Final Review

What to do:

- Read lectures 1-9, chapters 1-20 and 23 from the textbook. (Note: In chapter 20 read about hypotheses testing and $p$-value. You won't be tested on $z$-test for proportions, but you need to know how to get a confidence interval for $p$)

- Go over the midterm review

- Do the assigned exercises from the textbook

- Go over the quiz questions

- Use sample exams to practice

- Use extra TAs' office hours
Topics to review:

- Data collection

  - Observational study: retrospective or prospective
  - Experiment: randomized, comparative, single- or double-blind, placebo-controlled
  - Design of Experiment: subjects, factors, treatments; matched pairs, blocking
  - Randomization (use of random digits)
  - Sampling design
    - Population vs sample (parameter vs statistic)
    - SRS
    - Stratified random sample
    - Cluster sampling
    - Multistage sampling
    - Systematic sampling
Example: A maker of fabric for clothing is setting up a new line to “finish” the raw fabric. The line will use either metal rollers or natural-bristle rollers to raise the surface of the fabric; a dyeing cycle time is either 30 min or 40 min; and a temperature of either $150^\circ C$ or $175^\circ C$. An experiment will compare all combinations of these choices. Four specimens of fabric will be subjected to each treatment and scored for quality.

What are the factors and treatments? How many individuals does the experiment require? Outline a completely randomized design for this experiment.

3 factors:
- Roller type
  - Metal
  - Natural-bristle
- Cycle time
  - 30 min
  - 40 min
- Temperature
  - $150^\circ C$
  - $175^\circ C$

$2^3 = 8$ treatments

Thus, $4 \times 8 = 32$ individuals
• Probability

  o Probability model
  o Sample space
  o Event, probability of event
  o Probability rules
  o Venn diagrams
  o Independent events
  o Conditional probability
  o Tree diagrams
Example: Spelling errors in a text can be either nonword errors or word errors. Nonword errors make up 25% of all errors. A human proofreader will catch 90% of nonword errors and 70% of word errors. What percent of all errors will the proofreader catch? (Draw a tree diagram to organize the information given.)

\[
P(\text{caught nonword}) + P(\text{caught word})
= 0.25 \times 0.9 + 0.75 \times 0.7 = 0.75
\]
Example: Here are the projected numbers (in thousands) of earned degrees in the U.S. in the 2010-2011 academic year, classified by level and by the sex of the degree recipient.

<table>
<thead>
<tr>
<th></th>
<th>Bachelor’s</th>
<th>Master’s</th>
<th>Professional</th>
<th>Doctorate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>933</td>
<td>402</td>
<td>51</td>
<td>26</td>
</tr>
<tr>
<td>Male</td>
<td>661</td>
<td>260</td>
<td>44</td>
<td>26</td>
</tr>
</tbody>
</table>

(a) If you choose a degree recipient at random, what is the probability that the person you choose is a woman?

\[ P(\text{Female}) = \frac{1412}{2403} = 0.5876 \]
<table>
<thead>
<tr>
<th></th>
<th>Bachelor’s</th>
<th>Master’s</th>
<th>Professional</th>
<th>Doctorate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>933</td>
<td>402</td>
<td>51</td>
<td>26</td>
<td>1412</td>
</tr>
<tr>
<td>Male</td>
<td>661</td>
<td>260</td>
<td>44</td>
<td>26</td>
<td>991</td>
</tr>
<tr>
<td>Total</td>
<td>1594</td>
<td>662</td>
<td>95</td>
<td>52</td>
<td>2403</td>
</tr>
</tbody>
</table>

(b) What is the conditional probability that you choose a woman given that the person chosen received a professional degree?

\[
P(\text{Female} | \text{Prof.}) = \frac{51}{95} = 0.5368
\]

(c) Are the events “choose a woman” and “choose a professional degree recipient” independent?

No, since \( P(\text{Female} | \text{Prof.}) \neq P(\text{Female}) \)
Example: The table below describes the smoking habits of a group of asthma sufferers.

