Probability Distributions of Y and Simulated Sampling Distributions of $\bar{Y}$

In the following examples, consider a discrete random variable Y with the probability distribution $p(y)$. As usual $E(Y)=\mu_Y$ and $VAR(Y)=\sigma_Y^2$. The three examples each display tables and graphs illustrating $p(y)$, and then display simulated sampling distributions of $\bar{Y}$—the mean of an random sample of n independent observations of the random variable Y—at sample sizes of n = 5, 25, and 1000. The simulations were all constructed using the following process:

1. Specify some probability distribution $p(y)$.
2. Draw a sample of size n from the probability distribution.
3. Record the sample mean, $\bar{Y}$.
4. Repeat this process 5,000 times.[1]
5. Display a histogram of the 5,000 $\bar{Y}$'s with 10 bins.

The take-home point here: as N becomes large, the Central Limit Theorem tells us that the distribution of the sampling distribution of $\bar{Y}$ converges to the Normal with an ever-smaller variance. This is true, perhaps unsurprisingly, when the distribution of Y is itself nearly Normal (example 1). But it is also true for *any and all* possible distributions of Y, including those that are best described as "bimodal" (example 2) or skewed (example 3). Thus when n is large, no assumptions about the distribution of Y are necessary to fully describe the sampling distribution of $\bar{Y}$. Under this circumstance, $\bar{Y}$ is distributed Normal with mean μ_Y and variance σ_Y^2 / n.

[1] Note that I chose 5,000 as a large number that could nevertheless be done in a short amount of time on a standard computer. But this number doesn't matter. I could have picked 10,000 or 10 million such iterations: at higher numbers of iterations, the histograms would be smoother but would otherwise remain similar.

Example 1.

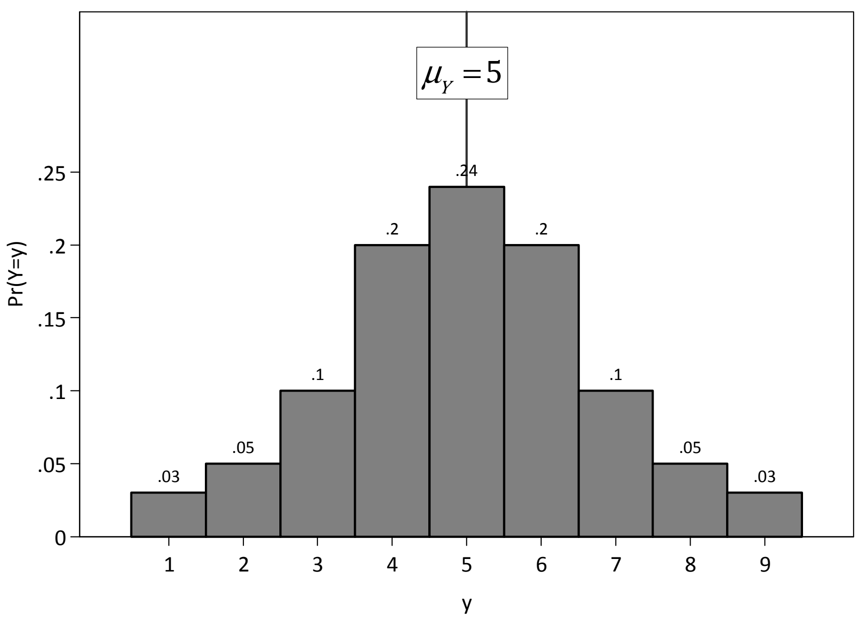
The Random Variable *Y* takes on a nearly Normal Distribution

the **probability distribution** of ***Y*** (table)

