Lecture Notes – Probability and Statistics – Ahmed Nouralla

Lecture 1

- **Probability Space** is a triple (Ω, A, P)
  - Ω: the set of all possible outcomes of an experiment.
  - A: the σ-algebra of events, a subset of Ω, also called the "favorable outcomes"
    - Closed under complement, union and intersection.
  - P: A ⟶ [0, 1]
    - A function that assigns a number to an event.
    - Probability of an event X happening = $\frac{Number\ of\ favorable\ outcomes\ in\ X}{Number\ of\ total\ outcomes\ in\ \Omega}$
- **Combination** (Binomial coefficient)
  - Number of ways to choose k elements from a set of n elements, where order doesn't matter.
  - $\binom{n}{k} = C(n,\ k) = \frac{n!}{k!(n-k)!}$
- **Permutation/Arrangement**
  - Number of ways to select and arrange k elements (in a specific order) from a set of n elements.
  - $A(n,k) = \frac{n!}{(n-k)!}$
- **Useful identities**
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ "∪ can be replaced with + and ∩ can be omitted"
  - $P(\overline{A+B}) = P(\bar{A}\bar{B}),\ P(\overline{AB}) = P(\bar{A} + \bar{B})$
  - $P(A \mid B) = \frac{P(AB)}{P(B)}$
  - $P(A + B + C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$
  - A, B are disjoint $\Longrightarrow P(AB) = 0, P(A + B) = P(A) + P(B)$
  - A, B are independent $\Longrightarrow P(AB) = P(A)P(B), P(A + B) = P(A) + P(B) - P(A)P(B)$
  - $P(AB) \geq P(A) + P(B) - 1$
  - $P(A_1A_2 \ldots A_n) \geq P(A_1) + P(A_2) + \cdots + P(A_n) - (n - 1)$

Lecture 2

- **Law of total probability**
  - Let $A_1, A_2, \dots A_n$ be a set of
    - Pairwise exclusive(disjoint) events = no two events happen together = $P\left(A_i A_j\right) = 0, i \neq j$.
    - Collectively exhaustive = at least one of them occurred = $\bigcup_{i=1}^{n} P(A_i) = 1$
  - Then, for any event B in the same probability space we have:
    - $P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n)$
- **Bayes' theorem**

$$P(A_k|B) = \frac{P(A_k B)}{P(B)} = \frac{P(B|A_k)P(A_k)}{P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n)}$$

- **Bernoulli trials (process)**
  - When doing N (fail/success) experiments with p $(0 < p < 1)$ being the probability of success and $q = 1 - p$ being the probability of failure. The probability of k success among N trials is denoted as $P(\mu_N = k) = \binom{n}{k} p^k q^{n-k}$
- **Bernoulli scheme(shift):**
  - A generalization of Bernoulli trials to more than two possible outcomes.
  - Instead of failure/success, we can have $\boldsymbol{n}$ possible outcomes.
  - When doing $\boldsymbol{m}$ experiments (each one has n possible outcomes, each outcome can happen with a probability p_i, such that $\sum_{i=1}^{n} p_i = 1$), the probability of getting the i^{th} outcome k_i times is given by

$$P_m(k_1, k_2, \dots, k_n) = \frac{m!}{k_1!\ k_2! \dots k_n!} p_1^{k_1} p_2^{k_2} \dots p_n^{k_n}$$

Lecture 3

- **Random variable (X/ξ)**
  - A variable that takes different values depending on the result of some random experiment, each value has some probability associated with it.
  - Mathematically speaking, it is a function X: Ω → E from the set of possible outcomes to a measurable space.
    - X (an event) = the random variable value associated with that event.
  - Has two types: **discrete** if E is countable, in this case, it can be described using **"Probability Mass Function (PMF)"**, otherwise it's **continuous** and can be described using **"Probability Density Function (PDF)".**
    - Countable means finite or countably infinite
      - A set is countably infinite if its elements can be put in one-to-one correspondence with the set of natural numbers.
    - **Common notation for the PMF**
      - The first row contains all the values the random variable can take (they can be infinite).
      - The second row contains the corresponding probability.
      - ω ∈ Ω represents an event/outcome from the set of all outcomes, for each ω, there is a corresponding value that X can take. ω can be written above the first row.

$$X \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix} = \begin{pmatrix} x_i \\ P_i \end{pmatrix}$$

- **Example:**
  - For rolling a die three times, let X be the random variable representing the sum of all outcomes, let $\Omega = \{[a, b, c] \mid a, b, c \in \mathbf{N} \ \& \ 1 \le a, b, c \le 6\}$ be the set of all outcomes, then $X([5, 1, 3]) = 9$
    - $[a, b, c]$ means that we get a, b, c in the first, second, third roll, respectively, it is not a standard representation; an event ω can be represented in any convenient way.
- **Cumulative distribution function (CDF): $F_X(x) = P(X < x)$** "some sources require the relation to be ≤, therefore, some properties will vary, but the concept is still the same"
  - A function describing a random variable X; takes a number $x \in \mathbf{R}$ and returns the probability that the random variable takes a value less than x.
  - For continuous random variables, $F_X(x)$ is the area under the curve of $P(X = \omega)$ and to the left of x
  - $F_X(x)$ is continuous on the left: it has a staircase graph, $\lim_{x \to \infty} F_X(x) = 1$, $\lim_{x \to -\infty} F_X(x) = 0$
  - $F_X(x)$ is increasing: $x_1 < x_2 \Longrightarrow F_X(x_1) \le F_X(x_2)$
- **Expected value of a random variable X** (denoted $E\ X$ or μ)
  - The expectation, arithmetic mean, average, or first-moment value we can get for X.
  - $E\ X = \sum_i x_i p_i = \sum_{\omega \in \Omega} X(\omega) P(\omega)$ for discrete case, exists only if the series converges
  - $E(cX) = c * E(X), E(c) = c. c = \text{const}$
  - $E(X \pm Y) = E\ X \pm E\ Y$ **"Expected value of a sum is a sum of expected values"**
    - This property is useful when it's hard to calculate the probability for each value that the random variable can take, in this case, we represent the random variable as a sum of n (preferably indicator) random variables x_i, and use $EX = EX_1 + \cdots + EX_n$

