EXAM HANDOUT

Counting Outcomes:

Permutations with replacement	$P_r(n,k) = \overbrace{n \cdot n \cdot \ldots \cdot n}^{k \text{ terms}} = n^k$
Permutations without replacement	$P(n,k) = \overbrace{n(n-1)(n-2) \cdot \ldots \cdot (n-k+1)}^{k \text{ terms}} = \frac{n!}{(n-k)!}$
Combinations without replacement	$C(n,k) = \binom{n}{k} = \frac{P(n,k)}{P(k,k)} = \frac{n!}{k!(n-k)!}$
Combinations with replacement	$C_r(n,k) = \binom{k+n-1}{k} = \frac{(k+n-1)!}{k!(n-1)!}$

Union of events:

Probability of a union	$\boldsymbol{P}\{A \cup B\} = \boldsymbol{P}\{A\} + \boldsymbol{P}\{B\} - \boldsymbol{P}\{A \cap B\}$ For mutually exclusive events, $\boldsymbol{P}\{A \cup B\} = \boldsymbol{P}\{A\} + \boldsymbol{P}\{B\}$

To scale this up for mutually exclusive events just add their probabilities together. To scale this up for non-mutually exclusive events, use principle of inclusion and exclusion

Intersection of independent events

Independent events	$\boldsymbol{P}\{E_1 \cap \ldots \cap E_n\} = \boldsymbol{P}\{E_1\} \cdot \ldots \cdot \boldsymbol{P}\{E_n\}$

Intersection of dependent events are dealt with by conditional probability.

Intersection, general case

$$\boldsymbol{P}\{A \cap B\} = \boldsymbol{P}\{B\}\,\boldsymbol{P}\{A \mid B\}$$

Conditional Probability

DEFINITION 2.15

Conditional probability of event A given event B is the probability that A occurs when B is *known to occur*.

Conditional probability

$$\boldsymbol{P}\{A \mid B\} = \frac{\boldsymbol{P}\{A \cap B\}}{\boldsymbol{P}\{B\}}$$

For independent events, P { A | B } = P { A }

Bayes Rule

Bayes Rule

$$\boldsymbol{P}\{B \mid A\} = \frac{\boldsymbol{P}\{A \mid B\}\,\boldsymbol{P}\{B\}}{\boldsymbol{P}\{A\}}$$

Total Probability

Law of Total Probability

$$\boldsymbol{P}\{A\} = \sum_{j=1}^{k} \boldsymbol{P}\{A \mid B_j\}\,\boldsymbol{P}\{B_j\}$$

In case of two events ($k = 2$),

$$\boldsymbol{P}\{A\} = \boldsymbol{P}\{A \mid B\}\,\boldsymbol{P}\{B\} + \boldsymbol{P}\left\{A \mid \overline{B}\right\}\boldsymbol{P}\left\{\overline{B}\right\}$$

Bayes Rule for two events

$$\boldsymbol{P}\{B \mid A\} = \frac{\boldsymbol{P}\{A \mid B\}\,\boldsymbol{P}\{B\}}{\boldsymbol{P}\{A \mid B\}\,\boldsymbol{P}\{B\} + \boldsymbol{P}\left\{A \mid \overline{B}\right\}\boldsymbol{P}\left\{\overline{B}\right\}}$$

Random Variables

Distribution	Discrete	Continuous
Definition	$P(x) = P\{X = x\}$ (pmf)	$f(x) = F'(x)$ (pdf)
Computing probabilities	$\boldsymbol{P}\{X \in A\} = \sum_{x \in A} P(x)$	$\boldsymbol{P}\{X \in A\} = \int_A f(x)dx$
Cumulative distribution function	$F(x) = \boldsymbol{P}\{X \leq x\} = \sum_{y \leq x} P(y)$	$F(x) = \boldsymbol{P}\{X \leq x\} = \int_{-\infty}^{x} f(y)dy$
Total probability	$\sum_x P(x) = 1$	$\int_{-\infty}^{\infty} f(x)dx = 1$

