Previously on 4SSPP109…

- Random variables
- Probability distribution
- Probability density function
- Expected value
- Variance & SD

Standardized random variables

- Random variable Y has mean μ_Y and variance σ_Y^2.
- *Standardized* version of Y:

$$Z = \frac{(Y - \mu_Y)}{\sigma_Y}$$

- By design, if Z is a standardized RV, we always have

$$E(Z) = 0$$

$$Var(Z) = SD(Z) = 1$$

TWO RANDOM VARIABLES

- In the US, are Democrats more likely to get vaccinated against Covid than Republicans?

- Are graduates more likely to find a job than non-graduates?
- How do women and men's average earnings differ?
- All these Qs involve the relationship between **two RVs**.

JOINT PROBABILITY DISTRIBUTION

- *Joint probability distribution* of X and Y:

$$\mathbf{p}(x_j, y_i) = Pr(X = x_j, Y = y_i)$$

- *Conditional distribution* of Y given X:

$$Pr(Y = y_i | X = x_j) = \frac{Pr(X = x_j, Y = y_i)}{Pr(X = x_j)}$$

- *Conditional expectation* (or conditional mean) of Y given X:

$$E(Y | X = x_j) = \sum_{i=1}^{k} y_i \, Pr(Y = y_i | X = x_j)$$

EXAMPLE: *Joint distribution of Covid VAX Status & Partisanship in US*

	Dem (X=1)	Rep (X=0)	**Total**
Vaccinated (Y=1)	0.37	0.21	0.58
Not Vaccinated (Y=0)	0.18	0.24	0.42
Total	0.55	0.45	1.00

Source: calculated and adapted from data in [New York Times "The Vaccine Class Gap", 5-24-2021]

Your Turn: Figure out the following

1. Pr(Y=1, X=0)
2. Pr(Y=1 | X=1)
3. E(Y|X=1)

Reminder:

- $Pr(Y = y|X = x) = \frac{Pr(X=x,Y=y)}{Pr(X=x)}$
- $E(Y|X = x) = \sum_{i=1}^{k} y_i Pr(Y = y_i|X = x)$

EXAMPLE: *Joint distribution of Covid VAX Status & Partisanship in US*

	Dem (X=1)	Rep (X=0)	**Total**
Vaccinated (Y=1)	0.37	0.21	0.58
Not Vaccinated (Y=0)	0.18	0.24	0.42
Total	0.55	0.45	1.00

1. Pr(Y=1, X=0) = 0.21
2. Pr(Y=1 | X=1) = 0.37/0.55=0.67
3. E(Y|X=1) = $0 * 0.33 + 1 * 0.67 = 0.67$

Reminder:

- $Pr(Y = y|X = x) = \frac{Pr(X=x, Y=y)}{Pr(X=x)}$
- $E(Y|X = x) = \sum_{i=1}^{k} y_i Pr(Y = y_i|X = x)$

Independence

- X and Y are *independently distributed* (or *independent*) if

$$\Pr(Y = y|X = x) = \Pr(Y = y) \qquad \text{for all possible X and Y}$$

- If X and Y are independent, then

$$\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y)$$

- Example: if rain does not affect Liverpool's performance, then

$$\Pr(Rain, Liverpool\ wins) = \Pr(Rain) * \Pr(Liverpool\ wins)$$

Covariance

- How much do X and Y move together?
- Covariance:

(Assuming X & Y are discrete RVs with *k* & *l* possible realizations)

$$cov(X,Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{l} (x_j - \mu_x)(y_i - \mu_Y) \Pr(X = x_y, Y = y_i)$$

