MTH 224 **Summer 2019**
Review Notes **Drew Armstrong**

Key Topics from Chapter 1

- Suppose an experiment has a finite set S of equally likely outcomes. Then the probability of any event $E \subseteq S$ is
$$P(E) = \frac{\#E}{\#S}.$$
- For example, if we flip a fair coin n times then the $\#S = 2^n$ outcomes are equally likely. The number of sequences with k H's and $n-k$ T is $\binom{n}{k}$, thus we have
$$P(k \text{ heads}) = \frac{\#(\text{ways to get } k \text{ heads})}{\#S} = \frac{\binom{n}{k}}{2^n}.$$
- If we flip a strange coin with $P(H) = p$ and $P(T) = q$ then the $\#S = 2^n$ outcomes are not equally likey. In this case we have the more general formula
$$P(k \text{ heads}) = \binom{n}{k} P(H)^k P(T)^{n-k} = \binom{n}{k} p^k q^{n-k}.$$
This agrees with the previous formula when $p = q = 1/2$.
- These *binomial probabilities* add to 1 because of the *binomial theorem*:
$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1^n = 1.$$
- In general, a *probability measure* P on a sample space S must satisfy three rules:
  1. For all $E \subseteq S$ we have $P(E) \geqslant 0$.
  2. For all $E_1, E_2 \subseteq S$ with $E_1 \cap E_2 = \varnothing$ we have
  $$P(E_1 \cup E_2) = P(E_1) + P(E_2).$$
  3. We have $P(S) = 1$.
- Many other properties follow from these rules, such as the *principle of inclusion-exclusion*, which says that for general events $E_1, E_2 \subseteq S$ we have
$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).$$
- Also, if E' is the complement of an event $E \subseteq S$ then we have $P(E') = 1 - P(E)$.

- Venn diagrams are useful for verifying identities such as *de Morgan's laws*:

$$(E_1 \cap E_2)' = E_1' \cup E_2',$$
$$(E_1 \cup E_2)' = E_1' \cap E_2'.$$

- Given events $E_1, E_2 \subseteq S$ we define the *conditional probability*:

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}.$$

- *Bayes' Theorem* relates the conditional probabilites $P(E_1|E_2)$ and $P(E_2|E_1)$:

$$P(E_1) \cdot P(E_2|E_1) = P(E_2) \cdot P(E_1|E_2).$$

- The events E_1, E_2 are called *independent* if any of the following formulas hold:

$$P(E_1|E_2) = P(E_1) \quad \text{or} \quad P(E_2|E_1) = P(E_2) \quad \text{or} \quad P(E_1 \cap E_2) = P(E_1) \cdot P(E_2).$$

- Suppose our sample space is partitioned as $S = E_1 \cup E_2 \cup \cdots \cup E_m$ with $E_i \cap E_j = \varnothing$ for all $i \neq j$. For any event $F \subseteq S$ the *law of total probability* says

$$P(F) = P(E_1 \cap F) + P(E_2 \cap F) + \cdots + P(E_m|F)$$
$$P(F) = P(E_1) \cdot P(F|E_1) + P(E_2) \cdot P(F|E_2) + \cdots + P(E_m) \cdot P(F|E_m).$$

- Then the general version of *Bayes' Theorem* says that

$$P(E_k|F) = \frac{P(E_k \cap F)}{P(F)} = \frac{P(E_k) \cdot P(F|E_k)}{\sum_{i=1}^m P(E_i) \cdot P(F|E_i)}.$$

- The *binomial coefficients* have four different interpretations:

$$\begin{aligned}\binom{n}{k} &= \text{entry in the } n\text{th row and } k\text{th diagonal of Pascal's Triangle,}\\ &= \text{coefficient of } x^k y^{n-k} \text{ in the expansion of } (x+y)^n,\\ &= \#(\text{words made from } k \text{ copies of one letter and } n-k \text{ copies of another letter}),\\ &= \#(\text{ways to choose } k \text{ unordered things without replacement from } n \text{ things}).\end{aligned}$$

