AMS 7
Sampling Distributions, Central limit theorem, Confidence Intervals
Lecture 4

Department of Applied Mathematics and Statistics, University of California, Santa Cruz

Summer 2014
Sampling Distributions!!!!!!

If we sample cookies, what is the distribution of the # of chips in a cookie? Poisson

This is the sampling distribution of the # of chips.

What if we think about the average # of chips per cookie in a bag?

Is this average the same for all bags?

No, it is random, and its distribution is derived from the sampling distribution of an individual cookie.

This is the sampling distribution of the mean
where $\bar{x}$’s are the mean number of chocolate chips in a cookie PER box, and $M$ is the total number of boxes (number of trials).

We are interested in what the distribution of the sample means looks like. $\bar{X} \sim ?$
Formal definitions:

♠ The **sampling distribution of the mean** is the probability distribution of the sample means, $\bar{x}$, with all samples having the same sample size $n$.

♠ The **sampling distribution of the proportion** is the probability distribution of the sample proportions, $\hat{p}$, with all samples having the same sample size $n$. 
The **Central Limit Theorem** tells us that the sampling distribution of the mean will be approximately normal no matter what the distribution of the individual observations is. Formally...

**Central Limit Theorem**

If samples of size $n$ are drawn from a population with mean $\mu$ and standard deviation $\sigma$, then the sampling distribution of the samples means, $\bar{x}$, will be approximately normally distributed with mean $\mu_{\bar{x}} = \mu$ and sd $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

$\sim X \sim ?$ with mean $\mu$ and sd $\sigma \Rightarrow \bar{X} \sim N\left(\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\right)$

where $\sigma_{\bar{x}}$ is called the (population) **standard error** (of the mean).
† **NOTE:** If the population of individuals is normally distributed, then $\bar{x}$ is exactly normally distributed. Otherwise $\bar{x}$ is *approximately* normally distributed, with the approximation getting better as $n$ increases.

→ In general, $n \geq 30$ is good.

☆ E.g. The population of the number of chocolate chips in a cookie follows a Poisson distribution, so the distribution of the mean number of chocolate chips in a box will be *approximately* normally distributed.

☆ E.g. The population of the length of sharks follows a normal distribution, so the distribution of the mean length of sharks is exactly normally distributed.

† Also note that as $n \to \infty$, $\sigma_{\bar{x}} \to 0$. 
**Example:** The weight of a cookie is normally distributed with mean $\mu = 11$ grams and sd $\sigma = 0.5$ grams.

- What is the distribution of the mean weight of cookies in a sample of size 32?

  \[ X \sim N(11, 0.5) \]
  \[ \Rightarrow \bar{X} \sim N\left(\mu_{\bar{X}} = \mu = 11, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{32}} = 0.088\right) \]

So \ldots

- What is the probability that the average weight of a cookie in a bag of 32 will be at least 10.8?
\[ \Rightarrow P(\bar{X} > 10.8) = P\left( \frac{\bar{X} - \mu}{\sigma} > \frac{10.8 - \mu}{\sigma} \right) \]

\[ = P\left( Z > \frac{10.8 - 11}{0.088} \right) = P(Z > -2.27) \]

\[ = 1 - P(Z < -2.27) = 1 - 0.0116 \]

\[ = 0.9884 \]

whereas the probability of selecting a SINGLE cookie from the population with weight at least 10.8 would be...

\[ \Rightarrow P(X > 10.8) = P\left( \frac{X - \mu}{\sigma} > \frac{10.8 - \mu}{\sigma} \right) \]

\[ = P\left( Z > \frac{10.8 - 11}{0.5} \right) = P(Z > -0.4) \]

\[ = 1 - P(Z < -0.4) = 1 - 0.3446 \]

\[ = 0.6554 \]
• Now, suppose the manufacturer wants to label their packages so that 99% of the bags will have an average cookie weight at least that large. What mean cookie weight should they specify?

\[ P(Z < z) = 0.01 \Rightarrow z = -2.33 \]

\[ \Rightarrow P(Z \sigma_{\bar{x}} + \mu_{\bar{x}} < -2.33\sigma_{\bar{x}} + \mu_{\bar{x}}) \]

\[ = P(\bar{X} < -2.33(0.088) + 11) \]

\[ = P(\bar{X} < 10.79) \]

\[ \Rightarrow \] Therefore, 99% of the bags will have an average cookie weight at least 10.79.

