

Sampling distributions

Probability distribution of Y : serves as a **model** of a **population** of quantities

μ, σ^2, π , etc. are **parameters**: constants (usually unknown) which characterize properties of the distribution

Y_1, Y_2, \dots, Y_n : a **random sample** (independent, identically distributed random variables)

Statistic: quantity calculated from Y_1, Y_2, \dots, Y_n (& possibly *known* parameters), usually for the purpose of estimating an unknown parameter

Examples of statistics

$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)$$

sample mean

$$S^2 = \frac{1}{(n-1)} \left[(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2 \right]$$

sample variance

Statistics are themselves **random variables** with probability distributions

TRUE FACTS about \bar{Y} :

1. If Y has *any* probability distribution with mean μ and variance σ^2 , then

$$E(\bar{Y}) = \mu_{\bar{Y}} = \mu$$

$$V(\bar{Y}) = \sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n}$$

2. If Y has a normal(μ, σ^2) distribution, then

$$\bar{Y} \sim \text{normal} \left(\mu, \frac{\sigma^2}{n} \right)$$

3. **Central Limit Theorem (CLT)**: if Y has *any* probability distribution with mean μ and variance σ^2 , then the distribution of \bar{Y} converges to a normal $\left(\mu, \frac{\sigma^2}{n}\right)$ distribution as $n \rightarrow \infty$

$$\left(P \left(\frac{\bar{Y} - \mu}{(\sigma/\sqrt{n})} \leq z \right) \rightarrow P(Z \leq z), \text{ where } Z \sim \text{normal}(0, 1) \right)$$

(If \bar{Y} is thought of as an estimate of μ , this property is called **asymptotic normality**)

4. **Law of Large Numbers (LLN)**: If Y has *any* distribution with mean μ and variance σ^2 , the probability that \bar{Y} is within ϵ of μ (where $\epsilon > 0$) converges to 1 as $n \rightarrow \infty$

$$(P(|\bar{Y} - \mu| < \epsilon) \rightarrow 1 \text{ as } n \rightarrow \infty)$$

In other words, the distribution of \bar{Y} *concentrates* around μ :

(If \bar{Y} is thought of as an estimate of μ , this property is called **statistical consistency**)

Variants of TRUE FACTS 1-3 for sums:

$$W = Y_1 + Y_2 + \cdots + Y_n$$

1. Y any distribution, mean μ , variance σ^2 , then

$$E(W) = n\mu$$

$$V(W) = n\sigma^2$$

2. $Y \sim \text{normal}(\mu, \sigma^2)$, then

$$W \sim \text{normal}(n\mu, n\sigma^2)$$

3. Y any distribution, mean μ , variance σ^2 , then the distribution of W converges to a normal($n\mu, n\sigma^2$) distribution

ex. Draw 10 students at random from UI and find out their SAT-math scores. In the nation, a randomly drawn SAT-math score has a normal distribution with a mean of 500 and a standard deviation of 100. If UI students were similar to students in the nation at large, what is the probability that \bar{Y} would be greater than or equal to 560?

$$Z = \frac{\bar{Y} - \mu}{(\sigma/\sqrt{n})} = \frac{560 - 500}{(100/\sqrt{10})} \approx 1.90; \text{ area} = 0.0287$$

CLT applied to the binomial distribution

Independent success/failure trials; $\pi = \text{prob. of success}$

$$I = \begin{cases} 1 & \text{if trial is a success} \\ 0 & \text{if trial is a failure} \end{cases}$$

$$E(I) = \pi \quad (= 0 \cdot (1 - \pi) + 1 \cdot \pi)$$

$$V(I) = \pi(1 - \pi) \quad (= (0 - \pi)^2 \cdot (1 - \pi) + (1 - \pi)^2 \cdot \pi)$$

$$Y = I_1 + I_2 + \cdots + I_n \sim \text{binomial}(n, \pi)$$

So Y is a *sum*; distribution of Y can be approximated by a normal($n\pi, n\pi(1 - \pi)$) distribution

Approximation is good if $n\pi \geq 5$ and $n(1 - \pi) \geq 5$

ex. Guess the suit of the top card in a shuffled deck.

Repeat (shuffle/guess) 100 times. What is the chance of 30 or more correct guesses? $\pi = 0.25$ $n = 100$

$$\begin{aligned} P(Y \geq 30) &\approx P\left(Z \geq \frac{30 - 100(.25)}{\sqrt{100(.25)(1-.25)}}\right) = P(Z \geq 1.15) \\ &= 0.125 \end{aligned}$$

Correction for continuity: improves normal approximation to binomial

$$\begin{aligned} P(Y \geq 30) &\approx P\left(Z \geq \frac{29.5 - 100(.25)}{\sqrt{100(.25)(1-.25)}}\right) = P(Z \geq 1.04) \\ &= 0.149 \end{aligned}$$

True probability is .14954