Lecture 7
Homework Update

Summary of Previous Lecture

Sampling Distributions
  Introduction
  Binomial Distribution
  Sample Distribution of the Mean
Part I

Homework Update
Issues when turning in SAS Homework

- Use proper variable names and units in tables and figures.
- Not all variables can be summarized the same way.
- Do **not** hand in a big pile of computer output; cut and paste *only* what is essential (typically a few figures and/or tables) from the output to the pages that you turn in.
Every Thursday from now on, you will have a homework due. The only exceptions will be:

- **Thurs Nov 23**: No homework due since this is Thanksgiving Day (no school). Indeed there are NO LAB SESSIONS THIS WEEK (so even if your lab is on Tuesday or Wednesday, you don’t have a lab this week).

- **Thurs Dec 14**: No homework due since this is after the semester has ended, and the final exam is the very next day. There are NO LAB SESSIONS THIS WEEK EITHER (so even if your lab is on Monday or Tuesday, you don’t have a lab this week either).
Part II

Review of Previous Lecture
Expectations (mean and variance) are properties of the distribution.

The average of $X$ and $Y$ has the same mean but half the variance of $X$.

Statistics are functions of data values, and are often used to estimate parameters, which are underlying theoretical characteristics of distributions (like the population mean $\mu$ and population variance $\sigma^2$).

Bayes’ Rule is helpful in reversing the order of conditioning in certain conditional probability problems.
Part III

Sampling Distributions
Flow Chart in Statistical Inference

- Class of Models
- Data
- Estimation
- Testing
- Model
- Interpretaion
- Decision
- Prediction
Definition
The probability distribution of a statistic (say, the sample mean $\bar{X}$) is called the *sampling distribution* of the statistic.

We would like to relate the variability of a statistic based on a random sample to the variability of the random variable on which the random sample is based.
Definitions

- Recall that a *Bernoulli* experiment has only two outcomes: success (1) and failure (0), with the probability of success denoted by $p$.

- When $n$ identical (same $p$) and independent Bernoulli experiments are conducted, the total number of successes $S$ has a *Binomial* distribution with parameters $n$ and $p$, denoted by $S \sim \mathcal{B}(n, p)$, or $S \sim \text{Bin}(n, p)$.

- Let $X_i$ denote the outcome of the $i$th Bernoulli experiment, then $S = \sum_{i=1}^{n} X_i$ is a statistic!
Binomial Distribution: Characteristics

- A response or trait takes on one and only one of precisely two possibilities.
- The response is observed a \textit{predetermined, known} number of times \((n)\).
- The probability of a success \((p)\) is the same for every trial \((\textit{identical} \text{ trials})\).
- The outcome of one trial is not influenced by the outcome of the other trials \((\textit{independent} \text{ trials})\).
Definition
When sampling is *with replacement*, after a sample is drawn from the population and recorded, it is put back into the population for possible resampling.
When sampling 8 balls from these 10 with replacement, the total number of black balls has a $B(8, 0.4)$ distribution.
Sampling Without Replacement

**Definition**
When sampling is *without replacement*, after a sample is drawn from the population and recorded, it is permanently removed from the population.
Sampling Without Replacement

When sampling 8 balls *without* replacement, the total number of black balls has a non-Binomial distribution.

To see this, consider \( \Pr(\text{black on draw 8} \mid 4 \text{ blacks drawn so far}) \)

But what if there were 10,000 balls of which 4000 were black?
When the population size is much larger than the sample size (say, more than 10 times as large), the number of a particular type in an SRS of size $n$ has \textit{approximately} a $\mathcal{B}(n, p)$ distribution, where $p$ is the population proportion of that type.
A binomial table gives the probabilities of various binomial events given particular $n$ and $p$ values.

A binomial table only lists values of $p \leq 0.5$, since if the probability of success of interest is greater than 0.5, we can just count the number of failures instead.

Example

For each individual there is a 45% chance a certain surgery will be successful. What is the probability of at most 6 successes out of twenty independent surgeries of this type?
Given $X \sim \mathcal{B}(20, 0.45)$, we seek values of the probability mass function (pmf; e.g., $\Pr(X = 6)$) and the cumulative distribution function (cdf; e.g., $\Pr(X \leq 6)$).

- **Excel:** `binomdist (6, 20, 0.45, 0)` and `binomdist (6, 20, 0.45, 1)`.