- **Variance** is a number that measures how far a set of numbers is spread out from their average.
    - The larger the variance, the more probable that the random variable can take a value far from it's expected (average) value.
    - $\mathrm{Var}\, X = \mathrm{E}(X - \mathrm{E}\, X)^2 = \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2$
    - $\sigma = \sqrt{Var\ X}$ is called the **standard deviation**.
    - $\mathrm{E}\, X^n$: n^{th} moment of the random variable X
    - $\mathrm{E}(X - \mathrm{Ex})^n$: n^{th} central moment.
    - $\mathrm{Var}\,(cX) = c^2\, Var\ X, c = const$
    - $Var\ (X + c) = Var\ X, c = const$
    - $Var\ c = 0, c = const$
    - $Var\ (X \pm Y) = Var\ X + Var\ Y \pm 2\ Cov(X, Y)$
    - $Var\ (X \pm Y \pm Z) = Var\ X + Var\ Y + Var\ Z \pm 2Cov(X, Y) \pm 2Cov(X, Z) \pm (\pm 2)Cov(Y, Z)$
- **For $\boldsymbol{n}$ identical random variables $\boldsymbol{X_i}$**
    - $E(\sum_{i=1}^{n} X_i) = n\, E(X_1)$
    - $Var\ (\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var\ X_i + \sum_{i \neq j} Cov(X_i, X_j)$
    - But, $X_1 + X_2 + \cdots + X_n \neq nX_1$
- **Indicator random variable:**
    - A random variable that takes one of two possible values: 0 or 1 depending on whether an event A happened or not.
    - $\mathrm{I}_A \sim \begin{pmatrix} 1 & 0 \\ \mathrm{P(A)} & 1 - \mathrm{P(A)} \end{pmatrix}$
    - $E\ I_A = E\ I_A^2 = P(A), Var\ I_A = P(A)(1 - P(A))$
- **Binomial distribution**: $\boldsymbol{Bin(n, p)}$
    - The discrete probability distribution of the number of successes in a sequence of n Bernoulli trials with p being the probability of success and $q = 1 - p$.
    - A random variable X following the binomial distribution is denoted as $X \sim Bin(n, p)$
    - **PMF** for X:

$$\mathrm{X} \sim \begin{pmatrix} 0 & \cdots & \mathrm{k} & \cdots & n \\ \mathrm{q}^n & \cdots & \binom{n}{k} p^k (1-p)^{n-k} & \cdots & p^n \end{pmatrix}$$

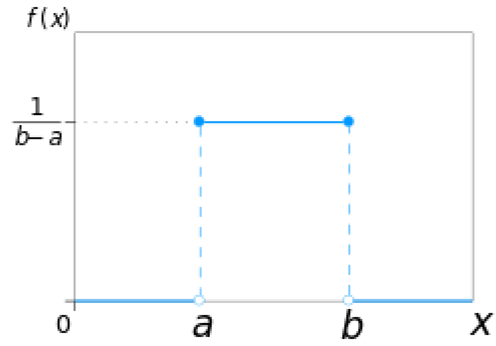
    - $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},\ E\ X = np,\ Var\ X = npq$
    - $X \sim c * Bin(n, p) \Longrightarrow X \sim N(cnp, c^2)$
    - $X_1, X_2, \dots, X_k \sim Bin(n, p) \Longrightarrow X_1 + X_2 + \cdots + X_k \sim Bin(nk, p)$

Lecture 4

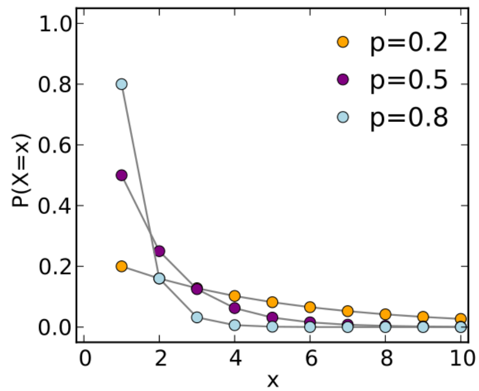
- **Probability distribution** is the function having the graph $p = P(X = x)$. With x on the x-axis "defined on some range" and p ($0 \leq p \leq 1$) on the y-axis.
  - It is a statistical function that describes all the possible values that a random variable can take within a given range.
  - It can be discrete (X can take a countable number of values) or continuous (otherwise), in both cases, $\sum_i p_i = \sum_{\omega \in \Omega} P(\omega)$ has to be equal (or to converge) to 1.
    - For discrete case, it's no use to draw a graph, although it's possible, but we use the PMF instead.
  - Avoid terms confusion.
  - There are several distribution functions, they serve different purposes and represent different data generation processes.
- **Continuous probability distribution**
  - CDF of a continuous random variable is given by
    - $F_X(x) = P(X < x) = \int_{-\infty}^{x} f_X(t)dt, x \in \mathbf{R}$
    - $F'_X(x) = f_X(x) \geq 0$ "Derivative of the CDF of X is the PDF of X"
  - $\int_a^b f_X(t)dt = P(a \leq x \leq b)$ gives the probability that a random variable is situated in between two different values (area under the PDF curve).
  - $\int_{-\infty}^{\infty} f_X(x)dx = 1$ "Area under the PDF is always 1"
  - $E\, h(X) = \int_{-\infty}^{\infty} h(x) f_X(x)dx$ "LOTUS"
    - Also works for discrete case.
- **Probability-generating function**: a power series representation of the PMF of a discrete random variable (X).
  - $g_X(t) = \sum_x t^x P(X = x) = E\, t^X, X \geq 0$
  - $g_X(0) = 0, g_X(1) = 1$
  - $E\, X = g'_X(1)$
  - $Var\, X = g''_X(1) + g'_X(1) - [g'_X(1)]^2$
- **Covariance**: a generalization of variance, measures how two variables tend to deviate from their expected values.
  - $\mathrm{Cov}(X, Y) = \sigma_{XY} = E\,(X - EX)(Y - EY) = E(XY) - E\,X * E\,Y$
  - $Cov(X, X) = Var\, X$
  - $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cox(X_2, Y)$
  - $Cov(aX + b, cY + d) = acCov(X, Y)$
- **Correlation coefficient**: similar to covariance, but has no unit, describes the degree of proportionality between the random variables
  - $\rho_{XY} = 1$ means: if X increases, Y will increase by the same amount.
  - $\rho_{XY} = -1$ means: if X increases, Y will decrease by the same amount and vice versa.
  - $\rho_{XY} = 0$ means that variables are not correlated.
  - $Corr(X, Y) = \rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var\, X * Var\, Y}}$
  - $-1 \leq Corr(X, Y) \leq 1$
- **Independent random variables**
  - X, Y are independent $\Leftrightarrow \forall (x \in X, y \in Y) : P(X = x, Y = y) = P(X = x)P(Y = y)$
  - **Properties**:
    - $E(XY) = (E\,X)(E\,Y)$
    - $\mathrm{Var}(X + Y) = \mathrm{Var}\, X + \mathrm{Var}\, Y$

