DISCRETE

Discrete (countable//countably infinite)

- Probability Distribution (list of all possible values of X and their probabilities of occurring)
- $0 \leq P(X=x) \leq 1$, for all of x
- $\sum P(X=x) = 1$, for all of x summed = 1
- Probability mass function: $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$

Discrete Random Variables (X):

- **expected value** of a random variable ➔ $E(X)$ ⇔ μ (theoretical mean) ⇔ $\sum x * P(X=x)$
- expectation of a function ➔ $E[g(X)]$ ⇔ $\sum g(x) * P(X=x)$
- **variance** ➔ $\sigma^2 = \sum[(x-\mu)^2 p(x)]$ ⇔ $E[(X-\mu)^2]$ ⇔ $E(X^2) - [E(X)]^2$ ⇔ $E(X^2) - \mu^2$)

  ↳ {distance from any x value}
- **standard deviation** ➔ $\sigma = \sqrt{\sum_x [(x-\mu)^2 * P(X=x)]}$ *OR* $\sqrt{\sigma^2}$

Bernoulli Distribution (for using values that can only 1 or 0, independent, success or failure)

- X = 1 is a success, while X = 0 is a failure (where the 1 or 0 is x)
- $\int_2^4 cx^3 \, dx$ $P(X=x) = p^x(1-p)^{1-x}$
- *D* $\sigma^2 = p(1-p)$

Binomial Distribution (number of successes in *n* (fixed) independent Bernoulli trials)

- P(Success) = p, P(Failure) = 1-p, *x* represents the number of successes in *n* trials
- **(PMF)** ➔ $P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$ *ALSO* $X \sim B(n,p)$
- $\mu = np$ *AND* $\sigma^2 = np(1-p)$

Geometric Distribution (number of trials needed to get the *first* success in repeated Bernoulli trials)

- X = number of trails needs to get first success
- *x -1* is the number of failures $\mathbf{(1-p)^{x-1}}$, *xth* trial is the first success **(p)**
- **(PMF)** ➔ $P(X = x) = (1-p)^{x-1} * p$, for x = 1,2,3,…
- $\mu = \frac{1}{p}$ *AND* $\sigma^2 = \frac{1-p}{p^2}$ *AND* mode is always 1
- cumulative distribution function ➔

  $F(x) = P(X \leq x) = 1 - (1 - p)^x$, for x = 1,2,3,…

**over a range (i.e. - $P(X \leq 3)$), add the values together

Poisson Distribution ([X] number of events in a fixed unit of *time, volume, unit of measure*) {Approximation}

- Events occur independently & randomly
- **(PMF)** ➔ $\frac{\lambda^x e^{-\lambda}}{x!}$ for x = 0,1,2,…
- $\mu = \lambda$ *AND* $\sigma^2 = \lambda$
- Probability histogram will be skewed right as λ gets closers to zero and symmetrical as λ gets larger

**over a range, add values together

******Binomial and Poisson Relationship******

- BD tends towards PD as $n \longrightarrow \infty$, $p \longrightarrow 0$, and np stays constant
- PD with $\lambda = np$ closely approximates the BD if *n* is large and *p* is small

Negative Binomial Distribution (number of trails needed to get a fixed number of successes)

- X is the trial number for the *rth* success

for the *r*th success to occur on the *x*th trial

PMF ➔ $\binom{x-1}{r-1} p^{r-1}(1-p)^{(x-1)-(r-1)}$ x = r, r+1,…

$\mu = \frac{r}{p}$ *AND* $\sigma^2 = \frac{r(1-p)}{p^2}$

CONTINUOUS

Continuous (every value in an interval, including infinite decimal places)

Continuous Probability Distribution

- probability density function (pdf) is a model of a continuous random variable with a curve *f*(*x*)
- probabilities are areas under the curve $P(a<X<b)$
- $P(X=a) = 0$
- *f*(*x*) must be greater than 0 for all of x
- total area cannot be greater than 1 (area under curve)
- integrating will be necessary, vs summations
- to find the median, integrate from lower to M so that the total value is 1/2
- for the cumulative distribution function, integrate from lower to x for f(t), so the integral from lower bound to x, integrate t to find the value of x
- $\mu = \int_{-\infty}^{\infty} x f(x)\, dx$ *AND* $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$

example:

random variable $x = f(x) = cx^3$ for $2 \leq x \leq 4$, and the integral from negative inf to pos inf but be equal to one.

➔ c = 1/60 (makes this entire thing sum to 1)

to find this same problem from P>3, integrate from 3 to

- Cumulative distribution function ➔ $\int_2^x \frac{1}{60} t^3 \, dt$

Uniform Distribution

- *f*(*x*) constant over the possible values of *x*
- Area = range from a to b multiplied by *f*(*x*), sums to 1
- PDF ➔ *f*(*x*) {1/b-a for c ≤ x ≤ d, and 0 else where}
- CDF ➔ $\frac{x-a}{b-a}$ for $a \leq x \leq b$

  0 for $x < a$

  1 for $x > b$
- *f*(*x*) = 1/(b-a), mean AND median = b+a/2
- $\sigma^2 = (1/12)\,(b-a)^2$

Weibull Distribution

-generalization of exponential distribution

- $f(x) = \kappa\lambda^\kappa x^{\kappa-1} e^{-(\lambda x)^\kappa}$

  x > 0, κ - scale param, λ - shape param
- PDF ➔ $\frac{k}{\lambda}\left(\frac{x-\theta}{\lambda}\right)^{k-1} e^{-(\frac{x-\theta}{\lambda})^k}$
- CDF ➔ $1 - e^{-(x/\lambda)^k}$