- And they have a nice formula:

$$\binom{n}{k} = \frac{n!}{k! \times (n-k)!} = \frac{n \times (n-1) \times \cdots \times (n-k+1)}{k \times (k-1) \times \cdots \times 1}.$$

- Ordered things are easier. Consider words of length k from an alphabet of size n:

$$\#(\text{words}) = n \times n \times \cdots \times n = n^k,$$

$$\#(\text{words without repeated letters}) = n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!}.$$

- More generally, the number of words containing k_1 copies of the letter "a_1," k_2 copies of the letter "a_2," ...and k_s copies of the letter "a_s" is

$$\binom{k_1 + k_2 + \cdots + k_s}{k_1, k_2, \ldots, k_s} = \frac{(k_1 + k_2 + \cdots + k_s)!}{k_1! \times k_2! \times \cdots \times k_s!}$$

- These numbers are called *multinomial coefficients* because of the *multinomial theorem*:

$$(p_1 + p_2 + \cdots + p_s)^n = \sum \binom{n}{k_1, k_2, \ldots, k_s} p_1^{k_1} p_2^{k_2} \cdots p_s^{k_s},$$

  where the sum is over all possible choices of $k_1, k_2, \ldots, k_s$ such that $k_1 + k_2 + \cdots + k_s = n$. Suppose that we have an s-sided die and p_i is the probability that side i shows up. If the die is rolled n times then the probability that side i shows up exactly k_i times is the *multinomial probability*:

$$P(\text{side } i \text{ shows up } k_i \text{ times}) = \binom{n}{k_1, k_2, \ldots, k_s} p_1^{k_1} p_2^{k_2} \cdots p_s^{k_s}.$$

- Finally, suppose that an urn contains r red and g green balls. If n balls are drawn without replacement then

$$P(k \text{ red}) = \frac{\binom{r}{k}\binom{g}{n-k}}{\binom{r+g}{n}}.$$

  More generally, if the urn contains r_i balls of color i for $i = 1, 2, \ldots, s$ then the probability of getting exactly k_i balls of color i is

$$P(k_i \text{ balls of color } i) = \frac{\binom{r_1}{k_1}\binom{r_2}{k_2}\cdots\binom{r_s}{k_s}}{\binom{r_1+r_2+\cdots+r_s}{k_1+k_2+\cdots+k_s}}.$$

  These formulas go by a silly name: *hypergeometric probability.*

Key Topics from Chapter 2

- Let S be the sample space of an experiment. A *random variable* is any function $X : S \to \mathbb{R}$ that assigns to each outcome $s \in S$ a real number $X(s) \in \mathbb{R}$. The *support of* X is the set of possible values $S_X \subseteq \mathbb{R}$ that X can take. We say that X is a *discrete* random variable is the set S_X doesn't contain any continuous intervals.

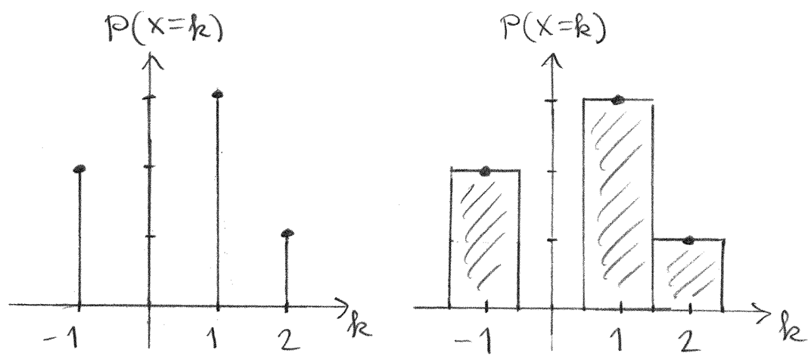
- The *probability mass function (pmf)* of a discrete random variable $X : S \to \mathbb{R}$ is the function $f_X : \mathbb{R} \to \mathbb{R}$ defined by

$$f_X(k) = \begin{cases} P(X = k) & \text{if } k \in S_X, \\ 0 & \text{if } k \notin S_X. \end{cases}$$