- Also, $cov(X,Y) = E[XY] - \mu_x \mu_Y$

Correlation

- The units of covariance are awkward (units of X * units of Y).
- Correlation:

$$corr(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

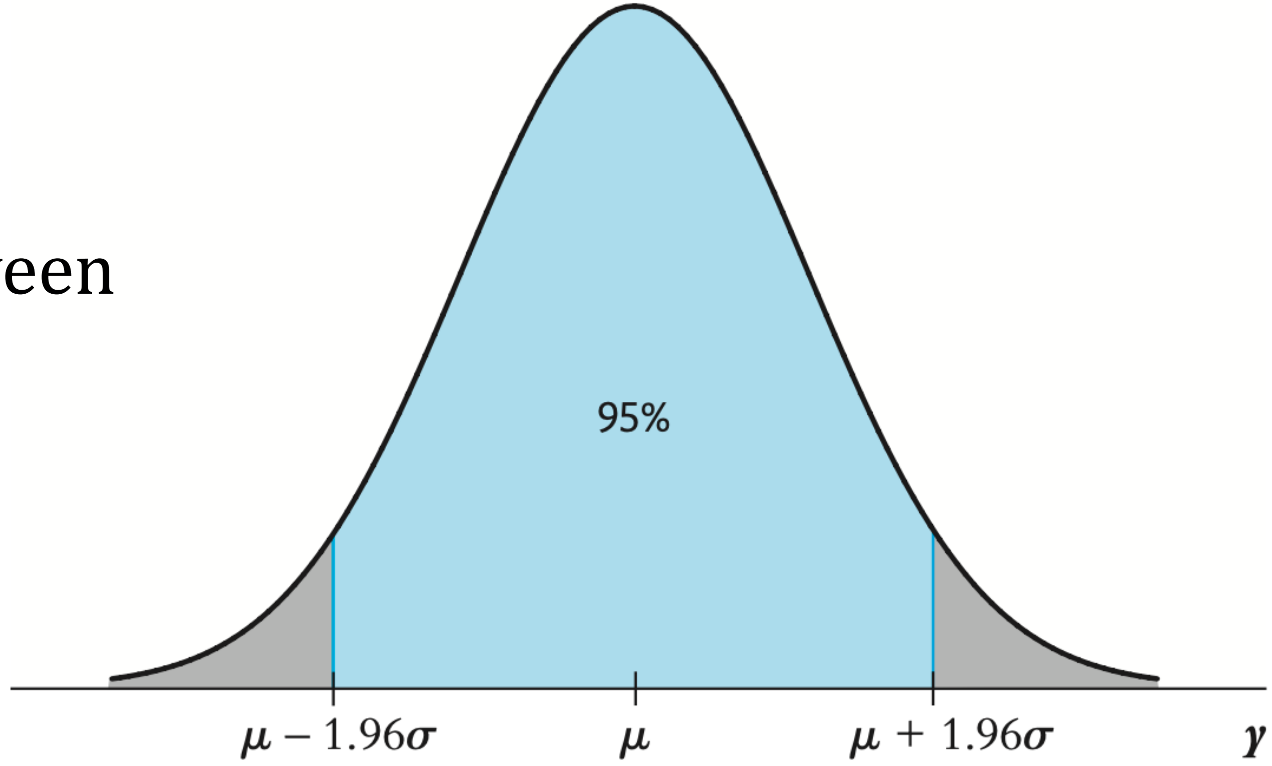
- Correlation is unit free and always between -1 and +1.

Sums of random variables

- $E(X+Y) = E(X) + E(Y) = \mu_X + \mu_Y$
- $Var(X+Y) = \mathrm{var(X)} + \mathrm{var(Y)} + 2\mathrm{cov(X,Y)}$
- If Y & Z are independent → $Var(X+Y) = \mathrm{var(X)} + \mathrm{var(Y)}$

The Normal distribution

- A particular type of p.d.f.
- Bell-shaped & symmetric.
- 95% of probability mass between $\mu - 1.96\sigma$ / $\mu + 1.96\sigma$
- Written as $N(\mu, \sigma^2)$
- Some random variables are distributed normally.