† whereas for the population of individual cookie weights, we can work out to see that 99% of the cookies will have a cookie weight of at least 9.84.
What if the manufacturer wants to claim that 95% of the bags will have an average weight in some interval? How do we find this interval? (assume symmetry)

We need to find $z$ such that $P(-z < Z < z) = 0.95$

$$P(Z < z) = \frac{1 - 0.95}{2} = 0.025$$

$\Rightarrow P(Z < -1.96) = 0.025$ (from z-table)

so, $P(-1.96 < Z < 1.96) = 0.95$

$\Rightarrow P(-1.96\bar{X} + \mu\bar{X} < Z\bar{X} + \mu\bar{X} < 1.96\bar{X} + \mu\bar{X})$

$= P(-1.96(0.088) + 11 < \bar{X} < 1.96(0.088) + 11)$

$= P(10.83 < \bar{X} < 11.17) = 0.95$

So, 95% of the bags will have an average cookie weight between 10.83g and 11.17g.
(Some more) Central Limit Theorem Examples

**Example:** IQ scores are normally distributed with a mean of 100 points and a sd of 15 points.

⇒ \( \mu = 100 \) and \( \sigma = 15 \)

1. Find the probability that a randomly selected person will have a score below 97.

\[
P(X < 97) = P \left( \frac{X - 100}{15} < \frac{97 - 100}{15} \right)
\]

\[
= P(Z < -0.2) = 0.4207
\]

2. If a random sample of 100 people is taken, find the probability that the mean score of the sample is below 97.

\[\sim\] By the CLT \( \bar{X} \sim N(\mu_{\bar{X}} = 100, \sigma_{\bar{X}} = \frac{15}{\sqrt{100}}) \):

\[
P(\bar{X} < 97) = P \left( \frac{\bar{X} - 100}{15/\sqrt{100}} < \frac{97 - 100}{15/\sqrt{100}} \right)
\]

\[
= P(Z < -2) = 0.0228
\]
Example: Suppose a batch of pepper seeds has a mean time to germination of 10.4 days with a sd of 2.13 days.

What is the probability that a random sample of 49 seeds will have a mean germination time between 10 and 11 days?

We need to find $P(10 \leq \bar{X} \leq 11)$...

$$P(10 < \bar{X} < 11) = P\left(\frac{10 - 10.4}{2.13/\sqrt{49}} < \frac{\bar{X} - 10.4}{2.13/\sqrt{49}} < \frac{11 - 10.4}{2.13/\sqrt{49}}\right)$$

$$= P(-1.31 < Z < 1.97) = P(Z < 1.97) - P(Z < -1.31)$$

$$= 0.9616 - 0.0951 = 0.8665$$
Normal Approximation to the Binomial

Recall the Binomial distribution...

- **Binomial Probability Distribution**: If the following are met
  1. Fixed number of trials, \( n \)
  2. Trials are independent
  3. Each trial has only two possible outcomes ("success", "failure")
  4. The probability of success, \( p \), is the same for each trial

where the mean is given by \( \mu = np \) and the sd is given by \( \sigma = \sqrt{np(1 - p)} \).
Although the binomial is a discrete distribution, and the normal is a continuous distribution, the normal distribution is a good approximation to the binomial distribution provided that,

\[ np \geq 5 \text{ and } n(1 - p) \geq 5, \text{ i.e. there are (on average) at least 5 successes and 5 failures} \]

Sort of like averaging over coin flips, so CLT applies.

