1. Sampling Distribution of a Proportion

In the last lecture we introduced the concept of a sampling distribution of sample means, i.e. the probability distribution that characterizes the fluctuation of means of samples drawn from the same population. There we considered *continuous* variables. Now we will see how the concept of a sampling distribution applies to a *binary* variable like true/false, male/female, success/failure etc.

This is called the **sampling distribution of a proportion**.

Example. A binary variable scored True/False (T/F):

Population: { T, T, F, T, F, F, T, F, T, T } (60% 'True', π = 6/10 = .6)
Sample 1: { T, T, F, F } (50% 'True', p = 2/4 = .5)

Rate of trait in *population* is π.
Rate of trait in *sample* is p.

Sampling. The repeated drawing of random samples of size n from a population. Here n = 4:

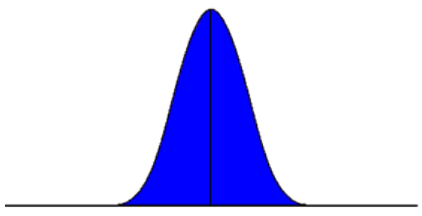
Population: { T, T, F, T, F, F, T, F, T, T} (60% 'True', π = 6/10 = .6)
Sample 1: { T, T, F, F } (50% 'True', p_1 = 2/4 = .5)
Sample 2: { F, F, T, F } (25% 'True', p_2 = 1/4 = .25)
Sample 3: { T, T, T, F } (75% 'True', p_3 = 4/4 = .75)
etc. …

Sampling distribution of a proportion

The proportion of 'Trues' changes from sample to sample. By repeated sampling, we can construct an observed distribution of sample proportions:

$$\{ p_1 \quad p_2 \quad p_3 \quad \ldots \} = \{ .5 \quad .25 \quad .75 \quad \ldots \}$$

If we considered all possible samples drawable from the population, the distribution of sample proportions (p) would be the **sampling distribution of the proportion**.



Sampling Distribution of the Proportion (*p*)

This distribution has a theoretical mean and variance/standard deviation.

The mean (expected value) for the above example would be .5, the same as the population proportion, namely π.

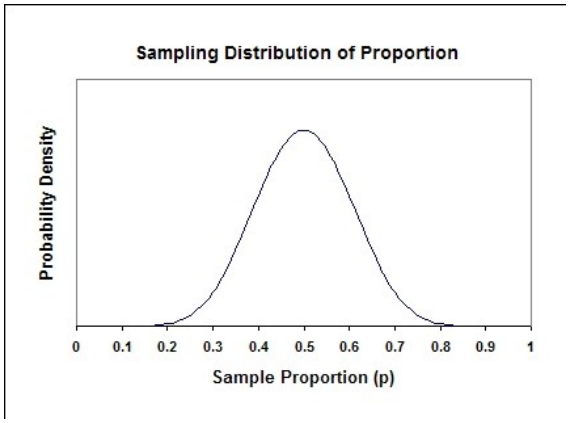
$$\mu_p = \pi$$

The standard deviation of this distribution, or the **standard error of the proportion** is

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

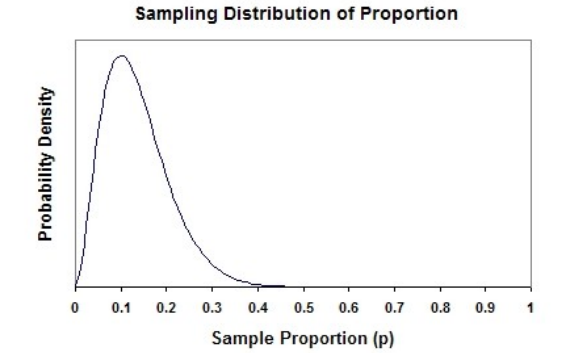
Normal Approximation

When the sample size n is sufficiently large, the sampling distribution of a proportion is well approximated by a normal (Gaussian) distribution. The usual rule-of-thumb is that $n\pi(1-\pi) \geq 5$. [However, when *p* is near 0 or 1.0, a rule of $\pi(1-\pi) \geq 105$ is safer.]



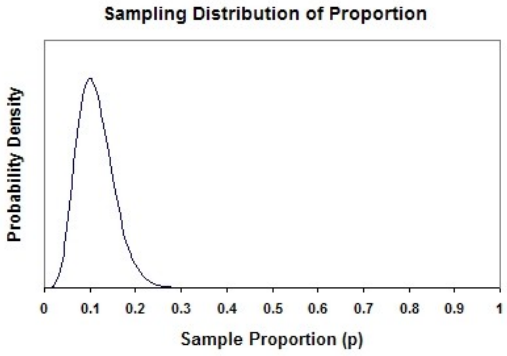
$\pi = .5$, $n = 20$, $n\pi(1-\pi) = 5$

Close to normal shape



$\pi = .1$, $n = 10$, $n\pi(1-\pi) = .9$

Very skewed



$\pi = .1$, $n = 60$, $n\pi(1 - \pi) = 5.4$

Approaching normal shape

Because of the normal approximation, we can apply what we learned earlier about calculating areas under the standard normal curve to compute cumulative probabilities to draw certain inferences about sample proportions.

Example: the population proportion (π) of CDs that fail to meet specifications is 0.10. What is the probability that, in a random sample of 400 CDs, the proportion of defective CDs will be between 0.10 and 0.12?

1. Compute z-scores corresponding to 0.10 and 0.12.

1.1 Consult z-score formula:

$$z = \frac{(X - \mu_x)}{\sigma_x}$$

1.2 Plug in appropriate values:

$$X = p, \quad \mu_x = \pi$$

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

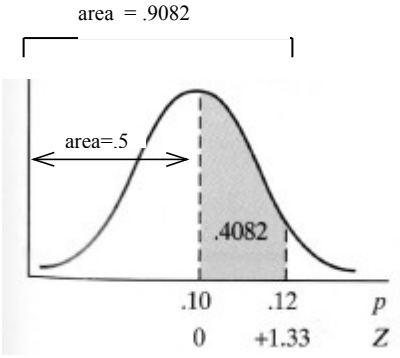
so that:

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}}$$

$$Z = \frac{0.12 - 0.10}{\sqrt{\frac{(0.10)(0.90)}{400}}} = \frac{0.02}{\sqrt{\frac{0.09}{400}}}$$

$$Z = \frac{0.02}{0.015} = 1.33$$

1.3 Determine area under standard normal curve from 0.10 to 1.33:



Pr(0.10 < p < 0.12) = area from z = 0.0 to z = 1.33

Area under standard normal curve from $-\infty$ to 1.333 is 0.9082:

Cumulative Standard Normal Distribution (z-to-p)	
NORM.S.DIST()	
z	1.3300
p (-inf to z)	0.9082

Area under standard normal curve from $-\infty$ to 0 = 0.50.

0.9032 – 0.50 = .4032.

Homework: Review lectures notes. Work sample problems. Review bullet points at end of Chapters 1 to 5.