STAT 340/341: Formula Sheet

Autumn 2023/Winter 2024

The following is a summary of formulas we have seen in STAT 340 and STAT 341. You will receive a copy of this summary with the final. You are not allowed to bring any notes/cheat sheets etc.

STAT 340 FORMULA SHEET

1. **Discrete Distributions (§ 5)**

a. A random variable is a *function* which maps each outcome in a sample space to a number. More informally, it is a variable whose value depends on the outcome of a random experiment.

Notation: uppercase X denotes the random variable as a function, lowercase x denotes a possible value or number.

b. PMF: probability of observing a specific value x

$$f(x) = P(X = x)$$

c. CDF: accumulated probability up till a specific value x

$$F(x) = P(X \leq x)$$

.....

d. Mean and variance (§ 7)

i. Mean: a number which represents the **average** value of random variable across separate replications of the experiment

Definition	$\mu = E[X] = \sum_{-\infty}^{\infty} x \cdot f(x)$
Linearity of Expectation	$E[aX + b] = aE[X] + b$
Law of the Unconscious Probabilist	$E[t(X)] = \sum_{-\infty}^{\infty} t(x) \cdot f(x).$

ii. Variance: a positive number which describes spread of the values of the random variable from the mean.

Definition	$\sigma^2 = Var[X] = \sum_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)$
Short cut for calculation	$\sigma^2 = E\left[X^2\right] - \mu^2.$
Variance of linear transformation	$Var[aX + b] = a^2 \cdot Var[X]$

iii. Standard deviation: positive square root of variance which is on the same units as data. It is interpretable as the *typical* deviation of the values from the mean.

iv. Chebychevs's inequality: a useful inequality which provides an upper bound for the probability that a random variable can be more than k standard deviations from the mean.

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

d. Binomial Random Variable (§ 6)

i. A binomial random variable counts the number of successes in n independent trials where each trial results in a success with probability π or in a failure with probability $1-\pi$. We write $X \sim Binom(n, \pi)$.

ii. Binomial PMF

$$f(x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}, \quad x = 0, 1, \ldots, n$$

iii. Mean of $X \sim Binom(n, \pi)$: $n\pi$

iv. Variance of $X \sim Binom(n, \pi)$: $n\pi(1-\pi)$

v. Relevant R functions:

- `dbinom(x, size, prob)` calculates $f(x) = P(X = x)$
- `pbinom(q, size, prob)` calculates $F(q) = P(X \leq q)$
- `pbinom(q, size, prob,lower.tail = F)` calculates $P(X > q)$.

e. Geometric random variable (§ 8.1)

i. A geometric random variable counts the number of failures *before* we see the first success when independent trials with probability π of observing a success are performed. We write $X \sim Geom(\pi)$.

ii. Geometric PMF

$$f(x) = \pi(1-\pi)^x, \ x = 0, 1, 2, \ldots$$

iii. For any non-negative integer k, we have the result

$$P(X \geq k) = (1-\pi)^k$$

iv. A geometric distribution is *memoryless*. This means for all non-negative integers x, k

$$P(X \geq x + k | X \geq k) = P(X \geq x)$$

v. Mean of $X \sim Geom(\pi)$: $\frac{1-\pi}{\pi}$.

vi. Relevant R functions:

- `dgeom(x, prob)` calculates $f(x) = P(X = x)$
- `pgeom(q, prob)` calculates $F(q) = P(X \leq q)$
- `pgeom(q, prob,lower.tail = F)` calculates $P(X > q)$.

f. Poisson (§ 8.2)

i. The Poisson random variable counts the number of occurrences of an event over a fixed time period or within a space. We write $X \sim Poisson(\lambda)$ where λ denotes the rate of occurrence.

ii. The PMF of a Poisson can be derived from a $Binom(n, \pi)$ by setting $\pi = \frac{\lambda}{n}$ in the binomial PMF and letting $n \to \infty$.

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2 \ldots$$

iii. Mean of $X \sim Pois(\lambda)$: λ

iv. Variance of $X \sim Pois(\lambda)$: λ

v. Relevant R functions:

- `dpois(x, lambda)` calculates $f(x) = P(X = x)$
- `ppois(q, lambda)` calculates $F(q) = P(X \leq q)$
- `ppois(q, lambda,lower.tail = F)` calculates $P(X > q)$.

