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)

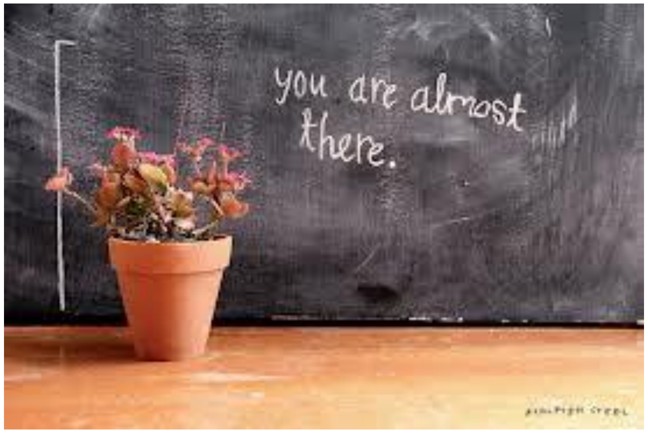
- Go over the midterm review
  + midterm test

- 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 150^0C or 175^0C. 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 $\to$ metal, natural-bristle

cycle time $\to$ 30 min, 40 min

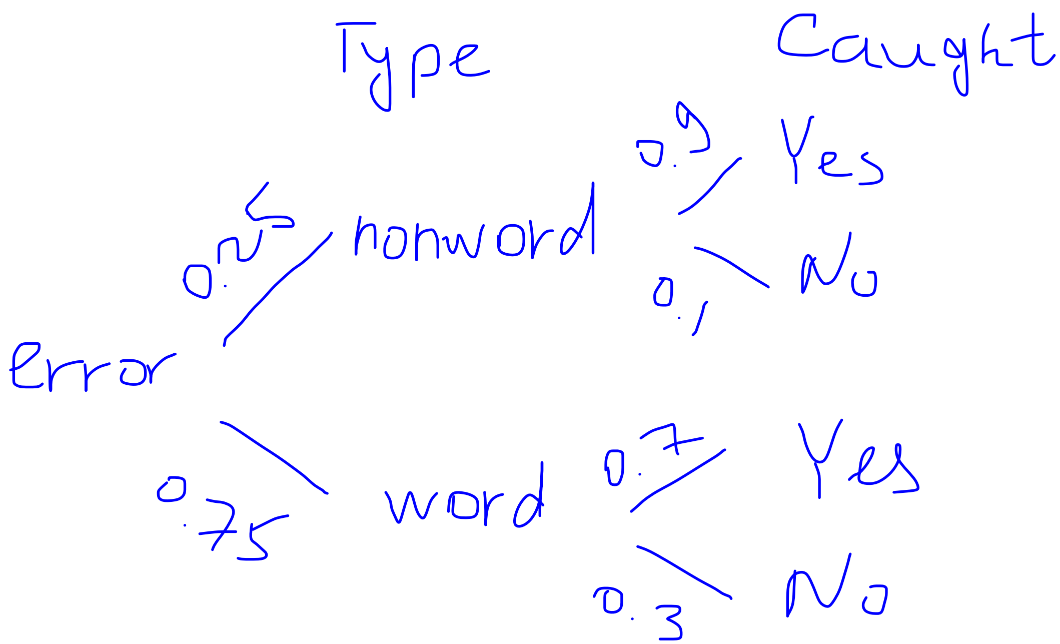
Tem-re $\to$ $150^\circ C$, $175^\circ C$

$2^3 = 8$ treatments

Thus, $4 \cdot 8 = 32$ individuals

- **Probability**
    - Probability model
    - Sample space
    - Event, probability of event
    - Probability rules
    - Venn diagrams
    - Independent events
    - Conditional probability
    - 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.)



$$P(\text{caught nonword}) + P(\text{caught word})$$

$$= 0.25 \cdot 0.9 + 0.75 \cdot 0.7 = 0.75$$

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.

	Bachelor's	Master's	Professional	Doctorate	
Female	933	402	51	26	1412
Male	661	260	44	26	991
	1594	662	95	52	2403

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

$$P(\text{Female}) = \frac{1412}{2403} = 0.5876$$

	Bachelor's	Master's	Professional	Doctorate	Total
Female	933	402	51	26	1412
Male	661	260	44	26	991
Total	1594	662	95	52	2403

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

$$P(\text{Female} \mid \text{Prof.}) = \frac{51}{95} = 0.5368$$

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

No, since $P(\text{Female} \mid \text{Prof}) \neq P(\text{Female})$

Example: The table below describes the smoking habits of a group of asthma sufferers.

HS

	Non-smoker	Occasional Smoker	Regular Smoker	Heavy Smoker	Total
Men	344	40	85	39	508
Women	370	31	88	48	537
Total	714	71	173	87	1045

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

$$P(\text{Man or HS}) = P(\text{Man}) + P(\text{HS}) - P(\text{Man and HS})$$

$$= \frac{508}{1045} + \frac{87}{1045} - \frac{39}{1045} = 0.532$$

Random variables:

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

<u>Example</u>: Let X_1 and X_2 be independent and have the same distribution given below:

x_i	-1	0
p_i	0.7	0.3

Let $X = X_1 + X_2$. Find the probability distribution of X, μ_X and σ_X.

X_1+X_2 : -2 $\quad -1 \quad$ 0

P : $0.7\cdot 0.7 = 0.49$ $\quad 2\cdot 0.7\cdot 0.3 = 0.42$ $\quad 0.3\cdot 0.3 = 0.09$

$$\mu_X = -2\cdot 0.42 + (-1)0.42 + 0\cdot 0.09 = -1.4$$

$$\sigma_X = \sqrt{(-2+1.4)^2 0.49 + (-1+1.4)^2 0.42 + (0+1.4)^2 0.09} = 0.65$$

- Binomial distribution

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

  $\mu_X = np \qquad \sigma_X = \sqrt{np(1-p)}$

<u>Example</u>: 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 = # of patients for who it was ~~effective~~ effective

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

$P(X=7) = 0.2013$ (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 μ and finite standard deviation σ. When n is large enough,

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

<u>Example</u>: 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 = # of customers who buy a brownie

$X \sim Bin(500, 0.2)$, $\underbrace{np = 100 > 10, \; nq = 400 > 10}$

$X \sim$ appr. $N(100, \sqrt{500 \cdot 0.2 \cdot 0.8}) = N(100, \sqrt{80})$

$P(X > 110) = P(Z > \frac{110 - 100}{\sqrt{80}}) = P(Z > 1.12)$

$= 0.1314$

z-score for 110

- **Statistical inference**
  - Confidence intervals
    - For population mean μ: $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ if σ is unknown
    - For single proportion p: $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$
  - Hypothesis testing
    - Test statistic
      - To test H_0: $\mu = \mu_0$ we find statistic $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ (σ is unknown)
    - P-value
    - Significance level