        - $\mathrm{Cov}(X, Y) = 0$ "The converse is not true: $Cov(X, Y) = 0$ doesn't necessarily imply that the random variables are independent".
        - $g_{X+Y}(t) = g_X(t) g_Y(t)$

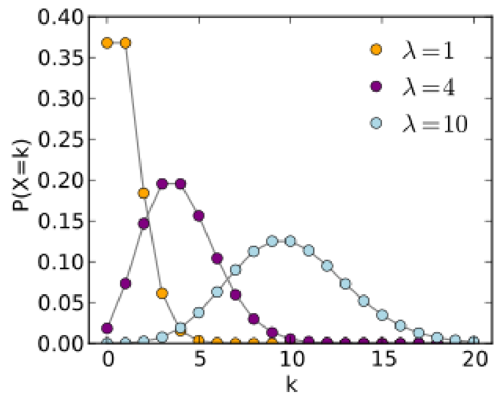
- **[Uniform distribution]** $X \sim U[a, b]$ - can be discrete or continuous
    - The simplest distribution in which all the events are equally likely to happen with probability p
    - **Examples:** tossing a fair coin ($p = 1/2$), rolling a standard fair die ($p = 1/6$).
    - For continuous case, the graph is a horizontal line, CDF calculations are easier.
        - CDF: $F_X(x) = P(X < x) = \int_a^x \frac{1}{b-a} I_{x \geq a} dt = \frac{x-a}{b-a} I_{x \geq a}$
        - PDF: $\mathrm{f}_X(\mathrm{x}) = \mathrm{P}(\mathrm{X} = \mathrm{x}) = \frac{1}{\mathrm{b}-\mathrm{a}} \mathrm{I}_{\mathrm{a} \leq \mathrm{x} \leq \mathrm{b}},$
        - $\mathrm{E\,X} = \frac{\mathrm{a}+\mathrm{b}}{2}, \mathrm{E\,X}^2 = \frac{\mathrm{a}^2+\mathrm{ab}+\mathrm{b}^2}{3}, \mathrm{Var\,X} = \frac{(\mathrm{b}-\mathrm{a})^2}{12}$



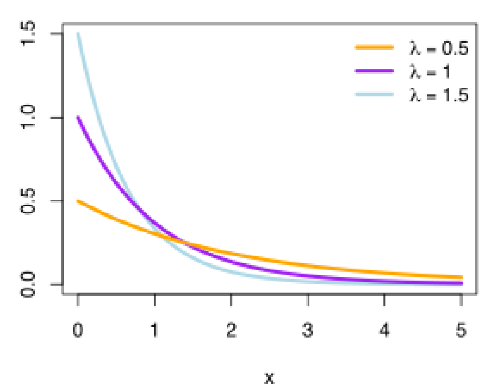
- **[Geometric distribution]** $X \sim G(p)$ - discrete
    - The geometric distribution function is the probability distribution of the number of Bernoulli trials needed to get the first success.
        - Recall that p is the probability of success, $q = 1 - p$.
    - A geometrically distributed random variable X can take values $\in \{1, 2, 3, \ldots\}$ with probabilities $P(\mu_X = 1)$,
    - PMF: $P(X = k) = G(k) = pq^{k-1}$ is the probability that we get the first success in the kth trial.
    - $\mathrm{E\,X} = \frac{1}{\mathrm{p}}, \mathrm{Var\,X} = \frac{\mathrm{q}}{\mathrm{p}^2}$
    - CDF: $\mathrm{F}_X(x) = P(X < x) = P(X \leq x - 1) = 1 - q^{x-1}$
    - G has the lack-of-memory property: the probability of an event happening is independent of previous results.
        - $\mathrm{P}(\mathrm{X} > \mathrm{a} + \mathrm{b} \mid \mathrm{X} > \mathrm{a}) = \mathrm{P}(\mathrm{X} > \mathrm{b}) = \mathrm{q}^{\mathrm{b}}$



- **[Poisson distribution]** $X \sim Po(\lambda), \lambda > 0$ - discrete
    - The Poisson distribution is popular for modeling the number of times an event occurs in an interval of time or space.
    - λ is the average number of times the event happens.
    - **PMF:** $\mathrm{P}(\mathrm{X} = \mathrm{k}) = \frac{\lambda^{\mathrm{k}}}{\mathrm{k}!} \mathrm{e}^{-\lambda}$, For $\mathrm{k} = 0, 1, 2, \ldots,$
    - **Properties:**
        - $\mathrm{E\,X} = \mathrm{Var\,X} = \lambda, \mathrm{E\,X}^2 = \lambda^2 + \lambda$
        - $Y \sim Po(\theta) \Rightarrow X + Y \sim Po(\lambda + \theta)$ "for independent X, Y"



