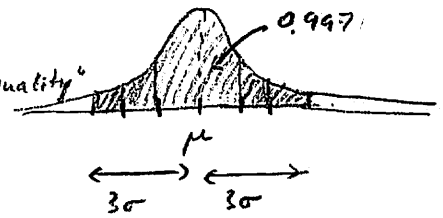
Ex $X \sim N(\mu, \sigma^2)$. Then

$$P\{|X-\mu| \le 3\sigma\} \approx 0.997.$$

"Deviation inequality"



Indeed, $Z := \frac{X-\mu}{\sigma} \sim N(0,1)$

$$P\{|X-\mu| \le 3\sigma\} = P\{|Z| \le 3\} \approx 0.997.$$

In words: "A normal random variable stays within 3 standard deviations from its mean with probability 99.7%".

Lec. 22 03/05

5.4.1 Normal approximation to Binomial

- Normal distribution arises in many applications. Here is one, where we can guarantee that normal distr. can be used.
- Recall Poisson approx. (4.7), which is valid for rare successes ($p \to 0$):

$$\text{Binom}(n,p) \approx \text{Poisson}(\lambda), \quad \text{if } n \to \infty \text{ and } np \to \lambda = \text{const}.$$

- Normal approx. holds when successes are not rare ($p = \text{const}$):

$$\text{Binom}(n,p) \approx N(\mu, \sigma^2) \quad \text{where } \mu = np,\ \sigma^2 = np(1-p),$$
$$\text{if } n \to \infty \text{ and } p = \text{const}$$

- In other words, if $X \sim \text{Binom}(n,p)$, its Z-score $Z = \frac{X-np}{\sqrt{np(1-p)}}$ is approximately $N(0,1)$
- Formally, this is the content of the Central Limit Theorem (CLT):

Thm (De Moivre-Laplace CLT):

Let $X \sim \text{Binom}(n,p)$; consider the Z-score $Z = \frac{X-np}{\sqrt{np(1-p)}}$.

Then, for every $a \in \mathbb{R}$,

$$F_Z(a) \longrightarrow \Phi(a) \quad \text{as } n \to \infty,\ p \text{ fixed}$$

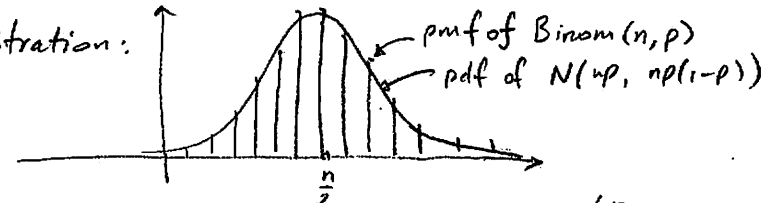
$F_Z(a)$: cdf of Z; $\Phi(a)$: cdf of $N(0,1)$

- In particular, $P\{a \le Z \le b\} \to \Phi(b) - \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\,dx$, as $n \to \infty$, p fixed

$$\| \quad P\left\{a \le \frac{X-np}{\sqrt{np(1-p)}} \le b\right\}.$$

- The proof (of a more general CLT) will be given later

- Illustration:



Ex Flip a fair coin 100 times. What is the probability that the # of heads is between 40 and 60?

$X \sim \text{Binom}(100, \frac{1}{2})$. $np = 50$, $\sqrt{np(1-p)} = 5$.

$$P\{40 \le X \le 60\} = P\left\{\underbrace{\frac{40-50}{5}}_{=-2} \le \underbrace{\frac{X-50}{5}}_{=Z\sim N(0,1)} \le \frac{60-50}{5}_{=2}\right\} = \Phi(2) - \Phi(-2) = 2\Phi(2) - 1 = 0.954$$

NOTE: Actual number (from exactly computating CDF of Binom) is 96.5 %.

Mention continuity correction;
$P\{39.5 \le X \le 60.5\} = \ldots = 0.964$,
closer to the actual number

Ex 51% of the newborn children are boys.
In a certain community, more girls than boys were born in 2011.
The total # of children born is 1,000.
Can this be caused by random fluctuations?

(What is the prob. that more girls than boys are among 1000 newborns?)

$X = \#\text{boys} \sim \text{Binom}(\underset{n}{1000}, \underset{p}{0.51})$

$$P\{\#\text{girls} \ge \#\text{boys}\} = P\{X < 500\} = P\left\{\frac{X-np}{\sqrt{np(1-p)}} < \frac{500-np}{\sqrt{np(1-p)}}\right\}$$

$$\overset{CLT}{\approx} P\{\underset{N(0,1)}{Z} < -0.63\} = \Phi(-0.63) = 1 - \Phi(0.63) \cong 0.264$$

So, with prob. 26.4% this can happen due to random fluctuations.

- NOTE: Actual number (by exactly evaluating CDF of Binom) is 27.4%

Ex (Predicting the outcome of elections).
In order to predict the outcome of a presidential election, a poll is taken.
The predicted percentage of votes for a candidate is compiled from the poll.
How many people need to be included in the poll to validate the following claim:
"The predicted percentage is accurate to within 1% with probability ≥ 0.95"?

Let p = actual probability, n = poll size.
S_n = # people in favor $\sim \text{Binom}(n, p)$.
Predicted percentage = $\frac{S_n}{n}$.
Claim: "$P\{|\frac{S_n}{n} - p| \le 0.01\} \ge 0.95$."

$$= P\left\{\left|\frac{S_n - np}{n}\right| \le 0.01\right\} = P\left\{\left|\frac{S_n - np}{\sqrt{np(1-p)}}\right| \le \underbrace{0.01\sqrt{\frac{n}{p(1-p)}}}_{=a}\right\} \overset{CLT}{\approx} \Phi(a) - \Phi(-a) = 2\Phi(a) - 1$$

Solving $2\Phi(a) - 1 = 0.95$ gives $\Phi(a) = 0.975 \Rightarrow a = 1.96$

$\Rightarrow 0.01\sqrt{\frac{n}{p(1-p)}} \ge 1.96$ $\quad n \ge 38,416\, p(1-p)$. Right Hand side is maximized for $p = \frac{1}{2} \Rightarrow$ $n \ge 9604$.