We approximate a binomial distribution with \( n \) and \( p \), \( Bin(n, p) \), with a normal distribution having mean \( \mu = np \) and sd \( \sigma = \sqrt{np(1 - p)} \):

\[ Bin(n, p) \approx N(\mu = np, \sigma = \sqrt{np(1 - p)}) \]

Note that because the binomial is discrete, but the normal is continuous, we typically use a continuity correction, moving the count by 1/2 such that the inequality still holds.
**Example:** A study found that 62% of the households in Alaska have a computer. If we take a random sample of 1000 Alaskan households, what is the probability that at least 640 have a computer?

Basic idea: \( X = \# \) households in the sample with a computer.

\[ X \sim Bin(n = 1000, p = 0.62) \]

\[ \Rightarrow \mu = 1000(0.62) = 620, \sigma = \sqrt{1000(0.62)(0.38)} = 15.35 \]

So, \( X \approx N(\mu = 620, \sigma = 15.35) \)

† Detail: \( X \) is discrete, so you could never get \( X = 639.6 \) or \( X = 639.78 \).

\[ \Rightarrow \text{Continuity Correction: } P(X \geq 640) = P(X > 639.5) \]

\[ \Rightarrow P \left( \frac{X - 620}{15.35} > \frac{639.5 - 620}{15.35} \right) = P(Z > 1.27) = 1 - P(Z < 1.27) \]

\[ = 1 - 0.8980 = 0.1020 \]
• How about a range? What is the probability of the number of households in the sample having a computer being between 615 and 625 non-inclusive?

\[
P(615 < X < 625) = P(615.5 < X < 624.5)
\]

\[
= P \left( \frac{615.5 - 620}{15.35} < \frac{X - 620}{15.35} < \frac{624.5 - 620}{15.35} \right)
\]

\[
= P(Z < 0.29) - P(Z < -0.29)
\]

\[
= 0.6141 - 0.3859
\]

\[
= 0.2282
\]
Recall the example of finding an interval such that the mean cookie weight of the cookies in a randomly chosen bag of cookies has probability 0.95 of being in that interval.

We want to find \( x_1 \) and \( x_2 \) such that \( P(x_1 < \bar{X} < x_2) = 0.95 \).

Instead we can find \( z \) such that \( P(-z < Z < z) = 0.95 \).
From z-table we get \( P(-1.96 < Z < 1.96) = 0.95. \)

\[ Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \quad \Rightarrow \quad \bar{X} = Z\sigma_{\bar{X}} + \mu_{\bar{X}} \]

\[ \Rightarrow \quad P(-1.96\sigma_{\bar{X}} + \mu_{\bar{X}} < Z\sigma_{\bar{X}} + \mu_{\bar{X}} < 1.96\sigma_{\bar{X}} + \mu_{\bar{X}}) \]

\[ = P(-1.96(0.088) + 11 < \bar{X} < 1.96(0.088) + 11) \]

\[ = P(10.83 < \bar{X} < 11.17) = 0.95 \]
♠ What if we don’t know the true population mean of a cookie, and want to learn it from the data (measurements of the sample)?

★ Inference: Our best guess (point estimate) of the population mean $\mu$ is the sample mean $\bar{x}$.

→ How good do we think this guess is? It depends on the data:

- How much data?
- How were they collected?
- How much variability in the data?

★ Example: Assume we know that the standard deviation of the weight of a cookie is 0.5g, but we don’t know the mean weight of a cookie. We get a bag (sample) of 32 cookies and find the average weight is 10.9g.

⇝ How confident are we in this estimate?