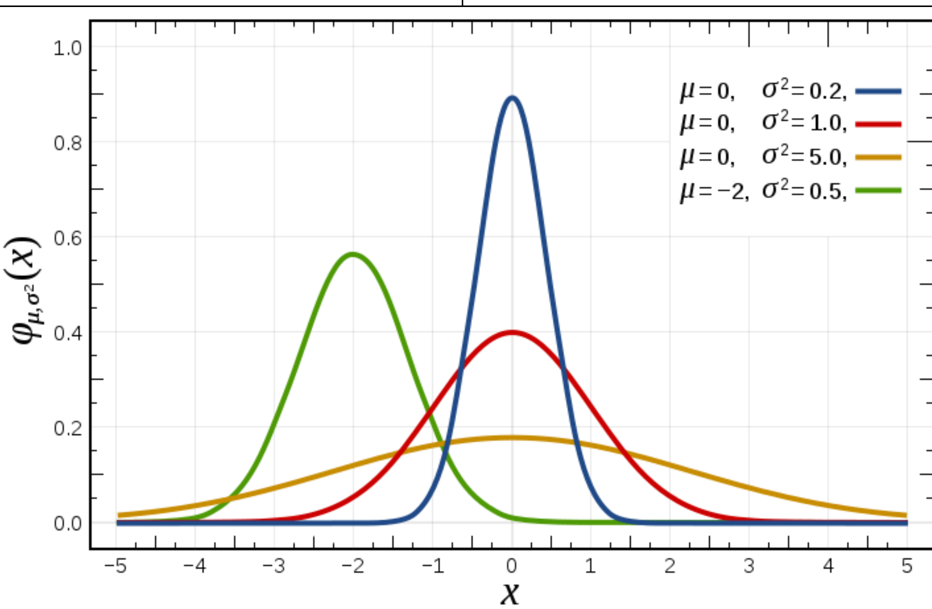
- **[Exponential distribution]** $X \sim Exp(\lambda), \lambda > 0$ – continuous
    - CDF: $F_X(x) = 1 - e^{-\lambda x} I_{x>0}$
    - PDF: $\mathrm{f}_X(\mathrm{x}) = \lambda\, \mathrm{e}^{-\lambda x} \mathrm{I}_{x>0}$
    - $E\,X = \frac{1}{\lambda}, E\,X^2 = \frac{2}{\lambda^2}, Var\,X = \frac{1}{\lambda^2}, E\,X^k = \frac{k!}{\lambda^k}$
    - Exponential distribution is the only continuous probability distribution to have the lack-of-memory property
        - $\mathrm{P}(\mathrm{X} > \mathrm{a} + \mathrm{b} \mid \mathrm{X} > \mathrm{a}) = \mathrm{P}(\mathrm{X} > \mathrm{b}) = e^{-\lambda b}$
    - **Note:** $X \sim U[a, b] \Longrightarrow -\ln X \sim Exp(1)$



Lecture 5

- **Normal distribution** $X \sim N(\mu, \sigma^2), \sigma > 0$ – continuous
  - Standard normal distribution: with $\mu = 0, \sigma = 1$
  - PDF: $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
    - A useful substitution while integrating this function is $t = \frac{x-\mu}{\sigma}, dx = \sigma dt$, this will transform the original integral to a standard one (with $\mu = 0, \sigma^2 = 1$), which can be calculated using the table of values of the function Φ or Φ_0
  - $Z = aX + bY + c \Longrightarrow Z \sim N(E\,Z, Var\,Z)$
  - $X \sim N(0,1) \Rightarrow E\,X^{2n-1} = 0,\; E\,X^{2n} = (2n-1)!! = (1)(3)(5)\dots(2n-1)$
- **Formulas table**

$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-0.5t^2} dt = 1$	Area under the curve is always = 1
$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-0.5t^2} = 0.5 + \Phi_0(x)$	Definition of Φ Relation between Φ and Φ_0
$\Phi_0(\mathrm{x}) = \int_0^x \frac{1}{\sqrt{2\pi}} e^{-0.5t^2} = 0.5\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)$	Definition of Φ_0 Relation between Φ_0 and erf
$\Phi(-\mathrm{x}) = 1 - \Phi(\mathrm{x}), \Phi_0(-x) = -\Phi_0(x)$	Negative arguments
$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < T < \frac{b-\mu}{\sigma}\right)$ $= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$	Probability using Φ
$E\,X = \mu, E\,X^2 = \sigma^2 + \mu^2, Var\,X = \sigma^2$	Expected value and Variance
$\Phi(x) = c \Longrightarrow x = \Phi^{-1}(c) = \sqrt{2}\,\mathrm{erf}^{-1}(2c-1)$	Inverse Φ function $\mathrm{erf}^{-1}(x) = inverf(x)$ in wolfram 😉



- **Joint probability distribution:**
  - Gives the probability that 2 or more random variables take specific values (discrete) or are situated in some domain (continuous).
  - The **discrete** case with 2 random variables can be visualized as a table:
  - Notice that the sum of all values (white cells) in the table should be 1

η \ ξ	1	2	3	Σ
-1	1/12	3/12	5/12	9/12
1	1/12	1/12	1/12	3/12
Σ	2/12	4/12	6/12	1

  - **Marginal distribution:** the distribution of one random variable, not considering the other ones.
    - **Example**: marginal distribution of $\xi \sim \begin{pmatrix} 1 & 2 & 3 \\ 2/12 & 4/12 & 6/12 \end{pmatrix}$
  - **Distribution of product**: includes all the values that the product can take with their respective probabilities.
    - **Example**: $\xi\eta \sim \begin{pmatrix} 1 & 2 & 3 & -1 & -2 & -3 \\ 1/12 & 1/12 & 1/12 & 1/12 & 3/12 & 5/12 \end{pmatrix}$
  - **Conditional expectation (expected value) for a joint probability distribution:** the expected value of a random variable given that the other variable(s) value(s) are known.
    - $E(\xi \mid \eta = \eta_0)$ is a number
      - $E(\xi \mid \eta = \eta_0) = \frac{1}{P(\eta=\eta_0)} \sum_{x_i} x_i P(\xi = x_i, \eta = \eta_0)\,, \xi \in \{x_i\}$
      - **Example:** $E(\xi \mid \eta = -1) = \frac{12}{9}\left(1 * \frac{1}{12} + 2 * \frac{3}{12} + 3 * \frac{5}{12}\right) = \frac{22}{9}$
    - $E(\xi \mid \eta)$ is a random variable (a function of x), whose expected value $= E\xi$
      - Here we are given that η is fixed, but we don't know its exact value.
      - Therefore, the expected value can take different values depending on the value of $\eta \in \{\eta_1, \eta_2, \dots, \eta_n\}$.
      - $E(\xi \mid \eta) \sim \begin{pmatrix} \frac{\sum_{x_i} x_i P(\xi = x_i, \eta = \eta_1)}{P(\eta = \eta_1)} & \dots & \frac{\sum_{x_i} x_i P(\xi = x_i, \eta = \eta_n)}{P(\eta = \eta_n)} \end{pmatrix}$
      - **Example:** $E(\xi \mid \eta) \sim \begin{pmatrix} \frac{1 * \frac{1}{12} + 2 * \frac{3}{12} + 3 * \frac{5}{12}}{9/12} & \frac{1 * \frac{1}{12} + 2 * \frac{1}{12} + 3 * \frac{1}{12}}{3/12} \end{pmatrix}$
  - **Continuous case**
    - Marginal distribution: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$
    - Expected value of a product: $E\,(XY) = \iint_{\boldsymbol{R}^2} xy f_{XY}(x,y)dy$
    - **Conditional expectation:**
      - $E(X \mid Y = y_0) = \int_{-\infty}^{\infty} x\, f_{X|Y=y_0}(x)dx\,,\; f_{X|Y=y_0}(x) = \frac{f_{X,Y}(x,y_0)}{f_Y(y_0)}$
      - $E(X \mid Y) = E(X \mid Y = y) = H(y)$
  - **Useful formulas:**
    - $E(E(X \mid Y)) = EX$ "expectation of a conditional expectation is unconditional expectation"
      - Useful for calculating $E(XY)$ used to find variance and covariance.
      - $E(XY) = E\big(E(XY|X)\big) = E(X\,E(Y|X))$
      - $E(XY) = E\big(E(XY|Y)\big) = E(Y\,E(X|Y))$
    - $E(Y \mid X) = \sum_j y_j P\big(y = y_j \mid X\big)$
    - $Var\,(Y \mid X) = \sum_j \big(y_j - EY\big)^2 P\big(y = y_j \mid X\big)$
    - $Var\,Y = E\big(Var(Y \mid X)\big) + Var\big(E(Y \mid X)\big)$ **"Law of total variance"**
  - Check Lecture 4: Covariance & Correlation coefficient & Independent random variables.

