# The Central Limit Theorem

In the past few lectures we have studied statistical inference and have often assumed we have normally distributed observations. Two reasons for this are

- Quite often real data is (approximately) normal (although this must be verified, for example, with a q-q-plot)
- The analysis is tractable.

This lecture discusses a third reason, which hinges on a key result called the <u>central limit theorem</u>.

# The Central Limit Theorem

Let $X_1, \ldots, X_n$ be a random sample (not necessarily normal) with mean $\mu$ and variance $\sigma^2 > 0$. ~~Denote~~

$$\cancel{W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma}}.$$

Then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{\sim} N(0,1)$$

as $n \to \infty$.

---

(The precise nature of the convergence is not covered in this course.)

Practically, we can use the CLT to say, for $n$ "sufficiently large,"

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1).$$

Recall, when we derived confidence intervals and tests for a normal sample ($\sigma^2$ known), we only used the fact that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

By the CLT, even if the sample is not normal, it is still true that

$$P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > z_\alpha\right) \approx \alpha.$$

Therefore, by the same analysis as before, we can show that

$$\left[ \bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

is an approximate $(100(1-\alpha))\%$ confidence interval for $\mu$.

Similarly, the test

$$\begin{cases} H_0 & \dfrac{|\bar{X}-\mu|}{\sigma/\sqrt{n}} < z_{\alpha/2} \\[2ex] H_1 & \dfrac{|\bar{X}-\mu|}{\sigma/\sqrt{n}} > z_{\alpha/2} \end{cases}$$

has level of significance approximately equal to $\alpha$.

~~The~~ Note that the CLT does not assume the data is continuous. It also hold for <u>discrete</u> data.

<u>Example</u> $X_1, \dots, X_n \sim \text{Poisson}(\lambda)$.

Then $\mu = \lambda$, $\sigma^2 = \lambda$.

By the CLT,

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \approx N(0,1).$$

Observe

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} = \frac{\sum_{i=1}^{n} X_i - \lambda n}{\sqrt{n\lambda}}$$

Thus, $\sum_{i=1}^{n} X_i \approx N(\lambda n, \lambda n)$

But $\sum_{i=1}^{n} X_i \sim \text{Poisson}(\lambda n)$. Making the substitution $\lambda n \to \lambda$, we conclude

$$\boxed{\text{Poisson}(\lambda) \approx N(\lambda,\lambda) \text{ for large } \lambda.}$$

How large does $n$ need to be for the CLT ~~to~~ approximation to hold? There are no guarantees, but here are some rules of thumb:

- If $n \approx 25$ or so, the approximation probably hold.

- If $X$ is symmetric, unimodal (only one peak), and continuous, then $n \approx 5$ may be sufficient (Example: $t$, $\chi^2$, Gamma).

See examples in book.