---

2. **Continuous Distributions**

a. PDF and CDF (§ 9)

i. The PDF is any function which satisfies two properties:

$$f(x) \geq 0 \ \forall \ x, \quad \int_{-\infty}^{\infty} f(x)dx = 1.$$

ii. Probabilities are calculated as areas under the PDF:

$$P(a \leq X < b) = \int_{a}^{b} f(x)dx.$$

Since a single value has no area, $P(X = x) = 0$ for any x however.

iii. The CDF $F(x)$ is again the accumulated probability up til a value x:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

The CDF has the following properties:

- it is non-decreasing
- it is right continuous
- $\lim_{x \to \infty} F(x) = 1$
- $\lim_{x \to -\infty} F(x) = 0$

iv. By the Fundamental Theorem of Calculus, we can write

$$f(x) = \frac{d}{dx} F(x).$$

. . . . . .

b. Mean and variance and higher moments (§ 12)

i. Mean: $\mu = E[X] = \int_{-\infty}^{\infty} x \cdot f(x)dx$

ii. Variance: $\sigma^2 = Var[X] = E\left[X^2\right] - \mu^2$

iii. The results stated in 1d. for Discrete Distributions hold in the continuous case as well.

iv. In addition to the mean and variance, we can also calculate percentiles for a continuous distribution.

- The $100p$ percentile of a continuous distribution is the number q such that $P(X < q) = p$

. . . . . .

c. Uniform random variable (§ 10)

i. The uniform random variable is the continuous analog of the equally likely model in a discrete sample space. We write $X \sim Unif(a, b)$.

ii. PDF of a uniform

$$f(x) = \frac{1}{b-a}, \quad a \leq x < b$$

iii. Mean of $X \sim Unif(a, b)$: $(a + b)/2$

iv. Variance of $X \sim Unif(a, b)$: $(b - a)^2/12$

v. The $100p$th percentile of $X \sim Unif(a, b)$ is given by $a + (b - a) \times p$

vi. Relevant R functions: For $X \sim Unif(min, max)$

- `dunif(x, min, max)` calculates PDF $f(x)$
- `punif(q, min, max)` calculates $F(q) = P(X \leq q)$
- `punif(q, min, max, lower.tail = F)` calculates $P(X > q)$.
- `qunif(p, min, max)` calculates the $100p$th percentile

. . . . . .

d. Exponential random variable (§ 11)

i. The exponential distribution arises as the inter-event time in a Poisson model. However, it can be used as a model for any non-negative random variable! We write $X \sim Exp(\lambda)$ where $\lambda(> 0)$ is called the *rate* parameter.

ii. PDF of an exponential random variable:

$$f(x) = \lambda e^{-\lambda x}, \quad 0 \leq x < \infty$$

iii. CDF of an exponential random variable:

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & 0 \leq x \end{cases} \end{aligned}$$

iv. The exponential distribution is *memoryless*: this means for $x, k > 0$ we have the result:

$$P(X \geq x + k | X \geq k) = P(X \geq x)$$

v. Mean of $X \sim Exp(\lambda)$: $\frac{1}{\lambda}$

vi. Variance of $X \sim Exp(\lambda)$: $\frac{1}{\lambda^2}$

vii. The $100pth$ percentile of $X \sim Exp(\lambda)$ is given by $-\frac{1}{\lambda}\ln(1-p)$.

viii. Relevant R functions: For $X \sim Exp(rate)$

- `dexp(x, rate)` calculates PDF $f(x)$
- `pexp(q, rate)` calculates $F(q) = P(X \leq q)$
- `pexp(q, rate, lower.tail = F)` calculates $P(X > q)$
- `qexp(p, rate)` calculates the $100p$th percentile

. . . . . .

e. Normal random variable (§ 13)

i. The normal distribution is often used as a model for biological measurements such as height, weight etc. It is also the limiting distribution for other models, such as the binomial, Poisson, etc. We write $X \sim Norm(\mu, \sigma)$.

ii. We can write $X = \mu + \sigma\, Z$ where $Z \sim Norm(0, 1)$.