Lecture 6

- **Let**
  - $y_1, y_2, \ldots.$ be an infinite list of independent, identically distributed $(i.i.d)$ random variables.
  - N be a random variable independent from $y_1, \ldots$
  - $S = \sum_{i=1}^{N} y_i$
- **Then**
  - $E(S) = E(y_1)E(N), Var(S) = Var(y_1)E(N) + Var(N) * (Ey_1)^2$

- If X $\geq$ 0 is a random variable, then $EX = \sum_{k=1}^{\infty} P(X \geq k) = \sum_{m=1}^{\infty} mP(X = m)$
  - **Infinite geometric series sum formula**: $\sum_{k=0}^{\infty} ar^k = \frac{a}{1-r}$
- **Markov's inequality:** $P(|X| \geq \varepsilon) \leq \frac{E|X|^t}{\varepsilon^t}, \varepsilon > 0, t > 0$
- **Chebyshev's inequality:** $P(|X - EX| \geq \varepsilon) \leq \frac{Var\ X}{\varepsilon^2}, \varepsilon > 0$
  - Estimates the probability for a random variable to deviate from its mean.

Lecture 7

- **Let**
  - $\boldsymbol{X}$ be a vector of random variables: $X = (X_1, X_2, \dots, X_n)$
  - x be a vector representing one possible outcome $x = (x_1, x_2, \dots, x_n)$ (for each random variable taking a specific value).
- **Then**
  - CDF: $F_X(x) = P(X_1 < x_1, X_2 < x_2, \dots, X_n < x_n)$
  - PDF: $f_X(x) = \frac{\partial^n F_X(x)}{\partial x_1 \partial x_2 \dots \partial x_n}$
- **For 2 variables case in a rectangular domain:**
  - $P(a \le X_1 \le b, c \le X_2 \le d) = F_{X_1,X_2}(b,d) + F_{X_1,X_2}(a,c) - F_{X_1,X_2}(b,c) - F_{X_1,X_2}(a,d)$
  - $P(a \le X_1 \le b, c \le X_2 \le d) = \iint_{\substack{a \le X_1 \le b \\ c \le X_2 \le d}} f_{X_1,X_2}(x_1,x_2)dx_1dx_2$
- **For independent random variables**
  - $f_{X,Y}(x,y) = f_X(x)f_Y(y)$
    - **Note:** if we can factorize the joint PDF, then the RVs are independent, if we can't, it doesn't imply anything.
- **Let**
  - $f_{X,Y}(x,y)$ be the joint PDF of two random variables X, Y
  - $U = U(X,Y), V = V(X,Y)$ be two functions of the random variables, $(U,V) \in D, (X,Y) \in G$
- **Then**
  - $\iint_D f_{U,V}(u,v)dudv = \iint_G f_{X,Y}(x,y)dxdy$
  - PDF: $f_{U,V}(u,v) = f_{X,Y}\big(x(u,v), y(u,v)\big)|J|$
    - $J = \frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} x_u & x_v \\ y_u & y_v \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}$ is the determinant of the Jacobean matrix.
    - A useful property for J: $\frac{\partial(x,y)}{\partial(u,v)} = 1/\frac{\partial(u,v)}{\partial(x,y)}$
  - **The convolution formula:** $f_{U+V}(t) = \int_{-\infty}^{\infty} f_{U,V}(t-y,y)dy$
- **A common task**: find $f_{h(X,Y)}(t)$ given $f_{X,Y}(x,y)$
  - **Solution #1:**
    - Substitute $U = h(x,y), V = Y,$
    - Calculate $f_{U,V}(u,v) = f_{X,Y}\big(x(u,v), y(u,v)\big)|J|$
    - $f_U(u) = \int_{-\infty}^{\infty} f_{U,V}(u,v)dv$
  - **Solution #2:** $f_{h(X,Y)} = \left(F_{h(X,Y)}(t)\right)' = \big(P(h(X,Y) < t)\big)'$
- **Covariance matrix of a vector of random variables** $X = (X_1, X_2, \dots, X_n)$
  - $C = \begin{pmatrix} Cov(X_1,X_1) & \dots & Cov(X_1,X_n) \\ \dots & \dots & \dots \\ Cov(X_n,X_1) & \dots & Cov(X_n,X_n) \end{pmatrix}, Cov(X,X) = Var\ X$
  - $Cov(a_1X_1 + \cdots + a_nX_n, b_1X_1 + \cdots + b_nX_n) = (a_1 \dots a_n)C\begin{pmatrix} b_1 \\ \dots \\ b_n \end{pmatrix} = ACB$
    - If $A = B$ then C is not a negative definite matrix.
  - $Var(a_1X_1 + \cdots + a_nX_n) = Cov(a_1X_1 + \cdots + a_nX_n, a_1X_1 + \cdots + a_nX_n) = (a_1 \dots a_n)C\begin{pmatrix} a_1 \\ \dots \\ a_n \end{pmatrix}$

Lecture 8 (Assignment 9)