If X is a discrete variable,

$P(X \leq x) = F(x)$ $\quad$ $P(X \geq x) = 1 - P(X < x) = 1 - P(X \leq x-1) = 1 - F(x-1)$

$P(X = x) = F(x) - F(x-1)$ $\quad$ $P(a \leq X \leq b) = F(b) - F(a)$

For a continuous distribution,

$P(X = x) = 0$ $\quad$ $P(X \leq x) = P(X < x) = F(x)$ $\quad$ $P(X \geq x) = P(X > x) = 1 - F(x)$

$P(a \leq X \leq b) = F(b) - F(a)$ $\quad$ $P(x-h \leq X \leq x+h) \approx 2h * f(x) \approx P(X = x)$(If h is really small)

Joint Distributions:

The following table shows how to calculate marginal distributions, check for independence and how to calculate probabilities of random vectors using joint probability distributions

Distribution	Discrete	Continuous
Marginal distributions	$P(x) = \sum_y P(x,y)$ $P(y) = \sum_x P(x,y)$	$f(x) = \int f(x,y)dy$ $f(y) = \int f(x,y)dx$
Independence	$P(x,y) = P(x)P(y)$	$f(x,y) = f(x)f(y)$
Computing probabilities	$\boldsymbol{P}\{(X,Y) \in A\}$ $= \sum\sum_{(x,y) \in A} P(x,y)$	$\boldsymbol{P}\{(X,Y) \in A\}$ $= \iint_{(x,y) \in A} f(x,y)\,dx\,dy$

Properties (or Moments) of a Distribution

Discrete	**Continuous**
$\mathbf{E}(X) = \sum_x xP(x)$	$\mathbf{E}(X) = \int xf(x)dx$
$\mathrm{Var}(X) = \mathbf{E}(X-\mu)^2 = \sum_x (x-\mu)^2 P(x) = \sum_x x^2 P(x) - \mu^2$	$\mathrm{Var}(X) = \mathbf{E}(X-\mu)^2 = \int (x-\mu)^2 f(x)dx = \int x^2 f(x)dx - \mu^2$
$\mathrm{Cov}(X,Y) = \mathbf{E}(X-\mu_X)(Y-\mu_Y) = \sum_x \sum_y (x-\mu_X)(y-\mu_Y)P(x,y) = \sum_x \sum_y (xy)P(x,y) - \mu_x\mu_y$	$\mathrm{Cov}(X,Y) = \mathbf{E}(X-\mu_X)(Y-\mu_Y) = \iint (x-\mu_X)(y-\mu_Y)f(x,y)\,dx\,dy = \iint (xy)f(x,y)\,dx\,dy - \mu_x\mu_y$

Std(X) is square root of Var(X)

DEFINITION 3.9

Correlation coefficient between variables X and Y is defined as

$$\rho = \frac{\mathrm{Cov}(X,Y)}{(\mathrm{Std}X)(\mathrm{Std}Y)}$$

Properties of expectations

$$\mathbf{E}(aX + bY + c) = a\,\mathbf{E}(X) + b\,\mathbf{E}(Y) + c$$

In particular,

$$\begin{aligned} \mathbf{E}(X+Y) &= \mathbf{E}(X) + \mathbf{E}(Y) \\ \mathbf{E}(aX) &= a\,\mathbf{E}(X) \\ \mathbf{E}(c) &= c \end{aligned}$$

For **independent** X and Y,

$$\mathbf{E}(XY) = \mathbf{E}(X)\,\mathbf{E}(Y)$$

Properties of variances and covariances

$$\mathrm{Var}(aX + bY + c) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y)$$

$$\begin{aligned}&\mathrm{Cov}(aX + bY, cZ + dW)\\ &\quad = ac\,\mathrm{Cov}(X, Z) + ad\,\mathrm{Cov}(X, W) + bc\,\mathrm{Cov}(Y, Z) + bd\,\mathrm{Cov}(Y, W)\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathrm{Cov}(Y, X)\\ \rho(X, Y) &= \rho(Y, X)\end{aligned}$$