iii. PDF of a normal:

$$f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \qquad -\infty < x < \infty$$

iv. The mean of $X \sim Norm(\mu, \sigma)$: μ

v. The variance of $X \sim Norm(\mu, \sigma)$: σ^2.

vi. The 68-95-99.7 rule states that regardless of the value of μ and σ the area within 1/2/3 standard deviations of the mean is 68%/95%/99.7%.

vii. The $100p$th percentile of $X \sim Norm(\mu, \sigma)$ is $\mu + \sigma q$ where q is the corresponding percentile for the standard normal distribution.

viii. Relevant R functions: For $X \sim Norm(\mu, \sigma)$

- `dnorm(x, mean, sd)` calculates PDF $f(x)$
- `pnorm(q, mean, sd)` calculates $F(q) = P(X \leq q)$
- `pnorm(q, mean, sd, lower.tail = F)` calculates $P(X > q)$.
- `qnorm(p, mean, sd)` calculates the $100p$th percentile

Sums and Series

Binomial Theorem For any real numbers a and b and integer $n > 0$

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$$

Geometric Series For any real numbers a and r ($|r| < 1$)

$$a + ar + ar^2 + ar^3 + \cdots = \frac{a}{1-r}$$

Taylor series for e^x:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots$$

STAT 341 FORMULA SHEET

The following is a summary of what we have covered in STAT 341 and the corresponding sections in the notes.

1. **Combining Random Variables** Data can be considered as realizations of random variables, particularly when some chance mechanism is employed to generate the data. Often, we summarize the data by considering an average. Chapters 15 and 16 describe some results relating to the distribution of a sum, minimum and maximum of a set of independent random variables. The important facts are summarized below.

a. **Expected value and variance of a linear combination of random variables:** Let X and Y be jointly distributed random variables and a, b, c be numbers. Then

   i. $E[aX + bY + c] = aE[X] + bE[Y] + c$.

   ii. $Var[aX + bY + c] = a^2 Var[X] + b^2 Var[Y]$ provided X,Y are independent.

   iii. The results are stated in terms of two random variables, but can be generalized to $X_1, X_2, \ldots, X_n$.

b. **Distribution of a sum of independent random variables**

   i. Suppose $X_1, X_2 \ldots, X_k$ are independent random variables with each $X_i \sim Binom(n_i, \pi)$. Then

$$S_k = X_1 + X_2 + \cdots + X_k \sim Binom(size = n_1 + n_2 + \cdots + n_k, prob = \pi).$$

   ii. Suppose $X_1, X_2 \ldots, X_k$ are independent random variables with each $X_i \sim Pois(\lambda_i)$. Then

$$S_k = X_1 + X_2 + \cdots + X_k \sim Pois(\sum_{i=1}^{k} \lambda_i).$$

   iii. Suppose $X_1, X_2 \ldots, X_n$ are independent random variables with each $X_i \sim Norm(\mu, \sigma)$. Then

$$S_n = X_1 + X_2 + \cdots + X_n \sim Norm(mean = n\mu, sd = \sqrt{n}\sigma).$$

   iv. The **Central Limit Theorem** says that result (iii) holds approximately when n is large whether or not the sample is drawn from a Normal distribution.

$$S_n = X_1 + X_2 + \cdots + X_n \approx Norm(mean = n\mu, sd = \sqrt{n}\sigma) \quad n \text{ large.}$$

   It may be equivalently re-stated as:

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \approx Norm\left(mean = \mu, sd = \frac{\sigma}{\sqrt{n}}\right) \quad n \text{ large.}$$

c. **Distribution of the minimum or maximum of independent random variables**

Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed *continuous* random variables, each having PDF f and CDF F. Let $x_1, x_2, \ldots, x_n$ denote the realized values for a sample. Then

i. The PDF of the random variable, X_{max} which takes as its value the largest of $x_1, x_2, \ldots, x_n$ is

$$f_{X_{max}}(x) = n\left[F(x)\right]^{n-1} f(x).$$

- Example: Suppose $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Unif(0, \theta)$,then $f_{X_{max}}(x) = \frac{nx^{n-1}}{\theta^n} \qquad 0 \leq x < \theta.$

ii. The PDF of the random variable, X_{min} which takes as its value the smallest of $x_1, x_2, \ldots, x_n$ is

$$f_{X_{min}}(x) = n\left[1 - F(x)\right]^{n-1} f(x).$$

- Suppose $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Exp(rate = \lambda)$,then $f_{X_{min}}(x) = n\lambda e^{-n\lambda x} \qquad 0 \leq x < \infty.$

. . . . . .