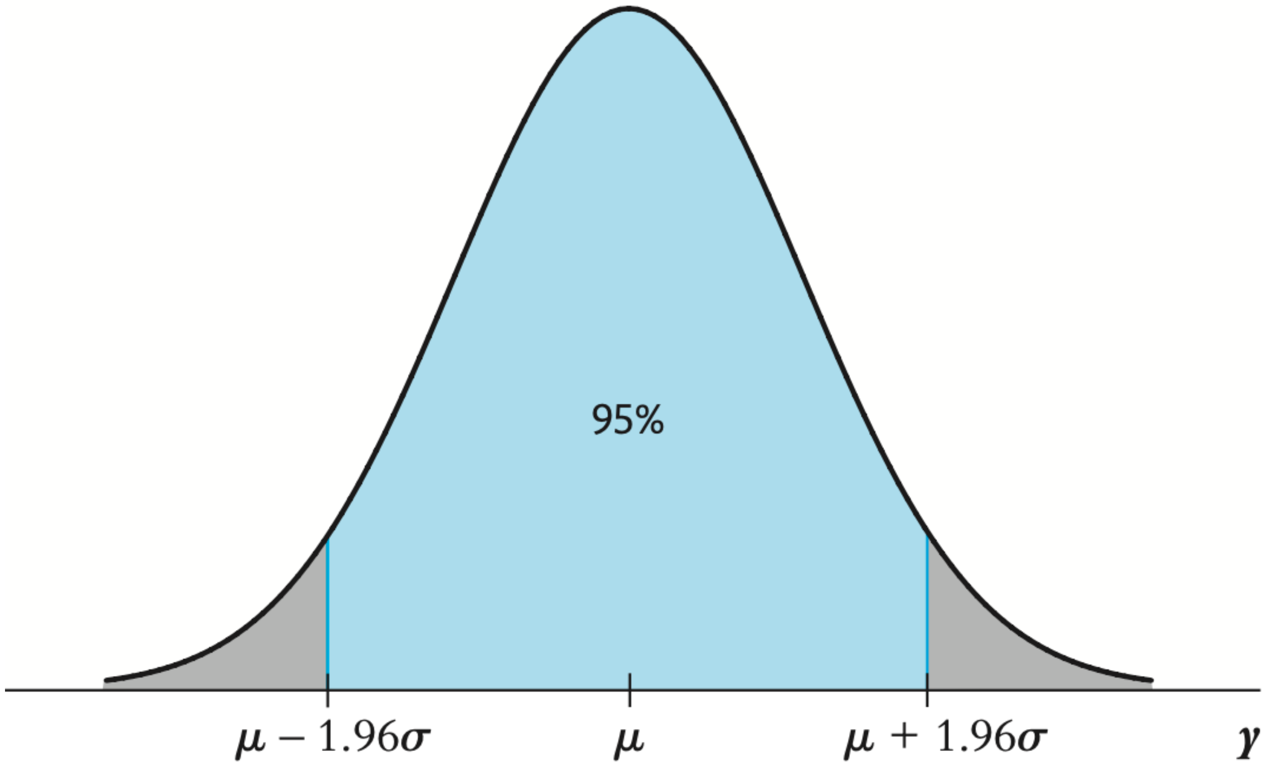
The *Standardized* Normal distribution

1. Take a variable Y distributed $N(\mu, \sigma^2)$.

2. Standardize it:

$$Z = \frac{(Y - \mu_Y)}{\sigma_Y}$$

3. Z is distributed $N(0,1)$.



The *Standardized* Normal distribution

- Z is distributed $N(0,1)$
- Then $\Pr(Z \leq z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}} du$
- To get $\Phi(z)$ for any z:
    - Excel: *NORMSDIST(z)*
    - STATA: *display normal (z)*
    - Table 6.1: look up the cumulative probability of the desired value.