- We can display a probability mass function using either a table, a line graph, or a probability histogram. For example, suppose that a random variable X has pmf f_X defined by the following table:

k	-1	1	2
$f_X(k)$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{1}{6}$

Here is the line graph and the histogram:



- The *expected value* of a random variable $X : S \to \mathbb{R}$ with support $S_X \subseteq \mathbb{R}$ is defined by either of the following formulas:

$$E[X] = \sum_{k \in S_X} k \cdot P(X = k) = \sum_{s \in S} X(s) \cdot P(s).$$

On the one hand, we interpret this as the center of mass of the pmf. On the other hand, we interpret this as the long run average value of X if the experiment is performed many times.

- Consider any random variables $X, Y : S \to \mathbb{R}$ and constants $\alpha, \beta \in \mathbb{R}$. The expected value satisfies the following algebraic identities:

$$\begin{aligned} E[\alpha] &= \alpha, \\ E[\alpha X] &= \alpha E[X], \\ E[X + \alpha] &= E[X] + \alpha, \\ E[X + Y] &= E[X] + E[Y], \\ E[\alpha X + \beta Y] &= \alpha E[X] + \beta E[Y]. \end{aligned}$$

In summary, the expected value is a *linear function*.

- Let $X : S \to \mathbb{R}$ be a random variable with mean $\mu = E[X]$. We define the *variance* as the expected value of the squared distance between X and μ:

$$\mathrm{Var}(X) = E[(X-\mu)^2].$$

Using the properties above we also have

$$\mathrm{Var}(X) = E[X^2] - \mu^2 = E[X^2] - E[X]^2.$$

Since we feel bad about squaring the distance, we define the *standard deviation* by taking the square root of the variance:

$$\sigma = \sqrt{\mathrm{Var}(X)}.$$

- For random variables $X, Y : S \to \mathbb{R}$ with $E[X] = \mu_X$ and $E[Y] = \mu_Y$, we define the *covariance* as follows:

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)].$$

Using the above properties we also have

$$\mathrm{Cov}(X,Y) = E[XY] - E[X] \cdot E[Y].$$

Observe that $\mathrm{Cov}(X,X) = E[X^2] - E[X]^2 = \mathrm{Var}(X)$.

- For any $X, Y, Z : S \to \mathbb{R}$ and $\alpha, \beta \in \mathbb{R}$ we have

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{Cov}(Y,X), \\ \mathrm{Cov}(\alpha X + \beta Y, Z) &= \alpha \mathrm{Cov}(X,Z) + \beta \mathrm{Cov}(Y,Z). \end{aligned}$$

We say that covariance is a *symmetric* and *bilinear* function.

- Variance by itself satisfies the following algebraic identities:

$$\begin{aligned} \mathrm{Var}(\alpha) &= 0, \\ \mathrm{Var}(\alpha X) &= \alpha^2 \mathrm{Var}(X), \\ \mathrm{Var}(X+\alpha) &= \mathrm{Var}(X), \\ \mathrm{Var}(X+Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y). \end{aligned}$$

- For discrete random variables $X, Y : S \to \mathbb{R}$ we define their *joint pmf* f_{XY} as follows:

$$f_{XY}(k,\ell) = P(X = k \text{ and } Y = \ell).$$

We say that X and Y are *independent* if for all k and ℓ we have

$$f_{XY}(k,\ell) = f_X(k) \cdot f_Y(\ell) = P(X=k) \cdot P(Y=\ell).$$

If X and Y are independent then we must have $E[XY] = E[X] \cdot E[Y]$, which implies that $\mathrm{Cov}(X,Y) = 0$ and $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. The converse statements are not true in general.

