EXAM 2 SUMMARY

Definition. Let X be a random variable.

(1) All random variables have a **cumulative distribution function (CDF)**: $F(x) = P(X \le x)$.
(2) Discrete random variables have a **probability mass function (PMF)**: $p(x) = P(X = x)$.
(3) Continuous random variables has a **probability density function (PDF)** $f(x)$ such that for any numbers a and b (with $a \le b$): $P(a \le X \le b) = \int_a^b f(x)dx$.

Theorem. *Let X be a continuous random variable.*

(1) If $F(x)$ is the CDF of X, then a PDF for X is $f(x) = F'(x)$.
(2) If $f(x)$ is a PDF for X, then the CDF of X is $F(x) = \int_{-\infty}^x f(t)dt$.

Definition. Expected value (or **mean**):

(1) If X is a discrete RV with PMF $p(x)$, then $\mu = E(X) = \sum_x xp(x)$
For any function g, $E(g(X)) = \sum_x g(x)p(x)$
(2) If X is a continuous RV with PDF $f(x)$, then $\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$
For any continuous function g, $E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$

Definition. The **variance** of X: $\sigma^2 = \text{Var}(X) = E\left[(X-\mu)^2\right] = E(X^2) - [E(X)]^2$. The **standard deviation** of X: $\sigma = \sqrt{\sigma^2}$.

Theorem. *For any random variable X and any constants a and b:*

(1) $E(aX + b) = aE(X) + b$ and
(2) $Var(aX + b) = a^2\,Var(X)$.

Theorem. *$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$ for any random variables $X_1, X_2, \ldots X_n$. If $X_1, X_2, \ldots X_n$ are independent, then it also follows that $Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + Var(X_2) + \cdots + Var(X_n)$.*

Definition. A **random sample of size** n is a set of independent identically distributed (iid) random variables $X_1, X_2, \ldots X_n$. Some **sample statistics**:

(1) The **sample total** $T = \displaystyle\sum_{i=1}^n X_i$

(2) The **sample mean**: $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^n X_i$

(3) The **sample variance**: $S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^n (X_i - \overline{X})^2$

Theorem. *For any random sample from a population with mean μ and variance σ^2:*

(1) $E(T) = n\mu$ and $Var(T) = n\sigma^2$
(2) $E(\overline{X}) = \mu$ and $Var(\overline{X}) = \frac{\sigma^2}{n}$
(3) $E(S^2) = \sigma^2$

Definition. The **standard error** of the sample mean $\overline{x}$ is $\frac{s}{\sqrt{n}}$

Definition. A sample statistic $\hat{\Theta}$ is an **unbiased estimator** of population parameter θ if $E(\hat{\Theta}) = \theta$. The **bias** of $\hat{\Theta}$ as an estimator for θ is $\text{bias}(\hat{\Theta}) = E(\hat{\Theta}) - \theta$.

Theorem. *If $\overline{X}$ is a the mean of a random sample from a normally distribution population, then $\overline{X}$ is normally distributed (with mean and variance given in the last theorem).*

Date: March 21, 2023.

Theorem (Central Limit Theorem)**.** *If $\overline{X}$ is a the mean of a large random sample from any population, then $\overline{X}$ is approximately normally distributed (with mean and variance given in the theorem above).*

Definition. Let X and Y be jointly distributed discrete RVs with joint PMF $p(x, y) = P(X = x, Y = y)$.

i. The **marginal PMF** of X is $p_X(x) = P(X = x) = \sum_y p(x, y)$.

ii. The **conditional PMF** of Y given $X = x$ is $p_{Y|X=x}(y) = P(Y = y|X = x) = \dfrac{p(x, y)}{p_X(x)}$.

iii. The **regression function** of Y on X is $\mu_{Y|X=x} = \displaystyle\sum_y y p_{Y|X=x}(y)$. Note that this is a function of x.

iv. X and Y are **independent** if $p(x, y) = p_X(x)p_Y(y)$ for all possible values of x and y.

Definition. Let X a Y be jointly distributed RVs. The **covariance** of X and Y is

$$\mathrm{Cov}(X, Y) = \sigma_{X,Y} = E\left[(X - \mu_X)(Y - \mu_Y)\right] = E(XY) - E(X)E(Y)$$

Pearson's correlation coefficient is $\rho = \dfrac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$. The correlation coefficient ρ takes values in the interval $[-1, 1]$.

Theorem. *If X and Y are independent, then $Cov(X, Y) = 0$.*

Definition. A random variable X has a **Bernoulli** distribution with parameter p (with $0 < p < 1$) if X has possible values 0 and 1 with $P(X = 1) = p$ and $P(X = 0) = 1 - p$. The outcome 1 is often referred to as "success" while 0 is "failure" and the experiment is often called a Bernoulli trial.

Definition. The total number of successes in n independent, identically distributed (iid) Bernoulli trials with parameter p is a random variable with a **binomial** distribution. The PMF of a random variable X having a binomial distribution with parameters n and p is

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x} \text{ for } x = 0, 1, \ldots, n$$

Proposition. *The mean and variance of a binomial distribution are $\mu = np$ and $\sigma^2 = np(1 - p)$.*

R Implementation. If $X \sim \mathrm{binom}(n, p)$, then the PMF is `dbinom(x, n, p)` and the CDF is `pbinom(x, n, p)`.