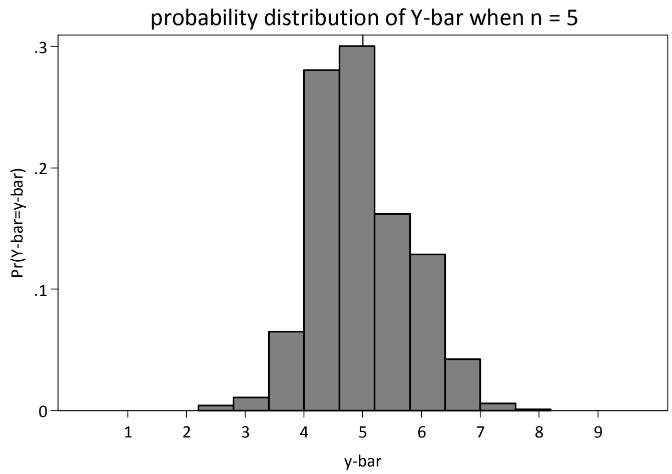
y	p(*Y*=*y*)
1	0.03
2	0.05
3	0.10
4	0.20
5	0.24
6	0.20
7	0.10
8	0.05
9	0.03
	1

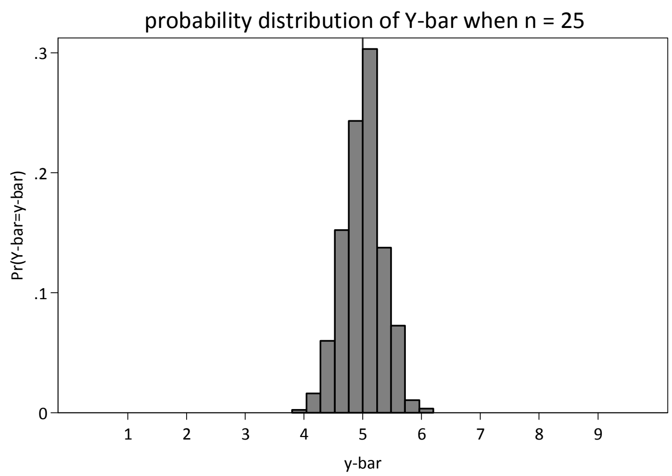
$\mu_Y = 5$
$\sigma_Y^2 \approx 3.09$

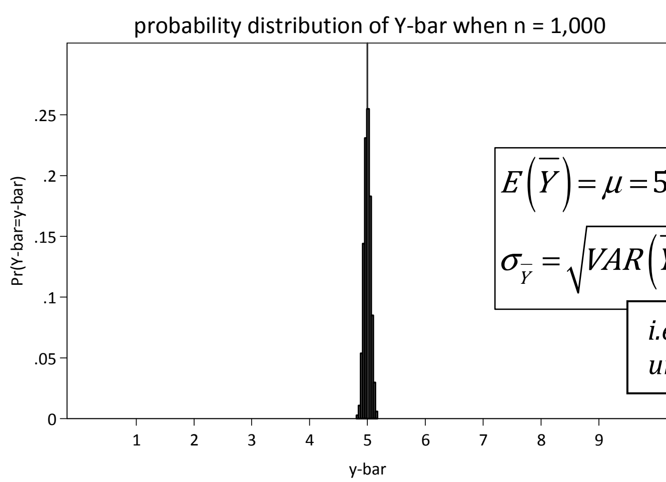
the **probability distribution** of ***Y*** (histogram)



the **sampling distribution** of ***Y*-bar** at different sample sizes







$$E\left(\overline{Y}\right) = \mu = 5$$

$$\sigma_{\overline{Y}} = \sqrt{VAR\left(\overline{Y}\right)} = \sqrt{\frac{\sigma^2}{n}} \approx \sqrt{\frac{3.09}{1{,}000}} \approx .056$$

i.e., on average, Y-bar is .056 units away from μ=5.