- **Normal distribution for an n-dimensional vector of normally distributed random variables**
  - $f_X(\boldsymbol{x}) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det \Sigma}} \exp(-0.5(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$
    - $\boldsymbol{X} = (X_1, X_2, \dots, X_n), X_i \sim N(\mu_i, \sigma_i), \boldsymbol{x} = (x_1, x_2, \dots, x_n)$
    - **The vector of expectations:** $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \cdots \\ \mu_n \end{pmatrix}, \mu_i = E\, X_i$
    - **The covariance matrix:** $\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \cdots & \cdots & \cdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}, \sigma_{ij} = Cov(X_i, X_j)$
      - Σ is a symmetric positive definite matrix: $\sigma_{ij} = \sigma_{ji}, \boldsymbol{x}^T \Sigma \boldsymbol{x} > 0, \boldsymbol{x} \neq \boldsymbol{0}$
      - $\det(\Sigma) = \prod_{i=1}^n \lambda_i^2$ the eigenvalues of Σ.
      - $\Sigma^{-1} = S\Lambda^{-1}S^{-1}, S = (\boldsymbol{x}_1 \quad \cdots \quad \boldsymbol{x}_n), \Lambda^{-1} = \begin{pmatrix} \lambda_1^{-1} & 0 & 0 \\ 0 & \cdots & 0 \\ 0 & 0 & \lambda_n^{-1} \end{pmatrix}$
        - $\boldsymbol{x}_i$ are the eigenvectors of Σ: solutions to $(\Sigma - \lambda I)x = 0$ for each eigenvalue λ.
    - X, Y are independent $\Leftrightarrow Cov(X,Y) = 0, \rho_{X,Y} = 0$
      - Only for normally distributed RVs, left implication is not always true.
    - $\begin{pmatrix} \xi \\ \eta \end{pmatrix} \sim N(\mu, \sigma^2) \Rightarrow (\xi \mid \eta = k) \sim N\left(\mu_\xi + \rho_{\xi,\eta}\left(\frac{\sigma_\xi}{\sigma_\eta}\right)(k - \mu_\eta),\ \sigma_\xi^2(1 - \rho_{\xi,\eta}^2)\right)$
- **Characteristic function of a random variable:**
  - A unique function that identifies the random variable, each random variable has a unique characteristic function.
  - The Fourier transform of the random variable PDF (if it's continuous).
  - **The general form:** $\phi_X(t) = E\left(e^{itX}\right), t \in \boldsymbol{R}$
    - X is discrete $\Rightarrow \phi_X(t) = \sum_k e^{itx_k} P(X = x_k)$
      - **Recall:** $X \sim \begin{pmatrix} x_1 & \cdots & x_n \\ P(X = x_1) & \cdots & P(X = x_n) \end{pmatrix}$
    - X is continuous $\Rightarrow \phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x)dx$
      - **Recall:** X can be uniformly, exponentially, normally, … distributed with some parameter(s) on some domain.
      - **Inverse formula:** $f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_X(t)dt$
  - **Properties**
    - $\phi_X(t)$ always exists, $\phi_X(0) = 1, |\phi_X(t)| \leq 1$
    - X_1, X_2 are independent $\Rightarrow \phi_{X_1+X_2}(t) = \phi_{X_1}(t)\phi_{X_2}(t)$
    - $\phi_{aX+b}(t) = e^{ibt}\phi_X(at), a, b = const$
    - $X \sim N(\mu, \sigma^2) \Rightarrow \phi_X(t) = \exp\left(i\mu t - \frac{\sigma^2 t^2}{2}\right)$

Lecture 9 (Assignment 8)

- **Random variable convergence:** a sequence of random variables can converge to a random variable
  - **Convergence in probability (P)**
    - $\lim_{n\to\infty} Y_n = Z$ in probability $if\ \forall\varepsilon > 0: \lim_{n\to\infty} P(|Y_n - Z| > \varepsilon) = 0$
  - **Convergence in distribution (weak convergence) (D)**
    - $\lim_{n\to\infty} Y_n = Z$ in distribution $if\ \lim_{n\to\infty} P(Y_n < x) = P(Z < x)$
  - **Almost sure convergence (A)**
    - $\lim_{n\to\infty} Y_n = Z$ almost surely $if\ P\left(\lim_{n\to\infty} Y_n = Z\right) = 1$
    - It means $\lim_{n\to\infty} Y_n(\omega) = Z(\omega)$
  - **Convergence in mean (M)**
    - $\lim_{n\to\infty} Y_n = Z$ in mean square $if\ \lim_{n\to\infty} E(Y_n - Z)^2 = 0$
- **Implications:** $(M) \Rightarrow (P), (A) \Rightarrow (P), (P) \Rightarrow (D)$
- **Let**
  - $X_1, X_2, \dots, X_n$ be **i**ndependent, **i**dentically **d**istributed **r**andom **v**ariables
  - X_i has $E\ X_i = \mu, Var\ X_i = \sigma^2$
  - $S_n = \sum_{i=1}^{n} X_i$
- **Then**
  - **Law of large numbers:** the more identical experiments we do, the closer the arithmetic mean of results gets to the expectation.
    - **Weak law:** For a large value of n, the "arithmetic mean" of the values of the random variables converges in probability to their common expectation
      - $\forall\varepsilon > 0: \lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) = 0$
    - **Strong law:** For a large value of n, the "arithmetic mean" of the values of the random variables converges almost surely to their common expectation
      - $P\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1$
  - **Central limit theorem**
    - $S_n \sim N(n\mu, n\sigma^2)$
      - It can also have other distributions depending on the distribution of X_i
      - This is what makes the normal distribution very common/popular; every variable that can be modelled as a sum of many small $iidrv$'s with finite mean and variance is approximately normal.
    - $\frac{S_n - ES_n}{\sqrt{Var\ S_n}} = \frac{S_n - n\mu}{\sqrt{n}\,\sigma} \sim N(0,1)$

Lecture 10 (Assignment 10)