In particular,

$$\begin{aligned}\mathrm{Var}(aX + b) &= a^2\,\mathrm{Var}(X)\\ \mathrm{Cov}(aX + b, cY + d) &= ac\,\mathrm{Cov}(X, Y)\\ \rho(aX + b, cY + d) &= \rho(X, Y)\end{aligned}$$

For independent X and Y,

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

Only a large variance will allow a Variable to vary greatly from the expected value. This is shown by Chebyshev's Inequality

Chebyshev's inequality

$$\boldsymbol{P}\left\{|X - \mu| > \varepsilon\right\} \leq \left(\frac{\sigma}{\varepsilon}\right)^2$$

for any distribution with expectation μ and variance σ^2 and for any positive ε.

This shows that a Variable with a high variance has bigger risk of varying from the expected amount by a large value.

Sample Statistics (Used as estimators for Population Parameters)

Mean

DEFINITION 8.3

Sample mean $\bar{X}$ is the arithmetic average,

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

Variance and Std. Deviation

DEFINITION 8.8

For a sample $(X_1, X_2, \ldots, X_n)$, a **sample variance** is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2. \tag{8.4}$$

It measures variability among observations and estimates the population variance $\sigma^2 = \mathrm{Var}(X)$.

Sample standard deviation is a square root of a sample variance,

$$s = \sqrt{s^2}.$$

It measures variability in the same units as X and estimates the population standard deviation $\sigma = \mathrm{Std}(X)$.

Alt formula for Variance:

$$\boxed{s^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}.}$$

Quantiles, percentiles and quartiles

DEFINITION 8.7

A p-**quantile** of a population is such a number x that solves equations

$$\begin{cases} \boldsymbol{P}\{X < x\} & \leq & p \\ \boldsymbol{P}\{X > x\} & \leq & 1-p \end{cases}$$

A **sample** p-**quantile** is any number that exceeds at most $100p\%$ of the sample, and is exceeded by at most $100(1-p)\%$ of the sample.

A γ-**percentile** is (0.01γ)-quantile.

First, second, and third **quartiles** are the 25th, 50th, and 75th percentiles. They split a population or a sample into four equal parts.

A **median** is at the same time a 0.5-quantile, 50th percentile, and 2nd quartile.

Shape of a distribution (comparing mean and median)

$$\begin{array}{lcl} \text{Symmetric distribution} & \Rightarrow & M = \mu \\ \text{Right-skewed distribution} & \Rightarrow & M < \mu \\ \text{Left-skewed distribution} & \Rightarrow & M > \mu \end{array}$$

IQR and outliers.

DEFINITION 8.10

An **interquartile range** is defined as the difference between the first and the third quartiles,

$$IQR = Q_3 - Q_1.$$

It measures variability of data. Not much affected by outliers, it is often used to detect them. IQR is estimated by the *sample interquartile range*

$$\widehat{IQR} = \hat{Q}_3 - \hat{Q}_1.$$

Any samples that are less than $Q_1 - 1.5(IQR)$ or more than $Q_3 + 1.5(IQR)$ can be treated as potential outliers.

Standard error of any estimator is its std deviation.

$$\begin{array}{rcl} \sigma(\hat{\theta}) & = & \text{standard error of estimator } \hat{\theta} \text{ of parameter } \theta \\ s(\hat{\theta}) & = & \text{estimated standard error } = \hat{\sigma}(\hat{\theta}) \end{array}$$

Graphical Statistics:

Can be used to visualize the given samples to make observations about the nature of the population

- Histograms are bar charts for columns for each bin
  - If height of bin is freq count: Frequency histogram
  - If height of bin is proportion of data: Relative Frequency histogram
  - Each sample value can have its own bin, or you can have multiple nearby values in one bin.
  - You can use histograms to guess shape of distribution.
- Stem and leaf plots
  - Choose the stem such that the values are not all limited to one stem.
  - Distribute values to each stem and sort he leaf values.
  - You can use this to calculate mean, median and guess shape of distribution
  - It can also be used to compare two distributions
- Boxplots
  - Boxplots are based on 5-point summaries
    - $< \min(X_i), \widehat{Q_1}, \widehat{M}, \widehat{Q_3}, \max(X_i) >$
  - Represent sample mean with small cross. Draw a box between sample Q1 and Q3 and draw a line for sample median. Draw whiskers to smallest sample and largest sample that is within the 1.5 IQR range. Draw dots for all samples outside the 1.5 IQR range.
  - Can be used to compare multiple distributions.

