Sets 22: Section 6.1, Confidence Intervals for Normal (variance known) and large samples

Statistical inference: **understand populations given sample data. We first study *confidence intervals*.**

The Problem: Given a statistical model
e.g. $X \sim \text{Normal}(\mu, \sigma^2)$, $Y \sim \text{Bin}(n, p)$, $W \sim \text{Poisson}(\theta)$,
we want to say something about unknown parameters, μ, σ, p, θ, **using observed data,** X**'s,** Y**'s,** W**'s.**

Idea 1: Estimate the population mean μ **with the *point estimate*** $\bar{X}$**.**

We want to say more ... quantify the margin of error in our estimate.

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

Consider the simplest scenario. Consider $X_1, \ldots, X_n$ **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.

Ignoring the criticism, $\bar{X} \sim \text{Normal}(\mu, \sigma^2/n)$**. A 95% confidence interval for** μ **is obtained via:**

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. ¡applet¿

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 $\mathrm{P}\left(\mu \in \bar{x}_{\text{obs}} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$.

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

- **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**

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 use the CLT where $\bar{X} \approx \text{Normal}(\mu, \sigma^2/n)$. 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?