

## 5.27 Central Limit Theorem

CLT

Example

Normal Approximation to the Binomial

The most important theorem in prob and stats.

**Central Limit Theorem:** Suppose  $X_1, \dots, X_n$  are i.i.d. with  $E[X_i] = \mu$  and  $\text{Var}(X_i) = \sigma^2$ . Then as  $n \rightarrow \infty$ ,

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} \text{Nor}(0, 1),$$

where “ $\xrightarrow{\mathcal{D}}$ ” means that the c.d.f.  $\rightarrow$  the  $\text{Nor}(0, 1)$  c.d.f.

Proof: Not in this class.

Remarks: (1) So if  $n$  is large, then  $\bar{X} \approx \text{Nor}(\mu, \sigma^2/n)$ .

(2) The  $X_i$ 's *don't have to be normal* for the CLT to work!

(3) You usually need  $n \geq 30$  observations for the approximation to work well. (Need fewer observations if the  $X_i$ 's come from a symmetric distribution.)

(4) You can almost always use the CLT if the observations are i.i.d.

Example: Suppose  $X_1, \dots, X_{100} \stackrel{\text{iid}}{\sim} \text{Exp}(1/1000)$ . Find  $\Pr(950 \leq \bar{X} \leq 1050)$ .

Solution: Recall that if  $X_i \sim \text{Exp}(\lambda)$ , then  $E[X_i] = 1/\lambda$  and  $\text{Var}(X_i) = 1/\lambda^2$ .

Further, if  $\bar{X}$  is the sample mean based on  $n$  observations, then

$$E[\bar{X}] = E[X_i] = 1/\lambda \quad \text{and}$$

$$\text{Var}(\bar{X}) = \text{Var}(X_i)/n = 1/(n\lambda^2).$$

For our problem,  $\lambda = 1/1000$  and  $n = 100$ , so that  $E[\bar{X}] = 1000$  and  $\text{Var}(\bar{X}) = 10000$ .

So by the CLT,

$$\begin{aligned} & \Pr(950 \leq \bar{X} \leq 1050) \\ &= \Pr\left(\frac{950 - E[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}} \leq \frac{\bar{X} - E[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}} \leq \frac{1050 - E[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}}\right) \\ &\approx \Pr\left(\frac{950 - 1000}{100} \leq Z \leq \frac{1050 - 1000}{100}\right) \\ &\approx \Pr\left(-\frac{1}{2} \leq Z \leq \frac{1}{2}\right) = 2\Phi(1/2) - 1 = 0.383. \end{aligned}$$

Example: Suppose  $X_1, \dots, X_{100}$  are i.i.d. from some distribution with mean 1000 and standard deviation 1000. Find  $\Pr(950 \leq \bar{X} \leq 1050)$ .

Solution: By exactly the same manipulations as in the previous example, the answer  $\approx 0.383$ .

Notice that we didn't care whether or not the data came from an exponential distrn. We just needed the mean and variance.

## Normal Approximation to the Binomial( $n, p$ )

Suppose  $Y \sim \text{Bin}(n, p)$ , where  $n$  is very large. In such cases, we usually approximate the Binomial via an appropriate Normal distribution.

The CLT applies since  $Y = \sum_{i=1}^n X_i$ , where the  $X_i$ 's are i.i.d.  $\text{Bern}(p)$ .

Then

$$\frac{Y - \mathbf{E}[Y]}{\sqrt{\text{Var}(Y)}} = \frac{Y - np}{\sqrt{npq}} \approx \text{Nor}(0, 1).$$

Why do we need such an approximation?

Example: Suppose  $Y \sim \text{Bin}(100, 0.8)$  and we want

$$\Pr(Y \geq 90) = \sum_{i=90}^{100} \binom{100}{i} (0.8)^i (0.2)^{100-i}.$$

Good luck with the binomial coefficients (they're too big) and number of terms to sum up (it's going to get tedious). I'll come back to visit you in an hour.

So how do we use the approximation?

Example: The Braves play 100 indep baseball games, each of which they have prob 0.8 of winning. What's the prob that they win at least 90?

$Y \sim \text{Bin}(100, 0.8)$  and we want  $\Pr(Y \geq 90)$  (as in the last example)...

$$\Pr(Y \geq 90) = \Pr(Y \geq 89.5) \quad (\text{“continuity correction”})$$

$$\approx \Pr\left(Z \geq \frac{89.5 - np}{\sqrt{npq}}\right) \quad (\text{CLT})$$

$$= \Pr\left(Z \geq \frac{89.5 - 80}{\sqrt{16}}\right) = \Pr(Z \geq 2.375) = 0.0088.$$

Use the continuity correction since the Binomial is a discrete distrn while the Normal is cts. If you don't want to use it, don't worry too much.