The Normal distribution: example

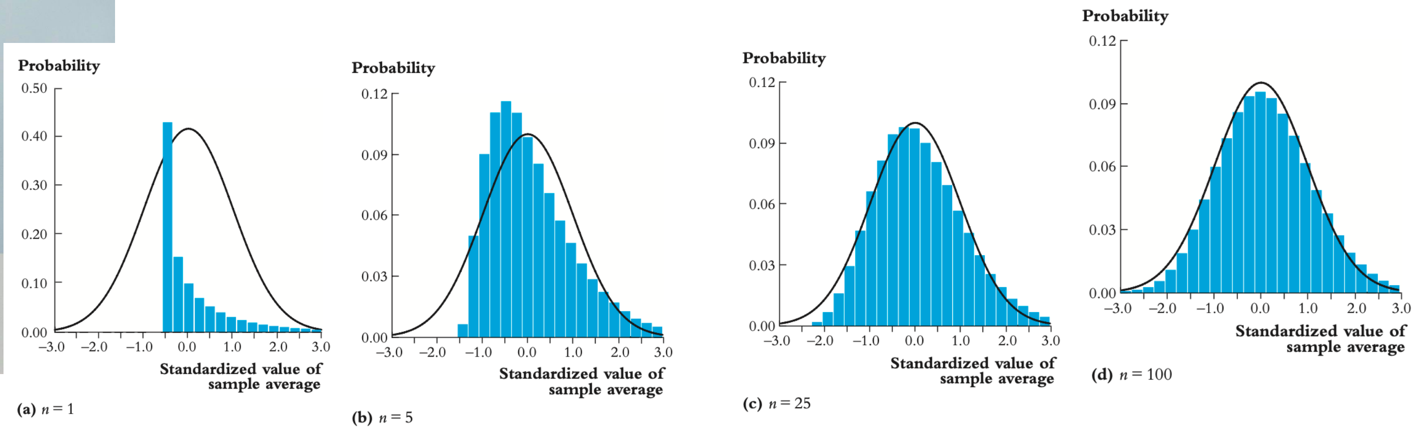
- Say $X \sim N(1,4)$
- What is the probability that $X \leq 2$?
- $\Pr(X \leq 2) = \Pr\left(Z \leq \frac{2-1}{\sqrt{4}}\right) = \Pr(Z \leq 0.5) = \Phi(0.5)$
- Look up 0.5 in the table or “display normal(0.5)” in STATA.

```
. display normal(0.5)
.69146246
```