- **SAS:**

```sas
DATA binom;
    a = PMF ('Binomial', 6, 0.45, 20);
    b = CDF ('Binomial', 6, 0.45, 20);
PROC PRINT data = binom;
RUN;
```
Fact

If $S$ has a Binomial distribution $\mathcal{B}(n, p)$, then

$$E[S] = \mu_S = np$$
$$\text{Var}[S] = \sigma^2_S = np(1 - p)$$

$$= npq \text{ where } q = 1 - p$$

*Remember:* The expectation of a random variable is the same as its population mean.
Could you derive these formulae given what we know about means and variances of sums of independent random variables?...
Example

Helsinki Heart Study (MM p.372).
Recall that for a Bernoulli variable $X$ with success probability $p$, $E[X] = p$ and $\text{Var}[X] = p(1 - p)$.

Since $S = \sum_{i=1}^{n} X$, we can derive its mean and variance by rules of expectation.
Note that $S$ is the sum of $X_i$. Suppose we instead consider the mean of $X_i$, i.e. the *sample proportion*,

$$\hat{p} \equiv \frac{S}{n}.$$ 

Then

$$E[\hat{p}] = p,$$

$$\text{Var}[\hat{p}] = \frac{p(1-p)}{n}.$$ 

Could you derive these formulae given what we now know about the mean and variance of the binomial sum $S$?...
Unbiasedness

Definitions

- An *estimator* is a statistic that is used to estimate a population/distribution parameter. Its value is often called an *estimate*.
- If the expectation of an estimator is the same as the population quantity it is trying to estimate, we say that the estimator is *unbiased*.

Example

Since we have just shown that $E(\hat{p}) = p$, this means that the sample proportion $\hat{p}$ is an unbiased estimator of the true population proportion $p$. 
Fact
When the sample size $n$ gets larger, the variance of the sample mean $\bar{X}$ becomes smaller. In fact, as $n$ tends to infinity, $\text{Var}(\bar{X})$ goes to 0.

This is a general property of sample means, and a manifestation of the Law of Large Numbers.
Approximations

For a binomial variable $S \sim \mathcal{B}(n, p)$, when $n$ is large,

$$S \approx \mathcal{N} \left( \mu = np, \sigma = \sqrt{np(1 - p)} \right),$$

and

$$\hat{p} \equiv \frac{S}{n} \approx \mathcal{N} \left( \mu = p, \sigma = \sqrt{\frac{p(1 - p)}{n}} \right).$$

These approximations work well provided $np \geq 10$ and $n(1 - p) \geq 10$. 
Normal Approximation Applet

http://www.stat.vt.edu/~sundar/java/applets/Distributions.html
Fact

Let $\bar{X}$ be the sample mean of $n$ independent and identical random variables $X_1, \ldots, X_n$. If the mean and standard deviation of $X_i$ are $\mu$ and $\sigma$, then the mean and standard deviation of $\bar{X}$ are

$$
\mu_{\bar{X}} = \mu,
$$

$$
\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.
$$
Fact

If $X_i$ has $\mathcal{N}(\mu, \sigma)$ distribution, then $\bar{X}$ has $\mathcal{N}(\mu, \sigma/\sqrt{n})$ distribution.

Note no approximation here: If the $X_i$ start out normal, then $\bar{X}$ is also exactly normal.
Central Limit Theorem

**Theorem**

Let $\bar{X}$ be the sample mean of $n$ independent and identically distributed (iid) random variables $X_1, \ldots, X_n$ having mean $\mu$ and standard deviation $\sigma$. When $n$ is large, the sampling distribution of $\bar{X}$ is approximately normal:

$$\bar{X} \overset{\text{approx}}{\sim} \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}).$$

Recall Brad’s symbolic version:

$$X_i \overset{iid}{\sim} ? (\mu, \sigma) \implies \bar{X} \overset{\text{approx}}{\sim} \mathcal{N}(\mu, \sigma / \sqrt{n})$$
Central Limit Theorem Applet

http://www.stat.vt.edu/~sundar/java/applets/CLT.html
How large is large?

**Rule of Thumb:** The normal approximation in the CLT is good provided \( n \) is larger than 25 or 30.

**Example**

Exercise 5.29 (MM p.402)
Exercise 5.35 (MM p.404)