- Let $\mathrm{Var}(X) = \sigma_X^2$ and $\mathrm{Var}(Y) = \sigma_Y^2$. If both of these are non-zero then we define the coefficient of coerrelation:
$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}.$$
We always have $-1 \leqslant \rho_{XY} \leqslant 1$.

- Let $p+q=1$ with $p \geqslant 0$ and $q \geqslant 0$. A *Bernoulli random variable* has the following pmf:

k	0	1
$P(X=k)$	q	p

We compute
$$\begin{aligned} E[X] &= 0 \cdot q + 1 \cdot p = p, \\ E[X^2] &= 0^2 \cdot q + 1^2 \cdot p = p, \\ \mathrm{Var}(X) &= E[X^2] - E[X]^2 = p - p^2 = p(1-p) = pq. \end{aligned}$$

- A sum of independent Bernoulli random variables is called a *binomial random variable*. For example, suppose that $X_1, X_2, \ldots, X_n$ are independent Bernoullis with $P(X_i = 1) = p$. Let $X = X_1 + X_2 + \cdots + X_n$. Then from linearity of expectation we have
$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_n] = p + p + \cdots + p = np$$
and from independence we have
$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) = pq + pq + \cdots + pq = npq.$$
If we think of each X_i as the number of heads from a coin flip then X is the total number of heads in n flips of a coin. Thus X has a *binomial pmf*:
$$P(X=k) = \binom{n}{k} p^k q^{n-k}.$$

- Suppose an urn contains r red balls and g green balls. Grab n balls without replacement and let X be the number of red balls you get. We say that X has a *hypergeometric pmf*:
$$P(X=k) = \frac{\binom{r}{k}\binom{g}{n-k}}{\binom{r+g}{n}}.$$
Let $X_i = 1$ if the ith ball is red and $X_i = 0$ if the ith ball is green. Then X_i is a Bernoulli random variable with $P(X_i = 1) = r/(r+g)$, hence $E[X_i] = r/(r+g)$, and from linearity of expectation we have
$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_n] = \frac{r}{r+g} + \frac{r}{r+g} + \cdots + \frac{r}{r+g} = \frac{nr}{r+g}.$$
Since the X_i are **not independent**, we can't use this method to compute the variance.[1]

[1] The variance is $\frac{nrg(r+g-n)}{(r+g)^2(r+g-1)}$ but you don't need to know this.

- Consider a coin with $P(H) = p$ and let X be the number of coin flips until you see H. We say that X is a *geometric random variable* with pmf

$$P(X = k) = P(T)^{k-1} \cdot P(H) = q^{k-1}p.$$

By manipulating the *geometric series*[2] we can show that

$$P(X > k) = q^k \qquad \text{and} \qquad P(k \leqslant X \leqslant \ell) = q^{k-1} - q^{\ell}.$$

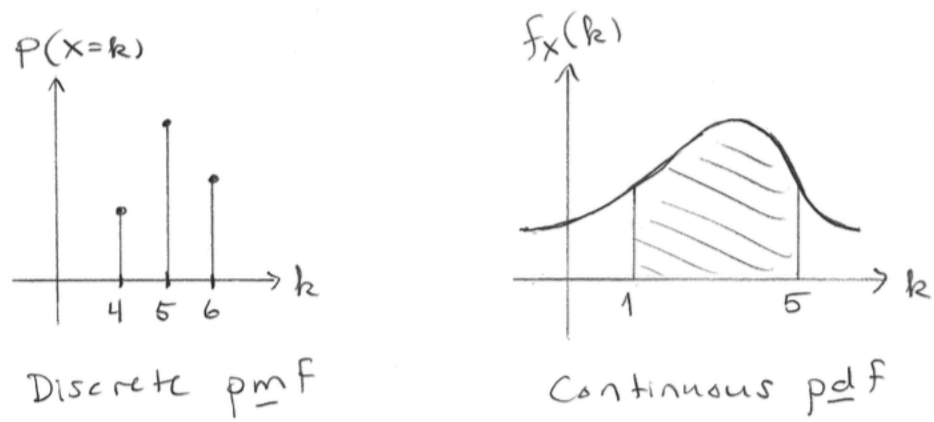
By manipulating the geometric series a bit more we can show that

$$E[X] = \frac{1}{p}.$$

In other words, we expect to see the first H on the $(1/p)$-th flip of the coin.[3]