⇝ What would be an interval of plausible values?
→ Assuming cookie weights are approximately normally distributed (or using the CLT for the mean with $n \geq 30$), that $\sigma = 0.5$ is known and that we have a simple random sample, starting with a standard normal, **before** we observe $\bar{X}$ we know

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

$$P(-1.96 \leq Z \leq 1.96) = P \left( -1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96 \right)$$

$$= P \left( -1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96 \right) = P \left( -1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

$$= P \left( 1.96 \frac{\sigma}{\sqrt{n}} \geq \mu - \bar{x} \geq -1.96 \frac{\sigma}{\sqrt{n}} \right) = P \left( -1.96 \frac{\sigma}{\sqrt{n}} \leq \mu - \bar{x} \leq 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

$$= P \left( \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right) = 0.95$$

★★ So, the **random** interval $\left( \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$ has a 95% probability of containing the true population mean, $\mu$!!

★★ 95% of the intervals constructed this way will contain the true population mean $\mu$!
• An axiom (assumption) of Frequentist Statistics is that unknown parameters have a fixed, right answer. So once we observe $\bar{x}$, the interval is fixed, and the value of $\mu$ is fixed. So $\mu$ is either in the interval, or it isn’t, but we don’t know which.

★ The interval we get when we use the observed $\bar{x}$ is called a confidence interval.

★ Example: For a sample bag of 32 cookies we have: $\bar{x} = 10.9$g and $\sigma = 0.5$g. Find the 95% CI for the mean weight of all cookies.

$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} = 10.9 - 1.96 \frac{0.5}{\sqrt{32}} = 10.73$

$\bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} = 10.9 + 1.96 \frac{0.5}{\sqrt{32}} = 11.07$

So $(10.73, 11.07)$ is a 95% CI for $\mu$.
★ Technical Interpretation: We are 95% confident that $\mu$ is in this interval.
• On average 95% of the intervals constructed this way will contain $\mu$, but we don’t know if this particular interval does contain it or not.

• Note that this is not a probability. The probability that $\mu$ is in this particular interval is 0 or 1, but we don’t know which.

• CI is an interval estimate for $\mu$. It provides a range of plausible values for $\mu$. 
More general form:

$$(1 - \alpha) \text{CI} \text{ is } \bar{x} \pm E, \text{ where } E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = \text{margin of error.}$$

$-z_{\alpha/2}$ is the $\frac{\alpha}{2}$ quantile of the standard normal distribution, i.e.

$$P(Z \leq -z_{\alpha/2}) = \frac{\alpha}{2}.$$
**Example:** Suppose a soda distributor is filling 20oz bottles and that from historical data, the sd of the contents of a bottle is known to be 0.03oz. Is the right amount of soda going into each bottle? Suppose a random sample of 34 bottles is found to have an average of 19.98oz. Find a 90% CI for the population mean contents.

\[
1 - \alpha = 0.9 \Rightarrow \alpha = 0.1 \Rightarrow \frac{\alpha}{2} = 0.05 = P(Z < -1.645) \text{ (z-table)}
\]

So, \( z_{\alpha/2} = 1.645 \) and \( E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1.645 \frac{0.03}{\sqrt{34}} = 0.0085 \)

\[
\bar{x} \pm E = 19.98 \pm 0.0085 = (19.9715, 19.9885).
\]

So (19.9715, 19.9885) is 90% CI for \( \mu \).

- Is it reasonable that 20oz are going to each bottle?
How do we determine the sample size needed for a desired margin of error?

**Example continued:** Suppose we want to be able to estimate soda contents with a margin of error of 0.001.

\[
0.001 = E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1.645 \frac{0.03}{\sqrt{n}}
\]

\[
\Rightarrow \sqrt{n} = \frac{1.645 \cdot 0.03}{0.001} = 49.35 \Rightarrow n = (49.35)^2 = 2435.42
\]

* We need to round up, so that the margin of error is no larger than specified, so \( n = 2436 \).

**In general,**

\[
n = \left( \frac{Z_{\alpha/2} \sigma}{E} \right)^2
\]

- **Note:** the sample size increases **rapidly** as the margin of error reduces.
Key Concepts!!!!!

- **Sampling Distribution of the mean**
- **Sampling Distribution of the proportion**
- **Central Limit Theorem**
- **Normal Approximation to the Binomial**
- **Confidence Interval**