- **Gamma distribution** $X \sim \text{Gam}(\alpha, \lambda)$ describes the time until *n* consecutive rare random events occur in a process with no memory.
  - $f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} I_{x\geq 0}, \alpha > 0, \lambda > 0, \Gamma(x) = \int_0^{+\infty} t^{\alpha-1} e^{-t}$
  - $E\,X = \frac{\alpha}{\lambda}, E\,X^2 = \frac{\alpha(\alpha+1)}{\lambda^2}, Var\,X = \frac{\alpha}{\lambda^2}, \phi_X(t) = \frac{\lambda^\alpha}{(\lambda - it)^\alpha}$
  - $X_1 \sim Gam(\alpha_1, \lambda), X_2 \sim Gam(\alpha_2, \lambda), X_1, X_2$ are independent $\Rightarrow X_1 + X_2 \sim Gam(\alpha_1 + \alpha_2, \lambda)$
- **Chi-Squared distribution** $X \sim \chi^2_n$ "with n-degrees of freedom"
  - Useful for inference regarding the sample variance of normally distributed samples
  - $\chi^2_n = X_1^2 + X_2^2 + \cdots + X_n^2, X_i \sim N(0,1), X_i$ are independent
  - $f_{\chi^2_n}(x) = \frac{1}{2^{0.5n}\,\Gamma(0.5n)} x^{0.5n\,-\,1} e^{-0.5x} I_{x>0} \sim Gam(0.5n, 0.5)$
    - Note: $\Gamma(0.5) = \sqrt{\pi}$
  - $E\,X = n, Var\,X = 2n$
- **Student's t-distribution** $X \sim t_n$ "with n-degrees of freedom"
  - Useful for inference regarding the mean of normally distributed samples with unknown variance
  - $t_n = \frac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}} = \frac{X}{\sqrt{\frac{Y}{n}}}, X \sim N(0,1), Y \sim \chi^2_n$
  - $f_{t_n}(x) = \frac{\Gamma\left(\frac{x+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad t_\infty = N(0,1)$
- **Mathematical Statistics**
  - Allows us to do estimations/analysis regarding the probability distributions of a data set.
  - Simple sample = $i.i.d.r.v$
    - $\boldsymbol{X}(X_1, X_2, \ldots, X_n)$
  - A **realization** of a sample: the actual values the sample variables took during an experiment.
    - $\boldsymbol{x}(x_1, x_2, \ldots, x_n)$, also called an **observation**.
    - Any function $f(\boldsymbol{x})$ on the realization is called a "**statistic**"
      - **Mean:** $f(\boldsymbol{x}) = \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$

In statistics, $E\,X$ is usually denoted as $\bar{X}$

- **Common types of problems in mathematical statistics**
  1. To find the distribution parameters **(parameter estimation)**
  2. To find the distribution, given a realization **(hypothesis testing)**
  3. Given two simple samples, determine whether they have the same distribution.
  4. Check whether all RVs from a sample have the same distribution.

An estimator for θ is usually be denoted as $\hat{\theta}(\boldsymbol{X})$ or $T_\theta(\boldsymbol{X})$

- **Estimator** is a rule for computing a value of a distribution parameter θ
  - An estimator is a statistic that depends on θ; a statistic is not necessarily an estimator.
  - **Bias of an estimato**r: $bias\ T_\theta(\boldsymbol{X}) = E\,T_\theta(\boldsymbol{X}) - \theta$
    - An estimator with $bias = 0$ is called unbiased.
  - **Consistency:** an estimator is called consistent if it approaches the exact value of θ as the size of the given sample approaches infinity.
    - $T_\theta(\boldsymbol{X})$ converges in probability to θ
      - $T_\theta(\boldsymbol{X})$ is unbiased
      - $\lim_{n\to\infty} Var\,T_\theta(\boldsymbol{X}) = 0$
  - **Mean squared error of an estimator**: $E\,(T_\theta(\boldsymbol{X}) - \theta)^2 = Var\,T_\theta(\boldsymbol{X}) + bias^2\,T_\theta(\boldsymbol{X})$

A good estimator is **unbiased** and **consistent** and has a small **mean square error**

Lecture 11 (Assignment 10)

- **Maximum likelihood method**
  - **Used to find the parameter of a known distribution given a realization of a sample.**
  - **Likelihood function:** $L(\boldsymbol{x}, \theta) = f_{\boldsymbol{X}}(\boldsymbol{x}) = f(x_1, \theta) f(x_2, \theta) \dots f(x_n, \theta) > 0$
    - For discrete case: $f(x_i, \theta) = P(X_i = x_i)$
    - For continuous case: $f(x_i, \theta) = f_{X_i}(x_i)$, the PDF of X_i
  - For a given realization, we choose the value of θ that maximizes L
    - **Recall:** derivative of a product of functions
      - $(f_1 f_2 \dots f_n)' = f_1' f_2 \dots f_n + f_1 f_2' \dots f_n + \dots + f_1 f_2 \dots f_n'$
    - Instead of maximizing L, we can maximize $\ln L$ "logarithm of likelihood"
      - It's easier as we can use the properties of logarithms.
- **Sufficient statistic:**
  - **If** knowing the value of a statistic for a sample is **sufficient** for us to estimate θ [without knowing the realization of the sample ($\boldsymbol{x}$)]
  - **Then** that statistic is called "sufficient" for parameter θ.
  - **Mathematically:**
    - $T_\theta(\boldsymbol{X})$ is sufficient for θ $\Leftrightarrow \forall D \subset \boldsymbol{R}^n : P\big(\boldsymbol{X} \in D \,\big|\, T_\theta(\boldsymbol{X})\big)$ doesn't depend on θ
      - $T(X)$ is sufficient for θ $\Leftrightarrow P(\boldsymbol{X} = \boldsymbol{x} \mid T_\theta(\boldsymbol{X}) = t)$ doesn't depend on θ.
    - $T_\theta(\boldsymbol{X})$ is sufficient for θ $\Leftrightarrow f_{\boldsymbol{X}}(\boldsymbol{x}, \theta) = g(T_\theta(\boldsymbol{x}), \theta) h(\boldsymbol{x})$ **"Factorization Criterion"**
      - The joint PDF can be factored into a product such that one factor, h, does not depend on θ and the other factor, which does depend on θ, depends on x only through $T_\theta(\boldsymbol{x})$
- **Rao-Blackwell theorem:**
  - The conditional expectation of an unbiased estimator given a sufficient statistic can never be a worse estimator.
  - **Let:**
    - T be a sufficient statistic for θ
    - T_1 be an unbiased estimator of θ
    - $T_2 = E(T_1 \mid T)$
  - **Then**
    - $E\,T_2 = E\,T_1 \Longrightarrow T_2$ is also unbiased
    - $Var\,T_2 \le Var\,T_1$
    - $T_1 = T_1(T) \Leftrightarrow Var\,T_2 = Var\,T_1$

Alternative notation
$T_1 = \theta^*$
$T_2 = \hat{\theta}^*$

Lecture 12 (Assignment 10)