2. **Estimation** An important step in being able to use probability distributions as models for data is "fitting the model". This means identifying values of parameters suggested by the data for that distribution. The **Method of Moments** is one method for estimating parameters.

a. Method of Moments Idea (**17.1**): Suppose random variables $X_1, X_2, \ldots, X_n$ are drawn from a distribution indexed by a parameter with true value equal to θ_0. The method of moments *estimator* - denoted by $\hat{\theta}_0^{mom}$ - is the value which solves the equation[1]:

$$E\left[X\right] = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

i. $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Exp(\lambda_0)$: $\quad \hat{\lambda}_0^{mom} = 1/\bar{X}.$

ii. $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Bernoulli(\pi_0)$: $\quad \hat{\pi}_0^{mom} = \bar{X}.$

iii. $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Unif(0, \theta_0)$: $\quad \hat{\theta}_0^{mom} = 2\bar{X}.$

iv. Extending the idea to t parameters: equate the first t moments of the distribution with their sample counterparts and solve for the unknown parameters.

$$E\left[X^r\right] = \frac{1}{n}\sum_{i=1}^{n} X_i^r \qquad r = 1, 2, \ldots, t$$

b. Properties of Estimators (**17.2**) An estimator is a random variable, its value in a particular sample is called an estimate. The probability distribution of an estimator is called a **sampling distribution**. The standard deviation of an estimator is called **standard error**.

i. The bias in an estimator is defined as

$$\textbf{Bias} = E\left[\hat{\theta}_0\right] - \theta_0.$$

ii. An estimator $\hat{\theta}_0$ for a parameter θ_0 is said to be **unbiased** if the Bias is exactly equal to 0. It is **asymptotically unbiased** if $Bias \to 0$ as $n \to \infty$.

iii. An estimator $\hat{\theta}_0$ for a parameter θ_0 is said to be **consistent** if for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\hat{\theta}_0 - \theta_0| > \epsilon) = 0.$$

In other words, the sampling distribution of $\hat{\theta}_0$ becomes increasingly concentrated about the true value θ_0.

[1] if $E\left[X\right] = 0$, then consider solving $E\left[X^2\right] = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ instead

iv. An estimator $\hat{\theta}_0$ for θ_0 which is unbiased (or at least asymptotically unbiased) is consistent provided $Var\left[\hat{\theta}_0\right] \to 0$ as $n \to \infty$.

v. The Mean Square Error (MSE) of an estimator is defined as

$$MSE(\hat{\theta}_0) = Bias^2(\hat{\theta}_0) + Var(\hat{\theta}_0).$$

Estimators with small MSE are preferred.

. . . . . .

3. **Interval Estimation:** A $100(1-\alpha)\%$ confidence interval for a parameter θ_0 is a a random interval $[L, U]$ such that

$$P(L \leq \theta_0 \leq U) = 1 - \alpha$$

a. Large sample confidence interval for mean (**18.2**) When n is large and σ_0 is known, the interval estimate for a population mean μ_0 is constructed using the estimator $\bar{X}$ and its approximate normal sampling distribution. A 100 $(1-\alpha)\%$ **large sample** confidence interval for μ_0 is then

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of a standard normal distribution.

i. The following table shows commonly used confidence levels and the corresponding values for $z_{\alpha/2}$ for the large sample confidence interval for μ_0. It is important to keep in mind that $z_{\alpha/2}$ - referred to as the *critical value* - is the $(1-\alpha/2)$ quantile of the standard normal distribution, that is

$$P(Z \leq z_{\alpha/2}) = 1 - \alpha/2.$$

α	confidence level $(1-\alpha)$	critical value $z_{\alpha/2}$	R expression
0.01	99%	2.58	qnorm(p = 0.995)
0.05	95%	1.96	qnorm(p = 0.975)
0.1	90%	1.65	qnorm(p = 0.95)

ii. The Margin Of Error (MOE) is the *half length* of the confidence interval. For the large sample confidence interval for μ_0, we have the following formula. We use this to make sample size calculations in order to achieve a desired level of accuracy.

$$MOE = z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}$$

iii. When σ_0 is unknown, it can be replaced by an estimate.