  $F\{x\} = \{0 \text{ for } x \leq 0;\ 1 - e^{-(\lambda x)^\wedge \kappa} \text{ for } x > 0\}$
- $\mu = \lambda\Gamma\left(1+\frac{1}{k}\right)$ *AND* $\sigma^2 = \lambda^2\left[\Gamma\left(1+\frac{2}{k}\right) - \left(\Gamma\left(1+\frac{1}{k}\right)\right)^2\right]$

Exponential Distribution (time taken between two events occurring)

- λ = average number of events in one unit of time
- $P(X>x) = e^{-\lambda x}$, $P(X<x) = 1 - e^{-\lambda x}$
- $\mu = 1/\lambda$ *AND* $\sigma^2 = 1/\lambda^2$
- PDF ➔ $\lambda e^{-\lambda x}$ $x \geq 0,$
- CDF ➔ $1 - e^{-\lambda x}$ $x \geq 0,$

  Otherwise 0, for x < 0 (For PDF and CDF)

Beta Distribution (modeling random probabilities and proportions)

- PDF ➔ $\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$

  for 0 ≤ x ≤ 1, α shape param, β > 0
- CDF ➔ $\frac{B(x;\alpha,\beta)}{B(\alpha,\beta)} = I_x(\alpha,\beta)$
- $\mu = \frac{\alpha}{\alpha+\beta}$ *AND* $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Normal Distribution

Z-score $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ (Distance from mean, area is probability)

T-score $T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

$\hat{p}$ - sample proportion (targets [estimate] the population proportion, p)

To re-standardize: $Z = \frac{x - \mu}{\sigma}$ (Maps to a ND with μ centered at 0, σ=1)

Central Limit Theorem

Requirements (with a population with μ and σ):

- n>30 (implies it is normally distributed)
- n≤30 and population is normally distributed, then sample is ND
- n≤30 and population distribution is unknown, other method

Test Statistic for Proportion, p:

When: Random Sample, np ≥ 10, nq ≤ 10, n, $\hat{p} = x/n$ (x = # of suc)

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p*q}{n}}}$$

p = population proportion (given)
q = 1-p

p-value = area in the tail(s) from the Z-value
Always use Z-score

Estimate of Population Mean, μ:

When: Random Sample, σ known or unknown, & n >30 (& pop. ND)

Point Estimate for μ = $\bar{X}$

Margin of Error: E= $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

The estimate = $\bar{X} \pm E$ $(\bar{X} - E < \mu < \bar{X} + E)$

When: Random Sample, σ not known , and n > 30 (OR pop ND)

Point Estimate for μ = $\bar{X}$

Margin of Error: E= $T_{\alpha/2}\frac{s}{\sqrt{n}}$

The estimate = $\bar{X} \pm E$ $(\bar{X} - E < \mu < \bar{X} + E)$

Confidence intervals with population variance, σ^2:

Chi-Square: $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

Confidence Interval $\frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$

If α = 0.05, X^2_L at .975 on tables, X^2_R at .025 on table, also uses degrees of freedom

Hypothesis Testing: (Testing whether or not a claim is valid)

Reject H_0 if p-value is ≤ α, *Fail to Reject* H_0 if p-value is > α
Reject H_0 if Z-value (calculated) falls in the Rejection Region, *FTR* otherwise (Traditional method); compare alpha Z-score with calculated Z-score

Proportions:

BASICS

$P(X=x) \Leftrightarrow p(x)$
$E(X) = \mu$
$var(X) = \sigma^2$
$sd(X) = \sigma$
Combinations formula ➔ $\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$ (AKA Binomial coefficient)
Calculate T-score by table with degrees of freedom (n-1) and α; α/2 for two-tailed test
iid - independent and identically distributed random variables

	n ≥ 30	n < 30
X_i ~ Normal	$Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	
	$Z_{\alpha/2}\frac{s}{\sqrt{n}}$	$T_{\alpha/2}\frac{s}{\sqrt{n}}$
X_i ~ Not Normal or unknown	$Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ (by CLT)	other method
	$Z_{\alpha/2}\frac{s}{\sqrt{n}}$ (by CLT)	

Examples:
Given 2 means and variances, the average of the two: $E(\bar{X})=(x_1+x_2)/2$ the variance being $(x_1+x_2)(1/2)^2$, the probability that the averages will be between two numbers = $(x_1 - \bar{X})$ / sqrt(Var($\bar{X}$) - $(x_2 - \bar{X})$ / sqrt(Var($\bar{X}$)
The probability that both are less than a number: x(the given number) - each x1 and x2 / sqrt(each of their variances). Take these two probability numbers and multiple them to get the probability.

Given a hypothesis (a claim), cut-off point, $\bar{X}$, and sd: the type 1 error is P($\bar{X}$ >= cut-off point) | H0 is true) = P (Z >= (cut-off - claim number) / sd / sqrt(n). 1 - calc number = type 1 error (alpha)
To find the power: P($\bar{X}$ > cut-off point) | mu = new claim #) = P (Z>= (cut-off - new claim #) / (sd) / sqrt(n).
The p-value: P($\bar{X}$ >= cut-off point) | H0) = 1-P(Z<sample average - mu) / sd / sqrt(n) (should be a percentage, between 0 and 1)

CI: check based on x_bar or x, calculate as usual

Claim: mu_a = mu_b, opp: mu_a < mu_b to find the n value, setup for z-score but solve for n, n > {(z_score)(sd)sqrt(2) / difference in mu)}^2