Entropy – The Mean Information

- For any random variable X with a probability mass function (pmf) p_i it is thus possible to measure the information content of each outcome

$$I(x_i) = I(p_i) = -\log_b(p_i)$$

- The expected information of an experiment (i.e. of the unknown value of X) can be computed

$$H(X) = E\left[I(X)\right] = \sum_{x \in X} p(x) I(x)$$
$$= -\sum_{x \in X} p(x) \log_b p(x)$$

Relative Entropy

- Relative Enthropy (Kullback-Leibler distance) is a measure of the difference of two distributions

$$D(p \| q) = E_p\left[\log_b\left(p(x)/q(x)\right)\right]$$
$$= \sum_x p(x) \log_b\left(p(x)/q(x)\right)$$

  - Measures not the difference in the amount of information but difference in the information itself
  - Both distributions have to be defined over the same domain
  - Is always positive and zero only if the two distributions are identical

Bernoulli distribution

- Used to model experiments with a binary outcome (yes/no, pass/fail, true/false)
  - Called Bernoulli Trials
- Random variable can take values 0 (fail) or 1 (pass)
- p = probability of success

$$P(x) = \begin{cases} 1-p & if \quad x = 0 \\ p & if \quad x = 1 \end{cases}$$
$$E(X) = p$$
$$Var(X) = p(1-p)$$

Binomial Distribution

- It is used to model number of success in a sequence of **independent** Bernoulli trials
- Models the probability of x successes in n trials
- p = probability of success; n = number of trials

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$E(X) = np$$
$$Var(X) = np(1-p)$$

Geometric Distribution

- It is used to model number trials needed to achieve the first success in a sequence of **independent** Bernoulli trials
- Models the probability of xth successive trial resulting in a success
- p = probability of success

$$P(x) = (1-p)^{x-1}p$$
$$E(X) = \frac{1}{p}$$
$$Var(X) = \frac{1-p}{p^2}$$

Negative Binomial Distribution

- It is used to model number of trials needed to obtain k success in a sequence of **independent** Bernoulli trials
- Models the probability of xth successive trial resulting in the kth success
- p = probability of success; k = number of success

$$P(x) = \binom{x-1}{k-1}(1-p)^{x-k}p^k$$
$$E(X) = \frac{k}{p}$$
$$Var(X) = \frac{k(1-p)}{p^2}$$

Side Note: Calculating Neg. Binomial in Practice

- If X follows Negative Binomial(k, p)
  - P(X = x) = prob of needing x trials for k success = prob of kth trial being success * prob of getting k-1 success in x-1 trials = p * P(Y = k-1)
    - Where Y follows Binomial(x-1, p)
  - P(X ≥ x) = prob of needing ≥ x trials for k success = prob of x-1 trials not having ≤ k-1 success = P(Y ≤ k-1)
    - Where Y follows Binomial(x-1, p)

Poisson Distribution

- It is used to model number rare events occurring within a fixed period of time
- Models the probability of x rare events occurring in a fixed period of time if we know the frequency at which the events occur on average
- λ = frequency (average number of events in a fixed time period)

$$P(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$
$$E(X) = \lambda$$
$$Var(X) = \lambda$$

Side Note: Poisson Approx. of Binomial

- If the number of trials is large and the probability of success is low, then we can use Poisson Distribution to approximate the Binomial Distribution
  - Also works if probability of failure is very low
- $\lim\limits_{\substack{n\to\infty \\ p\to 0 \\ np\to\lambda}} \binom{n}{x} p^x (1-p)^x = e^{-\lambda}\frac{\lambda^x}{x!}$
- Can use this approximation if n >= 30 and p <= 0.05