<table>
<thead>
<tr>
<th></th>
<th>Non-smoker</th>
<th>Occasional Smoker</th>
<th>Regular Smoker</th>
<th>Heavy Smoker</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>344</td>
<td>40</td>
<td>85</td>
<td>39</td>
<td>508</td>
</tr>
<tr>
<td>Women</td>
<td>370</td>
<td>31</td>
<td>88</td>
<td>48</td>
<td>537</td>
</tr>
<tr>
<td>Total</td>
<td>714</td>
<td>71</td>
<td>173</td>
<td>87</td>
<td>1045</td>
</tr>
</tbody>
</table>

If one of these 1045 people is randomly selected, find the probability that the person is a man or a heavy smoker.

A) 0.448  
B) 0.569  
C) 0.495  
D) 0.532  

\[
P(\text{Man or } H_s) = P(\text{Man}) + P(H_s) - P(\text{Man and } H_s)
\]

\[
= \frac{508}{1045} + \frac{87}{1045} - \frac{39}{1045} = 0.532
\]
Random variables:

- Discrete
  - Probability distribution
  - Mean and variance of a discrete random variable

**Example:** Let $X_1$ and $X_2$ be independent and have the same distribution given below:

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>-1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_i$</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Let $X = X_1 + X_2$. Find the probability distribution of $X$, $\mu_X$ and $\sigma_X$.

\[
X_1 + X_2 = \begin{array}{ccc}
-2 & -1 & 0 \\
0.7 & 0.7 & 0.3 & 0.3 \\
0.49 & 0.92 & 0.09
\end{array}
\]

\[
\begin{align*}
\mu_X &= -2 \cdot 0.42 + (-1) \cdot 0.92 + 0.009 = -1.4 \\
\sigma_X &= \sqrt{(-2+1.4)^2 \cdot 0.49 + (-1+1.4)^2 \cdot 0.92 + (0+1.4)^2 \cdot 0.09} = 0.65
\end{align*}
\]
Binomial distribution

\[ X \sim \text{Bin} (n, p), \; X = 0, 1, \ldots, n \]

\[ \mu_X = np \quad \sigma_X = \sqrt{np(1 - p)} \]

Example: Suppose that 80% of adults with allergies report symptomatic relief with a specific medication. If the medication is given to 10 new patients with allergies, what is the probability that it is effective in exactly seven?

\[ X = \# \text{ of patients for who it was effective} \]

\[ X \sim \text{Bin} (10, 0.8) \]

\[ P(X = 7) = 0.2013 \quad (table) \]
Continuous
  - Density curve
  - Normal distribution
  - Normal approximation (sampling distribution for counts and sample proportions)

Let $X$ be the count of successes in the sample and $\hat{p} = X/n$ be the sample proportion of successes. When $n$ is large, the sampling distributions of these statistics are approximately Normal:

$$X \text{ is approximately } N(np, \sqrt{np(1-p)})$$

$$\hat{p} \text{ is approximately } N(p, \sqrt{\frac{p(1-p)}{n}})$$

As a rule of thumb, we will use this approximation for values of $n$ and $p$ that satisfy $np \geq 10$ and $n(1 - p) \geq 10$. 
Sampling distribution of sample mean (CLT):
Draw an SRS of size \( n \) from any population with mean \( \mu \) and finite standard deviation \( \sigma \). When \( n \) is large enough,
\[
\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})
\]

Example: 20% of customers at a bakery will buy a brownie.

500 customers arrive at the bakery in a day. Assume that individual customers make their purchases independently.

What is the probability that more than 110 customers buy a brownie?

\[
X = \# \text{ of customers who buy a brownie}
\]

\[
X \sim \text{Bin}(500, 0.2), \ n p = 100 > 10, \ n q = 400 > 10
\]

\[
X \sim \text{appr. } N(100, \sqrt{500 \cdot 0.2 \cdot 0.8}) = N(100, 180)
\]

\[
P(X > 110) = P(Z > \frac{110 - 100}{180}) = P(Z > 1.12)
\]

2-score for 110

\[
= 0.1219
\]
• Statistical inference

  o Confidence intervals
    ▪ For population mean $\mu$: $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ if $\sigma$ is unknown
    ▪ For single proportion $p$: $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$

  o Hypothesis testing
    ▪ Test statistic

  ➢ To test $H_0: \mu = \mu_0$ we find statistic $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ ($\sigma$ is unknown)

    ▪ P-value

    ▪ Significance level