Definition. If independent, identically distributed Bernoulli trials with parameter p are repeated until the first success, then the total number of trials (counting the success) has a **geometric** distribution. The PMF of a random variable X having a geometric distribution with parameter p is

$$p(x) = (1 - p)^{x-1} p \text{ for } x = 1, 2, 3, \ldots$$

and the CDF is

$$F(x) = 1 - (1 - p)^x \text{ for } x = 1, 2, 3, \ldots$$

Proposition. *The mean and variance of a geometric distribution are $\mu = \frac{1}{p}$ and $\sigma^2 = \frac{1-p}{p^2}$.*

Definition. Suppose n elements are to be selected without replacement from a population of size $N = M_1 + M_2$ where M_1 is the number of successes and M_2 is the number of failures. The number of successes selected is a **hypergeometric** random variable and its PMF is

$$h(x) = \frac{\binom{M_1}{x}\binom{N-M_1}{n-x}}{\binom{N}{n}} = \frac{\binom{M_1}{x}\binom{M_2}{n-x}}{\binom{M_1+M_2}{n}}$$

Proposition. *The mean and variance of a hypergeometric distribution are $\mu = \frac{nk}{N}$ and $\sigma^2 = \frac{nk(N-k)(N-n)}{k^2(N-1)}$.*

R Implementation. If $X \sim \mathrm{hyper}(M_1, M_2, n)$, then the PMF is `dhyper(x, M₁, M₂, n)` and the CDF is `phyper(x, M₁, M₂, n)`.

Definition. A **Poisson** random variable with parameter $\lambda > 0$ has the PMF

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \text{ for } x = 0, 1, 2, \ldots$$

Proposition. *The mean and variance of a Poisson distribution are $\mu = \lambda$ and $\sigma^2 = \lambda$.*

R Implementation. If $X \sim \text{pois}(\lambda)$, then the PMF is `dpois(x, λ)` and the CDF is `ppois(x, λ)`.

Definition. A random variable X with a **uniform continuous** distribution on the interval $[\alpha, \beta]$ has the PDF: $f(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{if } \alpha < x < \beta \\ 0 & \text{elsewhere} \end{cases}$

Proposition. *The mean and variance of a uniform continuous distribution on $[\alpha, \beta]$ are $\mu = \frac{\alpha+\beta}{2}$ and $\sigma^2 = \frac{(\beta-\alpha)^2}{12}$.*

Definition. A random variable with an **exponential** distribution with parameter $\lambda > 0$ has PDF $f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{elsewhere} \end{cases}$ and CDF $F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{elsewhere} \end{cases}$

Proposition. *An exponential distribution with parameter λ has mean $\mu = \frac{1}{\lambda}$ and variance $\sigma^2 = \frac{1}{\lambda^2}$.*

R Implementation. If $X \sim \exp(\lambda)$, then the CDF is `pexp(x, λ)`.

The parameter λ is called the **rate**. For example, a rate of 2 events per minute corresponds to a mean of 0.5 minutes between events. If X is exponential with a mean of 2.5, then $P(1 < X \leq 3)$ can be calculated using

```
> pexp(3, 0.4) - pexp(1, 0.4)
```

Definition. A random variable with a **normal** distribution with parameters μ and $\sigma^2 > 0$ has the PDF: $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ for all $x \in \mathbb{R}$

Proposition. *A normal distribution with parameters μ and σ^2 has mean $\mu = \mu$ and variance $\sigma^2 = \sigma^2$.*

R Implementation. If $X \sim \text{N}(\mu, \sigma^2)$, then the CDF is `pnorm(x, μ, σ)` (note that R wants the standard deviation, not the variance).

Of these distributions, only the uniform continuous distribution and the exponential distribution allow one to easily compute probabilities by hand. Calculations of probabilities for the rest of the distributions generally rely on a table or computational device. For a normally distributed random variable X this often means **standardizing** the random variable. If $X \sim \text{N}(\mu, \sigma^2)$, then $Z = \dfrac{X - \mu}{\sigma}$ has a **standard normal distribution** with mean 0 and variance 1.

Challenge. Suppose that a population is uniformly distributed on the interval $[1, \beta]$. Let $X_1, X_2, \ldots, X_{16}$ be a random sample from this population. Find an unbiased estimator for β.

Challenge. Sometimes people conducting surveys want to ask questions with potentially embarrassing answers. They may use something called a *randomized response technique* to make sure that survey participants have no reason to lie. For example, suppose you wanted to determine how widespread cheating is on campus. You could print the phrase "I have cheated on a test" on 4 cards and "I have never cheated on a test" on 6 cards. Then you could ask each survey participant to select a random card and tell you if the statement on the card is true or false *without showing you the card.*

a) Suppose 55% of those surveyed said the statement on the card was true. What is your estimate of the percentage of Gonzaga students who have cheated?
b) Suppose 60% of those surveyed said the statement on the card was true. What is your estimate of the percentage of Gonzaga students who have cheated?
c) Suppose 65% of those surveyed said the statement on the card was true. What is your conclusion?