Key Topics from Chapter 3

- Instead of a pmf $f_X(k) = P(X = k)$, a continuous random variable X is defined by a *probability density function (pdf)* $f_X : \mathbb{R} \to \mathbb{R}$. Here is a picture:



By definition the pdf must satisfy

$$f_X(x) \geqslant 0 \text{ for all } x \in \mathbb{R} \qquad \text{and} \qquad \int_{-\infty}^{\infty} f_X(x)\,dx = 1.$$

Then for any real numbers $a \leqslant b$ we define

$$P(a < X < b) = \int_a^b f_X(x)\,dx.$$

Note that this implies $P(X = k) = P(k \leqslant X \leqslant k) = 0$ for any $k \in \mathbb{R}$.

[2] If $|q| < 1$ then $1 + q + q^2 + \cdots = 1/(1-q)$.
[3] The variance is q/p^2 but you don't need to know this.

- Let $f_X : \mathbb{R} \to \mathbb{R}$ be the pdf of a continuous random variable X. Then we define the expected value by the formula

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx.$$

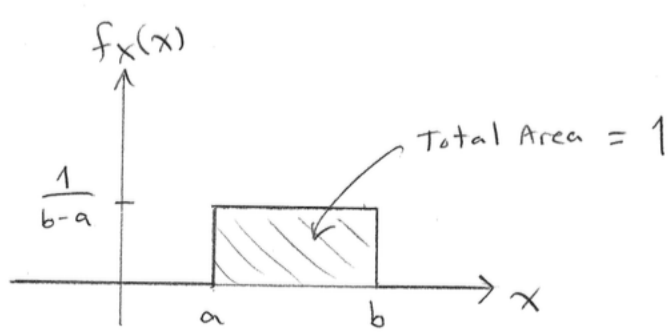
Just as in the discrete case, this integral represents the *center of mass* of the distribution. More generally, we define the rth moment of X by the formula

$$E[X^r] = \int_{-\infty}^{\infty} x^r \cdot f_X(x)\, dx.$$

As with the discrete case, the variance is defined as the average squared distance between X and its mean $\mu = E[X]$. That is, we have

$$\begin{aligned}
\mathrm{Var}(X) &= E[(X-\mu)^2] \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} (x^2 - 2\mu x + x^2) \cdot f_X(x)\, dx \\
&= \left(\int_{-\infty}^{\infty} x^2 \cdot f_X(x)\, dx\right) - 2\mu \left(\int_{-\infty}^{\infty} x \cdot f_X(x)\, dx\right) + \mu^2 \left(\int_{-\infty}^{\infty} f_X(x)\, dx\right) \\
&= E[X^2] - 2\mu \cdot E[X] + \mu^2 \cdot 1 \\
&= E[X^2] - 2\mu^2 + \mu^2 \\
&= E[X^2] - \mu^2 \\
&= E[X^2] - E[X]^2.
\end{aligned}$$

- The *uniform* distribution on a real interval $[a, b] \subseteq \mathbb{R}$ has the following pdf:

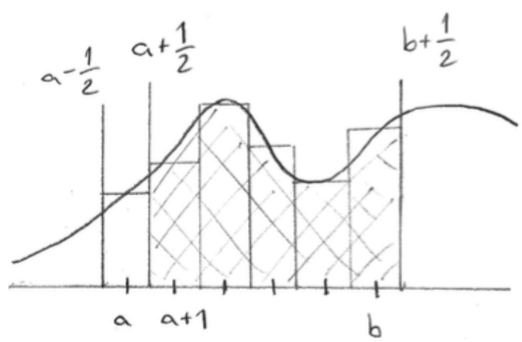


You should practice the definitions by proving that

$$E[X] = \frac{a+b}{2} \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

- Let X be a **discrete** random variable with pmf $P(X = k)$ and let Y be a **continuous** random variable with pdf f_Y. Suppose that for all integers k we have
$$P(X = k) \approx f_Y(k).$$
Then for any integers $a \leqslant b$ we can approximate the probability $P(a \leqslant X \leqslant b)$ by the area under the graph of f_Y, as follows:
$$\begin{aligned} P(a \leqslant X \leqslant b) &\approx \int_{a-1/2}^{b+1/2} f_Y(t)\,dt, \\ P(a < X \leqslant b) &\approx \int_{a+1/2}^{b+1/2} f_Y(t)\,dt, \\ P(a \leqslant X < b) &\approx \int_{a-1/2}^{b-1/2} f_Y(t)\,dt, \\ P(a < X < b) &\approx \int_{a+1/2}^{b-1/2} f_Y(t)\,dt. \end{aligned}$$

Here's a picture illustrating the second formula:



- Let X be a (discrete) binomial random variable with parameters n and p. If np and $n(1-p)$ are both large then de Moivre (1730) and Laplace (1810) showed that
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{1}{\sqrt{2\pi np(1-p)}}\, e^{-(k-np)^2/2np(1-p)}.$$
For example, let X be the number of heads in 3600 flips of a fair coin. Then we have
$$P(1770 \leqslant X \leqslant 1830) \approx \int_{1770-0.5}^{1830+0.5} \frac{1}{\sqrt{1800\pi}}\, e^{-(x-1800)^2/1800}\, dx \approx 69.07\%.$$

- In general, the *normal distribution* with mean μ and σ^2 is defined by the following pdf:
$$n(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}.$$
We will write $X \sim N(\mu, \sigma^2)$ for any random variable with this pdf.

- The *stability theorem* says that if X and Y are normal and if α, β, γ are constant then

$$\alpha X + \beta Y + \gamma \text{ is also normal.}$$

- A special case of the above fact says that normal random variables can be standardized:

$$X \sim N(\mu, \sigma^2) \qquad \Longleftrightarrow \qquad Z = \frac{X - \mu}{\sigma} \sim N(0,1).$$

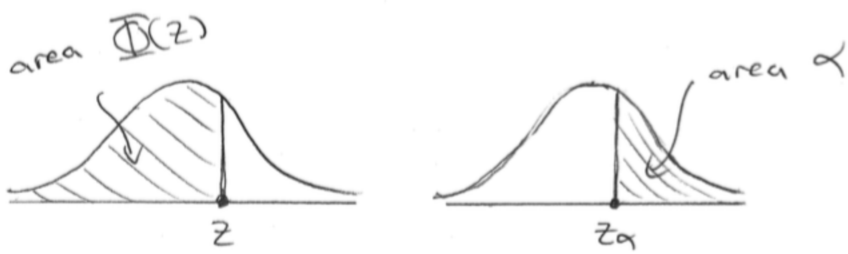
If Z is standard normal then it has the following *cumulative density function (cdf)*:

$$\Phi(z) = P(Z \leqslant z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx.$$

The values of $\Phi(z)$ can be looked up in a table. Furthermore, for any probability $0 < \alpha < 1$ we define the *critical value* z_α to be the unique number with the property

$$\int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx = P(Z \geqslant z_\alpha) = \alpha.$$

These numbers can also be looked up in a table. Here are some pictures:



- Let $X_1, X_2, \ldots, X_n$ be an *iid sample* with $\mu = E[X_i]$ and $\sigma^2 = \text{Var}(X_i)$. If $\overline{X} = (X_1 + \cdots + X_n)/n$ is the *sample mean* then we have

$$E[\overline{X}] = \mu \qquad \text{and} \qquad \text{Var}(\overline{X}) = \sigma^2/n.$$

The fact that $\text{Var}(\overline{X}) \to 0$ as $n \to \infty$ is called the *Law of Large Numbers (LLN)*. If n is large then the *Central Limit Theorem (CLT)* says that $\overline{X}$ is approximately normal:

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} \approx N(\mu, \sigma^2/n).$$

This is the most important theorem in all of (classical) statistics.