- **Linear Regression:**
  - A statistical approach for approximating a data set of n pairs (x_i, y_i) using a line $y = ax + b$
  - Given the realization $Y_i = \beta x_i + \gamma + \varepsilon_i$ where
    - Y_i is the value we get when doing the experiment i with input x_i
    - β, γ are unknown constants
    - ε_i are $i.i.d.r.v$ resembling the measurement error for each Y_i
      - $\varepsilon_i \sim N(0, \sigma^2) \Longrightarrow E\ \varepsilon_i = 0, Var\ \varepsilon_i = Var\ Y_i = \sigma^2$
  - The goal is to find the values of a, b using methods such as the Least Square Approximation.
  - Notice that the problem of approximation is reasonable only if $n \geq 3$
    - For $n = 2$ there is only one line that can be constructed with 0 error
    - For $n = 1$ there are infinitely many, with no way to prefer one over another.
- **Least Square Approximation**
  - A method used to fit an approximation line $y = ax + b$ given the data (x_i, y_i)
  - Substituting $\gamma = \alpha - \beta\bar{x}$ in $Y_i = \beta x_i + \gamma + \varepsilon_i$ we get $Y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i$
  - The approximation line can be determined by minimizing the sum of squared errors given by
    - $S = \sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} \big(y_i - \alpha - \beta(x_i - \bar{x})\big)^2$
  - Minimizing the sum, we get the estimators for α and β
    - $\hat{\alpha} = \bar{Y}, \hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
    - $\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2}{n}\right), \hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$
    - These estimators are good since they are **unbiased** and **consistent** and **uncorrelated** and have the minimum mean square error among all linear estimators of y_i
      - **Linear estimator of** y_i is a linear combination $T = \sum_{i=1}^{n} a_i y_i$
      - Strictly speaking, the estimator for β is consistent only for a proper choice of x_i in which the data is nearly uniformly distributed over the interval $[x_{min}, x_{max}]$
    - The MLE of σ^2 is $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \sim \frac{\sigma^2}{n}\chi_n^2$ "but this one depends on σ^2 itself"
    - $S^2 = \frac{1}{n-1} S_{xx}$ is an unbiased estimator for σ^2 and is usually used in practice
  - Finally, we get the approximation line $y = \hat{\alpha} + \hat{\beta}(x - \bar{x})$
  - **Let**
    - $R = \sum_{i=1}^{n}\left(y_i - \left(\hat{\alpha} + \hat{\beta}(x - \bar{x})\right)\right)^2$ be the squared difference between the measured value of y and the value we get using the estimators we constructed.
  - **Then**
    - $\frac{R}{\sigma^2} \sim \chi_{n-2}^2$
    - $R, \hat{\alpha}, \hat{\beta}$ are independent
- **Fisher's Lemma**
  - **Let**
    - $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, independent
    - $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
    - $S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2$
  - **Then**
    - S_{xx} and $\bar{X}$ are independent, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, $\frac{S_{xx}}{\sigma^2} \sim \chi_{n-1}^2$

Lecture 13 (Assignment 10)

- **Hypothesis testing**
  - **Given** a realization $\boldsymbol{x}$ of a simple sample $\boldsymbol{X}$**,** we need to determine the distribution of the sample so we make several hypotheses and we say that
    - A hypothesis H_0 is called the null hypothesis and is the default one to be accepted
      - A non-null hypothesis is sometimes called an alternative hypothesis H_1
    - H_0 is accepted if $\boldsymbol{x} \notin C$ where $C \in \boldsymbol{R^n}$ is the **critical region**
      - If $\boldsymbol{x} \in C$ then H_0 is rejected.
    - In practice, we don't compare $\boldsymbol{x}$ against C, but rather use a statistic $T(\boldsymbol{x})$ and check whether $T(\boldsymbol{x})$ falls in a particular range (for the critical region) or not.
  - **Hypothesis testing errors**
    - **Type I:** rejecting a true H_0 (more fatal and should be avoided)
    - **Type II:** accepting a false H_0
  - **Significance level (α):**
    - When constructing the critical region for H, the probability of making type I error shouldn't exceed α
    - Significance level accounts for the percentage of type I error we allow, choice of such value depends on how much we are afraid of making type I mistake.
    - The area under the PDF curve in the critical region = α
  - **Power of a test:**
    - A measurement of the hypothesis testing method to be used.
    - Indicates the probability that the test rejects a false H_0
      - This is the same as the probability of avoiding type II errors.
- **Tests for constructing the critical region**
  - **Likelihood ratio test**
    - The critical region is the one that satisfies $\Lambda_{H_1,H_0} > \beta$ (solve for β)
      - $\Lambda_{H_1,H_0}(\boldsymbol{x}) = \frac{f_1(x)}{f_0(x)}, f_0 = f_{\boldsymbol{X}}(\boldsymbol{x})$ given $H_0, f_1 = f_{\boldsymbol{X}}(\boldsymbol{x})$ given H_1
    - This gives the most powerful test, but it's impractical for high dimensional $\boldsymbol{x}$
  - **Chi-Squared test** (for a large realization vector $\boldsymbol{x}$ of N components)
    - Divide the real line **R** into n intervals such that
      - p_i is the probability that we are situated in the i^{th} interval
      - $\nu_i =$ number of x_i in the i^{th} interval
    - The test states that if the statistic $\sum_{i=1}^{n} \frac{(\nu_i - Np_i)^2}{Np_i} \sim \chi^2_{n-1}$ has a value greater than the quantile of χ^2_{n-1} at α, then we are in the critical region (and H_0 is rejected)
- **Consider a normal simple sample X (recall Fisher's lemma)**
  - **A good estimator for μ of a normal simple sample X**
    - **Given σ:** $T(\boldsymbol{x}, \mu) = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$
    - **Not given σ:** $T(\boldsymbol{x}, \mu) = \frac{(\bar{X}-\mu)\sqrt{n(n-1)}}{\sqrt{S_{xx}}} \sim t_{n-1}$
  - $(\mathbf{1}-\boldsymbol{\alpha})$ **confidence interval for μ: is the probability that μ**
    - **Given σ:** $P\left(\bar{X} - \frac{\sigma}{\sqrt{n}}t^* < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}t^*\right) = 1-\alpha$
    - **Not given σ:** $P\left(\bar{X} - \frac{\sqrt{S_{xx}}}{\sqrt{n(n-1)}}t^* < \mu < \bar{X} + \frac{\sqrt{S_{xx}}}{\sqrt{n(n-1)}}t^*\right) = 1-\alpha$

> **Quantiles of t-distribution (table):**
> Gives $[-t^*, t^*]$ for which the area under the PDF of $t_{\text{df}} = (\alpha/2)$

> **Quantiles of χ^2 distribution (table):**
> Gives (h) for which the area under the PDF of $\chi^2_{\text{df}} = (\alpha)$