Name	Parameter	Point estimator	SE	95% confidence interval estimate
$X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mu_0, \sigma_0$ $\mu_0 = E[X_i]$ $\sigma_0^2 = Var[X_i]$	μ_0	$\bar{X}$	$\frac{\sigma_0}{\sqrt{n}}$	$\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$
$X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Binom(1, \pi_0)$ $X = X_1 + X_2 + \cdots + X_n$ $\mu_0 = E[X_i] = \pi_0$ $\sigma_0^2 = Var[X_i] = \pi_0(1-\pi_0)$	π_0	$\hat{\pi}_0 = \frac{X}{n}$	$\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$	$\hat{\pi}_0 \pm 1.96\sqrt{\frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n}}$

. . . . . .

4. **Significance Testing:** In this form of inference, we choose between two claims, called the null (H_0) and alternative (H_1) about a population distribution. A significance test analyzes whether the data assert H_1 using the indirect approach of proof by contradiction. Specifically H_0 is presumed to be true. Under this presumption, if the observed value of the **test statistic** would be considered very unusual, as judged by the **P-value**, we have our contradiction.

   Some important definitions in this context are stated below.

i. The P-value is the probability, **presuming H_0 is true**, of observing our test statistic value or a value even more extreme/unusual in the direction predicted by H_1.

ii. Decision errors

- The α level or significance level is a number between 0 and 1 such that we reject H_0 if P-value $\leq \alpha$.
- Type 1 error rate is the probability of rejecting H_0 when H_0 is true.
- Type 2 error rate is the probability of failing to reject H_0 when a particular value in H_1 is true.

Practice

The following questions are for practice. They are not to be interpreted as a reflection of what I will ask on the test. They are just here to improve your problem solving skills and to reinforce your familiarity with the material.

1. Ariel (A) and Lucia (L) are playing a game in which the higher score wins. A's scores are (approximately) normally distributed with a mean of 100 and a standard deviation of 20. L's scores are (approximately) normally distributed with a mean of 110 and a standard deviation of 15.

a. Who is more likely to score above 150?

b. Assuming their scores are independent, what is the approximate probability that A beats L?

c. Now suppose they decide to play 3 games and declare the winner to be the one who gets the highest total score for the three games together. What is the approximate probability that A beats L in this format?

d. One more change. This time they will play the best of three format, meaning the winner is the one who wins two of the three games. What is the approximate probability that A beats L in this format?

2. Suppose a particle moves along the x axis starting at 0. Each minute, it is twice as likely to move a step to the right that the left. You may assume all steps are of unit length and that successive step directions are independent.

   Let X_i equal +1 if the particle moves to the right on the ith minute and -1 otherwise. The location, S, of the particle after an hour can then be written as the sum of 60 i.i.d. random variables:

$$S = X_1 + X_2 + \cdots + X_{60}.$$

a. Write the PMF of X_i.

b. Use the Central Limit Theorem to approximate the probability distribution of S.

c. Where is the particle most likely to be in an hour? Explain. Also, does your answer make sense intuitively? Why or why not?

3. Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with PDF

$$f(x) = \frac{2x}{\theta_0^2} \quad 0 \leq x < \theta_0$$

a. Find $\hat{\theta}_0^{mom}$, the method of moments estimator of θ_0.

b. Is your estimator unbiased?

c. Is your estimator consistent? Find the variance of your estimator. You will need to find $Var[X]$ first for this part.

d. Suppose we have the following $n = 8$ observations from this probability distribution for some unknown θ_0.

x
0.7139740
0.8836683
1.0231620
0.6150005
1.4112628
0.7713351
0.9855480
0.5629832

```
## xbar 0.8708668
```

Calculate your estimate of θ_0 based on this sample. Also find the estimated standard error.