- *Application: Estimating a proportion.* Let p be proportion of yes voters in a population. To estimate p we take a random sample of n voters and let Y be the number who say yes. Then the *sample proportion* $\hat{p} = Y/n$ is an *unbiased estimator for* p because $E[\hat{p}] = p$. Furthermore, since $\text{Var}(\hat{p}) = p(1-p)/n$ we know that $(\hat{p} - p)/\sqrt{p(1-p)/n}$ is approximately $N(0,1)$.

Thus we obtain the following approximate $(1-\alpha)100\%$ intervals for the unknown p:

$$\begin{aligned} p &< \hat{p} + z_\alpha \cdot \sqrt{\hat{p}(1-\hat{p})/n}, \\ p &> \hat{p} - z_\alpha \cdot \sqrt{\hat{p}(1-\hat{p})/n}, \\ |p-\hat{p}| &< z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n}. \end{aligned}$$

If we want to test the hypothesis $H_0 =$ "$p = p_0$" at the α level of significance then we use the following rejection regions:

$$\begin{aligned} \hat{p} &> p_0 + z_\alpha \cdot \sqrt{p_0(1-p_0)/n} && \text{if } H_1 = \text{“}p > p_0,\text{”} \\ \hat{p} &< p_0 - z_\alpha \cdot \sqrt{p_0(1-p_0)/n} && \text{if } H_1 = \text{“}p < p_0,\text{”} \\ |\hat{p} - p_0| &> z_{\alpha/2} \cdot \sqrt{p_0(1-p_0)/n} && \text{if } H_1 = \text{“}p \neq p_0.\text{”} \end{aligned}$$

- *Application: Estimating a mean.* Let $X_1, X_2, \ldots, X_n$ be an iid sample from a normal distribution with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. The sample mean $\overline{X}$ is an *unbiased estimator* for μ because $E[\overline{X}] = \mu$. Furthermore, since $\text{Var}(\overline{X}) = \sigma^2/n$ we know from the stability theorem that $\overline{X}$ is exactly $N(\mu, \sigma^2/n)$.

  If σ^2 is known then we obtain the following exact $(1-\alpha)100\%$ intervals for μ:

$$\begin{aligned} \mu &< \overline{X} + z_\alpha \cdot \sqrt{\sigma^2/n}, \\ \mu &> \overline{X} - z_\alpha \cdot \sqrt{\sigma^2/n}, \\ |\mu - \overline{X}| &< z_{\alpha/2} \cdot \sqrt{\sigma^2/n}. \end{aligned}$$

  If we want to test the hypothesis $H_0 =$ "$\mu = \mu_0$" at the α level of significance then we use the following rejection regions:

$$\begin{aligned} \overline{X} &> \mu_0 + z_\alpha \cdot \sqrt{\sigma^2/n} && \text{if } H_1 = \text{“}\mu > \mu_0,\text{”} \\ \overline{X} &< \mu_0 - z_\alpha \cdot \sqrt{\sigma^2/n} && \text{if } H_1 = \text{“}\mu < \mu_0,\text{”} \\ |\overline{X} - \mu_0| &> z_{\alpha/2} \cdot \sqrt{\sigma^2/n} && \text{if } H_1 = \text{“}\mu \neq \mu_0.\text{”} \end{aligned}$$

  If σ^2 is unknown then we replace it with the *sample variance*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

  If n is small then we also replace z_α with $t_\alpha(n-1)$. This is because the random variable $(\overline{X} - \mu)/\sqrt{S^2/n}$ has a *t-distribution with $n-1$ degrees of freedom.*

- Chi-squared distributions are not on the exam.