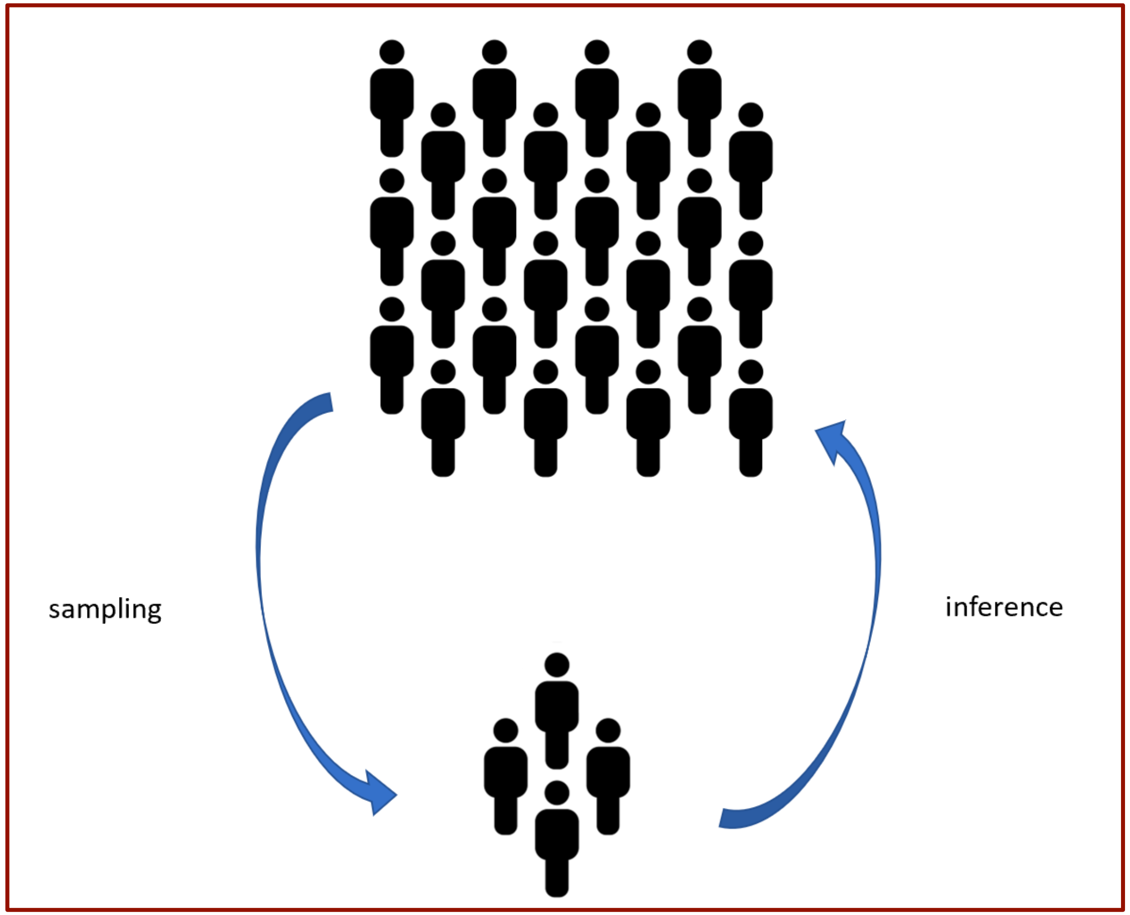
Other important distributions

- Chi-Squared distribution
- Student-t distribution
- F distribution
- …
- We don't really need to study them, at least for now.

3. The distribution of sampling statistics



Population, sample, inference



Random sampling

- *n* randomly drawn observations of Y:

$$Y_1, Y_2, Y_3, \dots, Y_n$$

- $Y_1, \dots, Y_n$ are *random variables*: different from one random sample to the next.
- $Y_1, \dots, Y_n$ are identically & independently distributed (i.i.d.).
- The sample mean

$$\bar{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n) = \frac{1}{n}\sum_{i=0}^{n} Y_i$$

is also a random variable.

- *Sampling distribution*: the probability distribution of $\bar{Y}$.