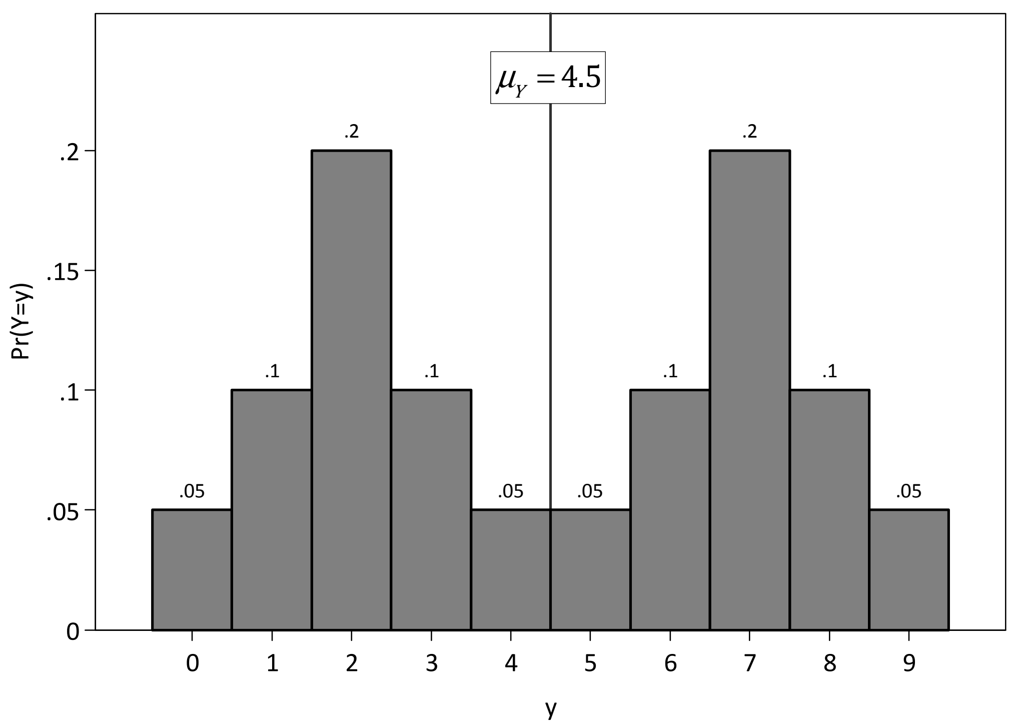
Example 2.
the Random Variable *Y* takes on a <u>Bimodal Distribution</u>

the **probability distribution** of ***Y*** (table)

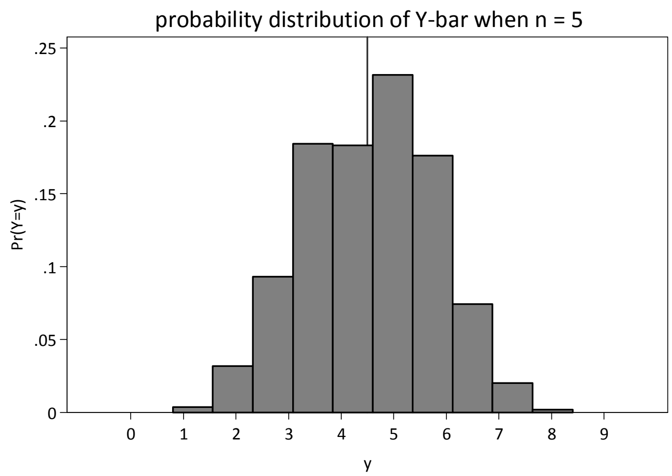
y	p(Y=y)
0	0.05
1	0.10
2	0.20
3	0.10
4	0.05
5	0.05
6	0.10
7	0.20
8	0.10
9	0.05
	1

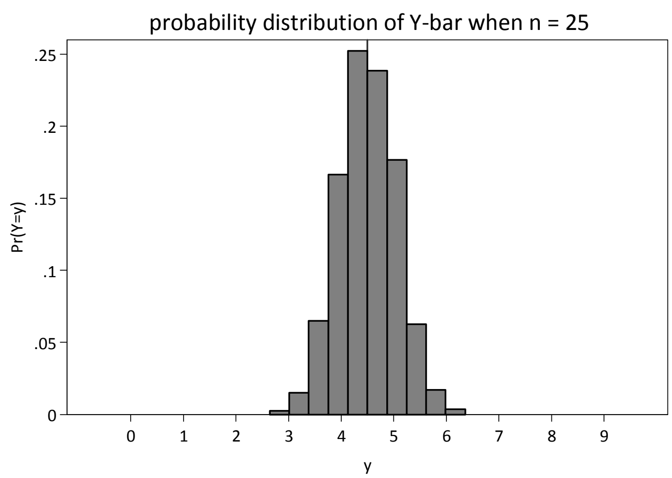
$$\mu_Y = 4.5$$
$$\sigma_Y^2 \approx 7.53$$