Uniform Distribution

- Used to model scenarios where outcome lies within a given interval (a,b) and all outcomes are equally likely.
- If interval is (0,1) it is called standard uniform distribution

$$f(x) = \frac{1}{b-a}, a < x < b$$
$$E(X) = \frac{a+b}{2}$$
$$Var(X) = \frac{(b-a)^2}{12}$$

Exponential Distribution

- Used to model time (or separation) between events occurring at frequency λ (rate at which the events occur)

$$f(x) = \lambda e^{-\lambda x}, x > 0$$
$$E(X) = \frac{1}{\lambda}$$
$$Var(X) = \frac{1}{\lambda^2}$$

Gamma Distribution

- Used to model total time of multistage processes with α steps (shape parameter) where time of each step can be modeled as a Exponential distribution with frequency λ.

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \quad if\ \alpha > 0\ x > 0$$
$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy \quad if\ \alpha > 0$$
$$\text{also } \Gamma(\alpha) = (\alpha - 1)! \quad if\ \alpha\ is\ a\ positve\ integer$$
$$E(X) = \frac{\alpha}{\lambda}$$
$$Var(X) = \frac{\alpha}{\lambda^2}$$

- Please note that Gamma(1, λ) = Exponential(λ)

Side-Note: Gamma-Poisson Formula

- Can be used to simplify calculation of probabilities of RV T with Gamma Distribution.

$$\{T > t\} = \{X < \alpha\}$$

- Where is T has Gamma distribution with parameters α (number of events) and λ (frequency of each event).
- X models the number of events that occurs before time t. It has Poisson distribution with parameter λt.

  So,

$$P\{T > t\} = P\{X < \alpha\}$$
$$P\{T \le t\} = P\{X \ge \alpha\}$$

  Where T has Gamma(α, λ) distribution and X has Poisson(λt) distribution

Normal Distribution

- Used to model a large number of scenarios
  - Sums, averages or errors: Mainly due to CLT
  - Naturally occurring phenomena
- Allows you to model a scenario on the basis of expectation μ (location parameter) and standard deviation σ (scale parameter)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)} - \infty < x < \infty$$
$$E(X) = \mu$$
$$Var(X) = \sigma^2$$

Central Limit Theorem

- If you have a Random variable that is expressed as a sum of large number of independent random variables (usually >= 30), then you can use this theorem to model them as a Normal Distribution.

Theorem 1 (Central Limit Theorem) *Let $X_1, X_2, \ldots$ be independent random variables with the same expectation $\mu = \mathbf{E}(X_i)$ and the same standard deviation $\sigma = \mathrm{Std}(X_i)$, and let*

$$S_n = \sum_{i=1}^{n} X_i = X_1 + \ldots + X_n.$$

As $n \to \infty$, the standardized sum

$$Z_n = \frac{S_n - \mathbf{E}(S_n)}{\mathrm{Std}(S_n)} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

converges in distribution to a Standard Normal random variable, that is,

$$F_{Z_n}(z) = \boldsymbol{P}\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} \le z\right\} \to \Phi(z)$$

for all z.

Side-Note: Using Normal to Approx. Binomial

- A binomial distribution is a sum of n Bernoulli trials. So if n is large (>=30) but p is not small enough (or large enough) to use Poisson approximation (0.05 <= p <= 0.95) then we can model the binomial as a sum of Bernoulli distributions with mean p and variance p(1 - p).
- So by Central Limit Theorem,

$$\mathrm{Binomial}(n, p) \approx Normal\left(\mu = np, \sigma = \sqrt{np(1-p)}\right)$$

- This normal distribution can be calculated by converting it to a standard normal distribution

Side-Note: Using Normal to Approx. Gamma

- Similarly if α is large enough the we can use CLT to approx. a Gamma distribution using a normal distribution

$$Gamma(\alpha, \lambda) \approx Normal\left(\frac{\alpha}{\lambda}, \sqrt{\frac{\alpha}{\lambda^2}}\right)$$