4. Suppose we have two unbiased estimators T_1 and T_2 of θ.

a. Show that $aT_1 + (1-a)T_2$ is also an unbiased estimator of θ for $0 \leq a \leq 1$.
b. If T_1 and T_2 are also independent, e.g., determined from independent samples, then calculate $Var(aT_1 + (1-a)T_2)$ in terms of $Var(T_1)$ and $Var(T_2)$.
c. For the situation in part (b), determine the best choice of a in the sense that $Var(aT_1 + (1-a)T_2)$ is smallest.
d. What is the effect on this combined estimator of T_1 having a very large variance relative to T_2?

5. A psychologist is teaching a class of 100 students. He administers a test of passivity to them, and finds that 20 students score over 50. His conclusion is that about 20% of students would score over 50 on the passivity test. He estimates the standard error of the number as $\sqrt{100 \times 0.2 \times 0.8} = 4$. What does statistical theory say?

6. The US Commission on Crime wants to estimate the proportion of crimes related to firearms in an area with one of the highest crime rates in the country. They intend to draw a random sample of n files of recently committed crimes in the area. How many files do they need to look at if they require 90% confidence that $\hat{\pi}_0$ - the proportion of cases in the sample with firearms - is within 5% of π_0, the true proportion in the area? (As in HW 6 problem 3, you can set $\pi_0 = \frac{1}{2}$ in the standard error formula)

   You are given the following normal quantiles to help you do your calculation. Pick the right one and proceed.

```
qnorm(p = 0.9)
```

```
## [1] 1.281552
```

```
qnorm(p = 0.95)
```

```
## [1] 1.644854
```

7. Suppose that X is a random variable that has PDF

$$f(x) = (\theta_0 + 1)\, x^{\theta_0} \quad 0 \leq x < 1$$

   where $\theta \geq 0$ is an unknown parameter.We want to perform a test of the hypotheses[2]

$$H_0 : \theta = 0 \quad H_1 : \theta > 0.$$

   We are feeling pretty lazy and so we only take one value randomly from this distribution. Let's call the value that results x.

a. What should the form of the test be? Should we reject when x is small? large? both large and small?
b. Calculate the P-value if $x = 0.2$.
c. Suppose we want to conduct the test at $\alpha = 0.05$. What values of x would lead to rejection of H_0?
d. For a level $\alpha = 0.05$ test, calculate the Type II error rate when $\theta_0 = 1$.

8. An urn contains ten marbles: an unknown number of them are white, the rest red. We wish to test:

$$H_0 : \text{exactly half are white}$$

   versus

$$H_1 : \text{more than half are white}$$

   We will draw randomly, without replacement, three marbles and reject H_0 if two or more are white.

[2]Note: the null is saying that $X \sim Unif(0, 1)$

a. Find the Type I error rate of this test.

Hint: we reject if X, the number of white marbles drawn is ≥ 2. You will need to write the PMF of X under H_0 to calculate the P-value. To find it, remember this is a problem where you are sampling from a finite population where the items can be one of two types. Think back to the goldfish (tagged, untagged) problem from STAT 340.

b. Find the Type II error rate in two situations.

- 60% of the marbles in the urn are white
- 70% of the marbles in the urn are white.

9. True or false and support your answer. If something is true, it must always be true and you must explain why it is so. To show something is false, you can either explain why or give a counter example. **How you explain your answers matters on this type of question. Also, we will only go by what you have written, not what you meant by what you wrote.**

a. _____

Suppose we have one observation, X, from the PDF

$$f(x) \quad = \quad \theta_0 x^{\theta_0 - 1}, \quad 0 \leq x < 1.$$

A 95% confidence interval estimator for θ_0 is $[\frac{\ln(0.025)}{\ln(X)}, \frac{\ln(0.975)}{\ln(X)}]$.

b. _____

A Type I error occurs anytime the test statistic falls in the rejection region of a test.

c. _____

Suppose $X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} Exp(\lambda_0)$. Then

$$T = n\, X_{min}$$

is an unbiased estimator of $\frac{1}{\lambda_0}$

d. _____

A butcher weighs steaks by placing his thumb on the scale in addition to the steaks. This results in an error in the measured weight of the steak. We call this type of error "bias".

e. _____

A P-value is a random variable.