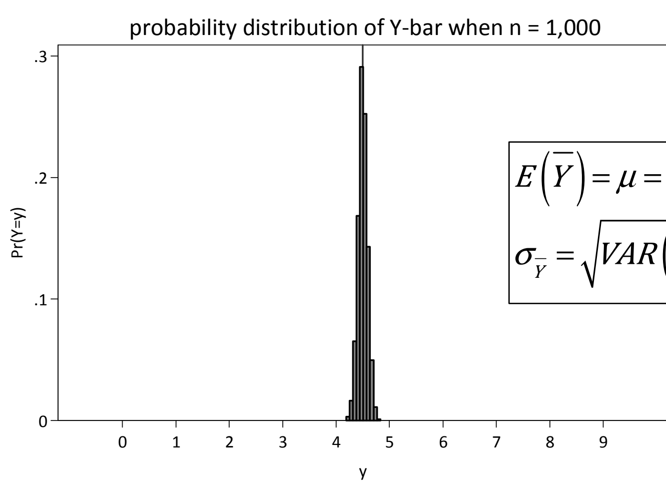
the **probability distribution** of ***Y*** (histogram)



the **sampling distribution** of ***Y*-bar** at different sample sizes







$$E\left(\overline{Y}\right) = \mu = 4.5$$

$$\sigma_{\overline{Y}} = \sqrt{VAR\left(\overline{Y}\right)} = \sqrt{\frac{\sigma^2}{n}} \approx \sqrt{\frac{7.53}{1{,}000}} \approx .087$$

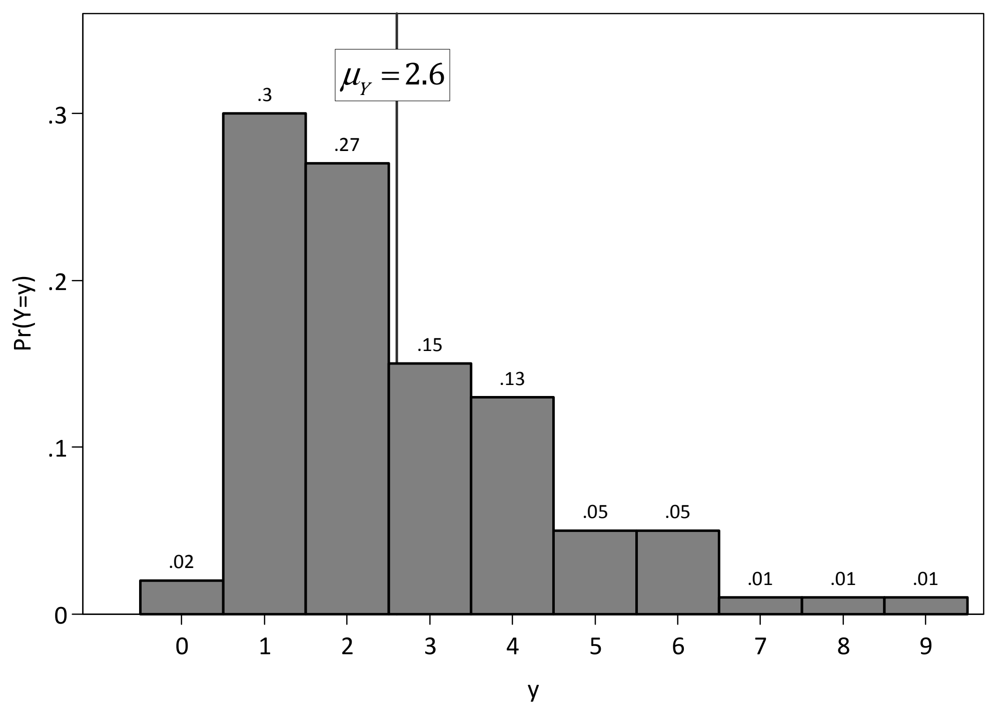
Example 3.
the Random Variable *Y* takes on a <u>Skewed Distribution</u>

the **probability distribution** of ***Y*** (table)

