Leftrightarrow\ & \text{P}\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \\
\Rightarrow\ & \bar{X}_{\text{obs}} \pm 1.96\frac{\sigma}{\sqrt{n}} \quad \textbf{is a 95\% CI for } \mu
\end{aligned}$$

More generally,

$$\bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad \textbf{is a } (1-\alpha)100\% \textbf{ CI for } \mu.$$

Interpretation of CI's: The explanation is subtle and you need to pay close attention.

Consider many hypothetical replications of an experiment.

A common but incorrect interpretation for CI'S:

If $\bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ **is a** $(1-\alpha)100\%$ **CI for** μ**, it is incorrect to write** $\mathrm{P}\left(\mu \in \bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$**.**

Discussion points wrt the CI $\bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$:

- **as n increases, the width of the CI decreases**
- **as our confidence increases (ie. $1-\alpha$ bigger), the width of the CI increases**
- **tradeoff: we want narrow CI's with large confidence**
- **a CI of a given confidence $1-\alpha$ is not unique**

The simple but unrealistic CI setting previously presented is extended to more realistic scenarios.

We begin by assuming that our sample $X_1, \ldots, X_n$ **is large (ie.** $n \geq 30$**) as is often the case in practice.**

Case 1: Since n **is large, we can invoke the CLT where approximately** $\bar{X} \sim \text{Normal}(\mu, \sigma^2/n)$**. What is great about this is that we no longer need to assume that the** X**'s are normal. In this case,**

$$\bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

is an approximate $(1-\alpha)100\%$ **CI for** μ **where** σ **is still assumed known.**

Case 2: We have the same conditions as Case 1 except that σ **is unknown. In this realistic case,**

$$\bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

is an approximate $(1-\alpha)100\%$ **CI for** μ **where** s **is the sample standard deviation.**

Example: Consider heat measurements taken in degrees Celsius where $\mu = 5$ and $\sigma = 4$. A change is made in the process such that μ changes but σ remains the same. We observe $\bar{X}_{\text{obs}} = 6.1$ based on $n = 100$ observations.

(a) Construct a 90% CI for μ.

(b) How big should the sample size be such that the CI is less than 0.6 degrees wide?

Problem: Consider the CI $\bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.

(a) How much should the sample size n increase to reduce the width of by half?

(b) What is the effect of increasing the sample size by a factor of 25?