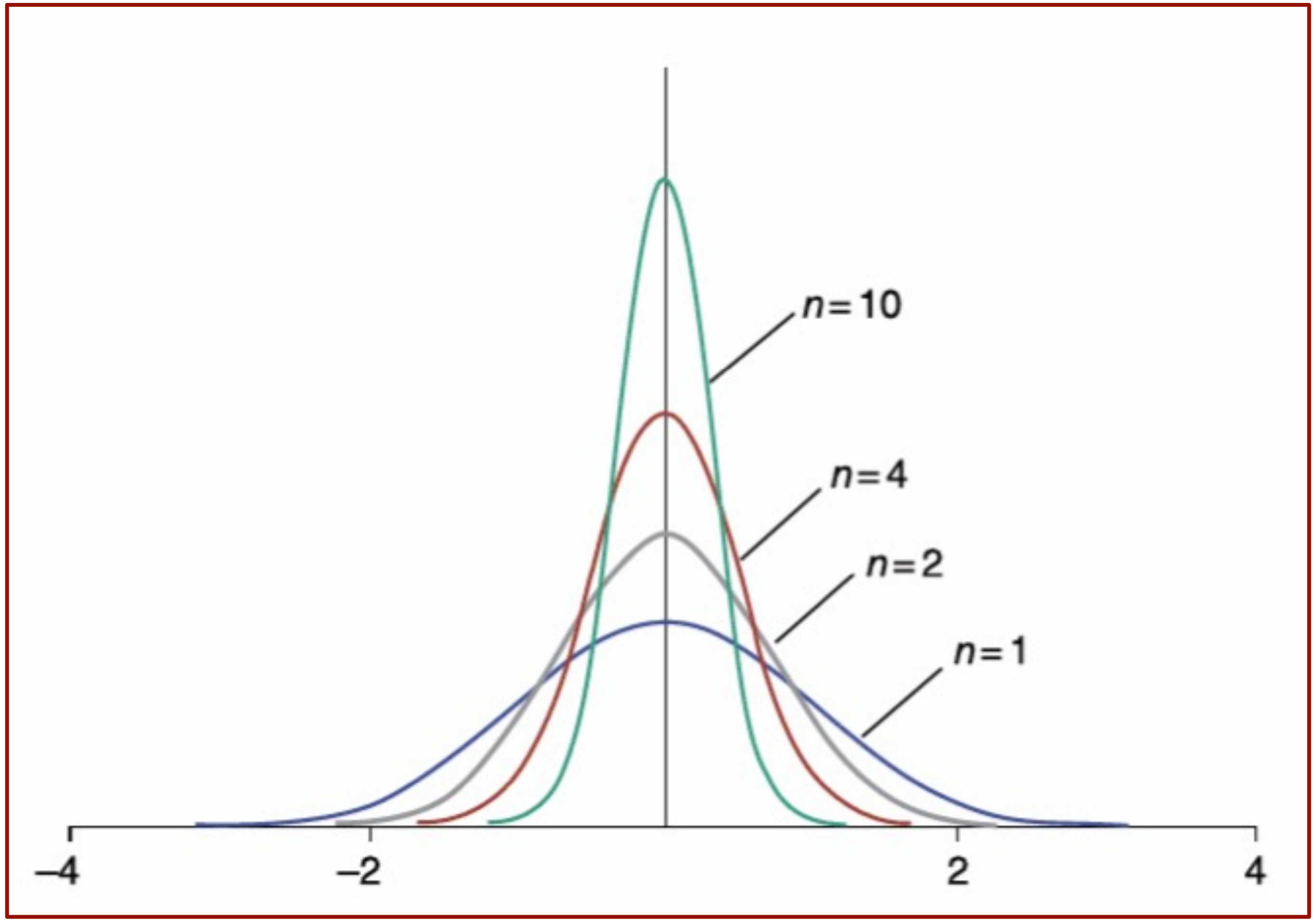
The Sampling Distribution of $\bar{Y}$

If sample observations $Y_1, \ldots, Y_n$ are i.i.d.,

- $E(\bar{Y}) = \mu_Y$
- $var(\bar{Y}) = \sigma^2_{\bar{Y}} = \frac{1}{n}\sigma^2_Y$
- $std.dev(\bar{Y}) = \sigma_{\bar{Y}} = \frac{1}{\sqrt{n}}\sigma_Y$

The law of large numbers

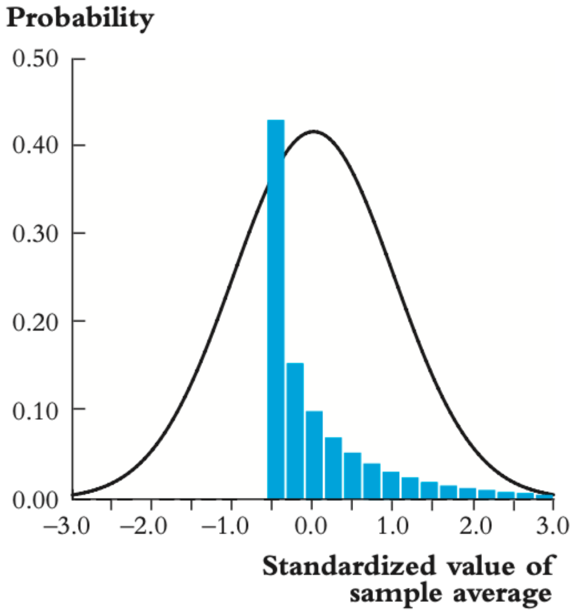
- **Law of large numbers**:
  - If n is larger, $\bar{Y}$ is more likely to be close to μ_Y.
  - $\sigma^2_{\bar{Y}}$ goes down as n increases
- → $\bar{Y}$ is **consistent** estimator of μ_Y