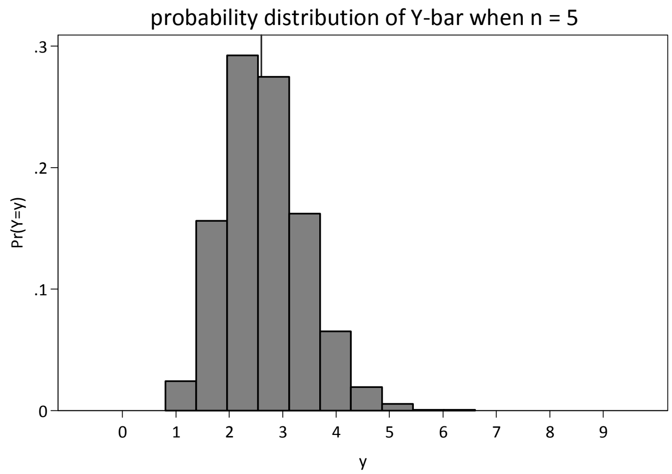
y	p(Y=y)
0	0.02
1	0.30
2	0.27
3	0.15
4	0.13
5	0.05
6	0.05
7	0.01
8	0.01
9	0.01
	1

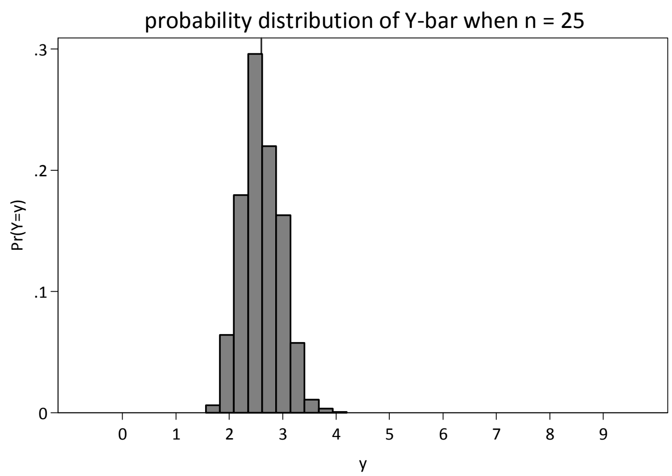
$$\mu_Y = 2.6$$
$$\sigma_Y^2 \approx 3.07$$

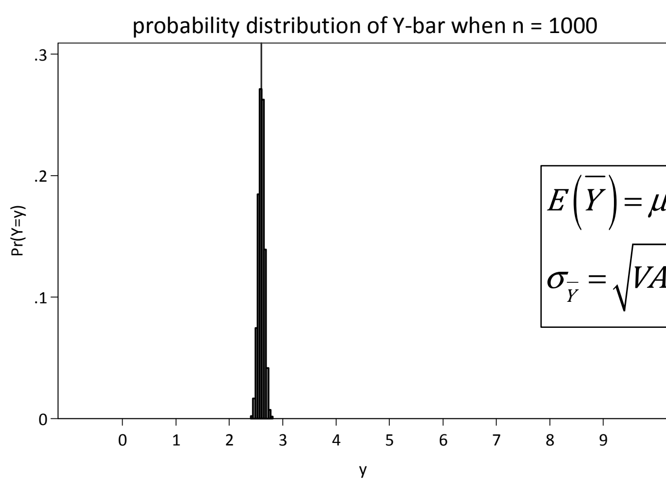
the **probability distribution** of ***Y*** (histogram)



the **sampling distribution** of ***Y*-bar** at different sample sizes







$$E\left(\overline{Y}\right)=\mu=2.6$$

$$\sigma_{\overline{Y}}=\sqrt{VAR\left(\overline{Y}\right)}=\sqrt{\frac{\sigma^2}{n}}\approx\sqrt{\frac{3.07}{1,000}}\approx.055$$