Old Business
- variance/std. dev. of binomial distribution
- mid-term (day, policies)
- class strategies (problems, etc.)
- exponential distributions

New Business
- Central Limit Theorem, standard error of mean
- Standard error of proportions

1. Mean and Standard Deviation of Binomial Process

Clarification
These formulas cannot be used to get the mean and standard deviation of any binary variable (e.g., coded 1/2 or –1/1). Rather they give the mean and variance of the total number of 'positive' outcomes of binary variables.

Expected Value (Mean) of a Binomial Distribution

\[ \mu_x = np \]

Standard Deviation of a Binomial Distribution

\[ \sigma_x = \sqrt{np(1 - p)} \]

2. The Exponential Distribution (ctd)

Just be clear, in the last lecture we gave two formulas – one for the probability distribution of an exponential process, and one for the cumulative distribution function. These formulas look similar, so let's review them and how they differ in application.

The probability density function for the length of time between events in an exponential process is:

\[ \Pr(X = x) = f(X) = \lambda e^{-\lambda x} \]

Note that this is denoted by a lower-case \( f(x) \), which is our convention for a probability distribution. This gives the probability density at a single value of \( X = x \), e.g. the ordinate value below where \( x = 5 \):
The cumulative distribution function for an exponential process is:

$$\Pr(X \leq x) = F(x) = 1 - e^{-\lambda x}$$

Note how we denote this with a capital $F(x)$. As a cumulative function, this calculates the area under the probability distribution less than or equal to $x$ (e.g., the white-shaded region in the plot above).

**Homework**

Read pp. 209–210. Prob. 5.34. Make an Excel spreadsheet like the example.

### 3. The Central Limit Theorem, Sampling Theory, Standard Error

**Goal**

Our goal in this section is to understand and relate to each other three concepts:

- sampling distribution of the mean
- Central Limit Theorem
- Standard error of the mean

As we shall see, all three of these are integrally connected.

**Motivation**

Most of the statistical inferences we will be making from this point forward will involve the principles covered here, so it is important we understand them.

**Definitions**
Sample. A set of objects or measurements drawn at random from a larger population

Population: \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}
Sample: \{1 4 7\}

Sampling. The repeated drawing of random samples of size \(n\) from a population. Here \(n = 3\).

Population: \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}
Sample 1: \{1 4 7\} \(X_{\bar{1}} = 4\)
Sample 2: \{2 5 8\} \(X_{\bar{2}} = 5\)
Sample 3: \{7 9 10\} \(X_{\bar{3}} = 8.67\)
etc. …

Sampling distribution of the mean. Every sample has a mean. The sample mean (X-bar) changes from sample to sample:

Population: \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}
Sample 1: \{1 4 7\} \(X_{\bar{1}} = 4\)
Sample 2: \{2 5 8\} \(X_{\bar{2}} = 5\)
Sample 3: \{7 9 10\} \(X_{\bar{3}} = 8.67\)
etc. …

The sampling distribution of the mean is the probability distribution that describes the variability in the means of all samples of size \(n\) that we could draw from the population.

Probability Distribution of Sample Means

Central Limit Theorem

Central Limit Theorem. Regardless of the shape of the underlying population of the continuous variable \(X\) having mean \(\mu\), and standard deviation \(\sigma\), the sampling distribution of the mean formed by taking all possible samples of a given size \(n\) will more and more closely approximate a normal distribution with mean \(\mu_{\bar{X}} = \mu\) and standard error of the mean \(\sigma_{\bar{X}} = \sigma/\sqrt{n}\) as the sample size \(n\) increases.

The central limit theorem says, that as sample size \((n)\) becomes large, the sampling distribution will approximate a normal distribution, regardless of the
shape of the original distribution. In the example below, the population distribution is very irregular (black curve). However the sampling distribution (distribution of sample means draw from the population) starts to approximate a normal distribution, even with \( n = 5 \).

As sample size \( n \) increases, the sampling distribution of the mean will not only get close to the form of a normal distribution, it's variance will get smaller.

### Mean of the Sampling Distribution

Not surprisingly, the mean of the distribution of sample means is the same as the population mean.

\[
\mu_{\bar{x}} = \mu_x
\]

### Standard Error of the Mean

There is a very simple formula to estimate the standard deviation of sample means:

\[
\sigma_{\bar{x}} = \sigma_x/\sqrt{n}
\]

This value is known as the **standard error of the mean**. The variance of the sample mean is obtained by squaring the formula above.
4. Sampling Distribution of a Proportion

The same principles above apply to a binary variable for which we may take repeated samples from a population. For example, there is a certain proportion of males (= 1) and females (= 0) in the US population as a whole. If we take random samples of $n = 5$ people, these proportions will vary by chance alone.

- Sample 1: {1, 0, 1, 0, 0} Proportion ($p_1$) of males = .40
- Sample 2: {0, 0, 1, 1, 1} Proportion ($p_2$) of males = .60
- Sample 3: {1, 0, 0, 0, 0} Proportion ($p_3$) of males = .20

As before, this distribution of sample proportions is characterized by a probability distribution function. And again the Central Limit Theorem applies: as $n$ becomes large, the probability distribution of sample proportions will approximate a normal distribution. The standard deviation of this distribution, termed the **standard error of a proportion** is:

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

where $\pi$ is the proportion in the entire population.

In the remainder of Chapter 5 we will use what we have learned about sampling distributions and standard errors to draw inferences about population and sample means using $z$-scores computed from standard errors.