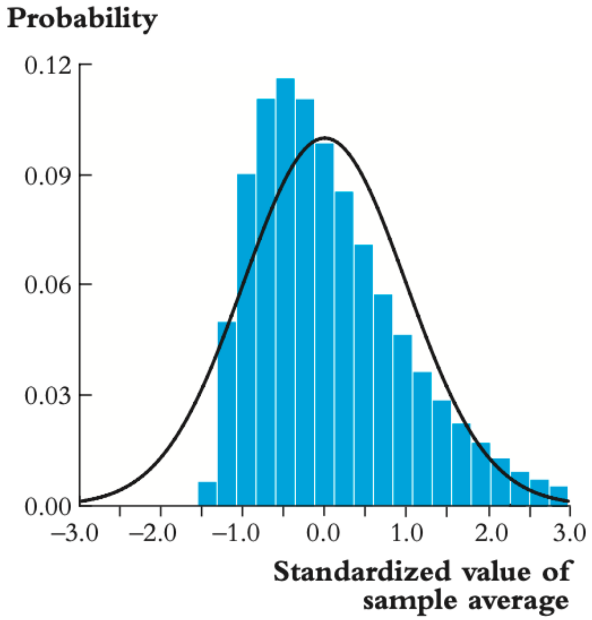
Sampling distributions in large samples

- How does the probability distribution of $\bar{Y}$ look like?
- If $Y \sim N(\mu_Y, \sigma_Y^2)$ then $\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$ *irrespective of sample size*.
- **Central Limit Theorem**:
  - When n is large, $\bar{Y}$ is (approximately) normally distributed *even if Y is not*.
  - The larger n, the closer the distribution of $\bar{Y}$ to a normal.
  - → $\bar{Y}$ is asymptotically normally distributed.

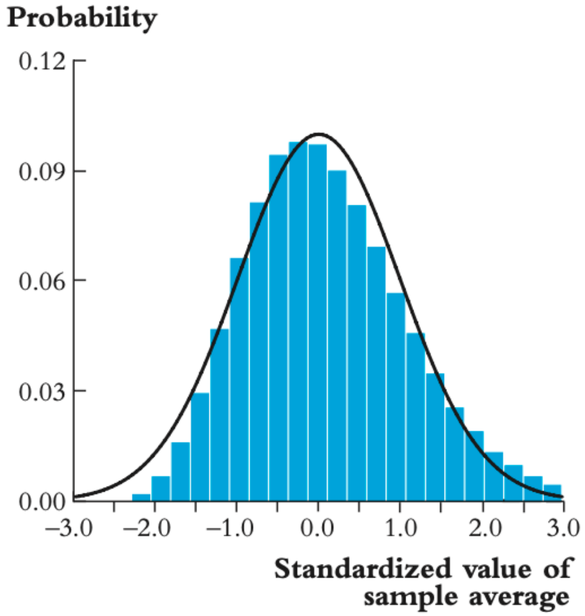
The Central Limit Theorem



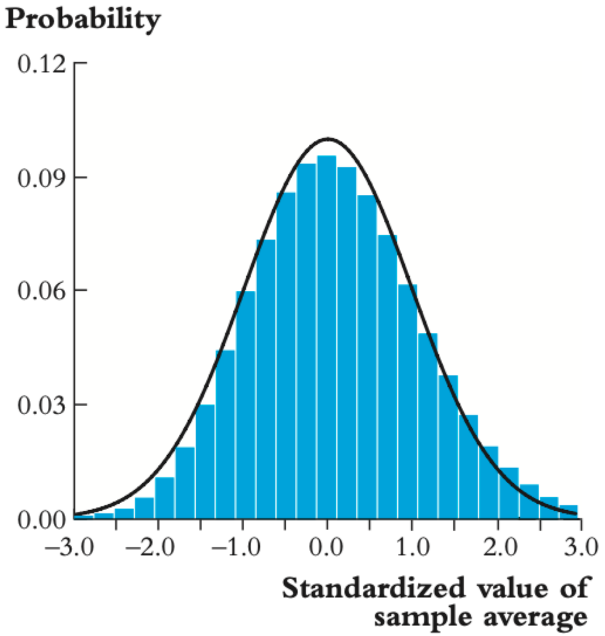
(a) $n = 1$



(b) $n = 5$



(c) $n = 25$



(d) $n = 100$

Thank you for your attention