# APPLIED STATISTICAL METHODS

Mathematics Institute

Faculty of Mathematics and Computer Science

Wroclaw University – Fall 2015

LECTURE 1

PROBABILITY BASICS- A QUICK REVIEW

Wroclaw, 5 - 9 October, 2015

- **Random variable:** A function that assigns numerical values to outcomes of an experiment.
- **E.g.** Toss a coin, outcomes: H or T. Take r.v. X such that

$$X = \begin{cases} 1 & \text{if H came up,} \\ 0 & \text{if T came up.} \end{cases}$$

  Possible values are 0 and 1.
- **E.g.** Measure daily maximum temperature in Wroclaw for a year. Get 365 values (one per day). Assign X= daily max temp. Possible values in an interval (-20, +40)C.

# Types of random variables (rv's):

- Discrete- finite or countable number of possible outcomes, e.g. number of dots on a die, survey results (Male/Female, Married/Not Married), etc.

- Continuous - set of values contains an interval, e.g. temperatures, weight, height, etc.

# Probability distributions - descriptions of random variables

- Descriptions of random variables depend on the rv being continuous or discrete.
- Generally, for **discrete** random variables we list their values with the probabilities of their occurence, which is often called their "probability mass function".
- For a continuous random variable $X$ probability distribution is given by **probability density function (pdf)** $f$ such that

  1. $$P(a \leq X \leq b) = \int_a^b f(x)dx, \tag{1}$$
  2. $f(x) \geq 0$ for all real $x$, and
  3. $\int_{-\infty}^{\infty} f(x)dx = 1.$

# Cumulative Distribution Function

- For any random variable $X$, the **cumulative distribution function (cdf)** $F_X(x)$ of $X$ is given by

$$F_X(x) = P(X \leq x) = \begin{cases} \int_{-\infty}^{x} f(t)dt, \text{ for all real } x & \text{if X continuous,} \\ \sum_{t \leq x} P(X = t), \text{ for all real } x & \text{if X discrete.} \end{cases} \tag{2}$$

- **Properties of a cdf.** For any rv $X$, its cdf is
  1. nondecreasing,
  2. between 0 and 1 (including the endpoints), and
  3. describing $X$ uniquely.
- Note, that cdf describes a rv uniquely, but pdf does not.

Let $X$ be a random variable.

- Mean (expected value): $\mu_X$ or EX. Mean is the "center of gravity" for a distribution.

- For continuous X, we have $\mu_X = EX = \int_{-\infty}^{\infty} xf(x)dx$, where $f(x)$ is the pdf of X. Note, that EX does not always exist.

- For discrete X, we have $\mu_X = EX = \sum_{allx} xP(X = x)$.

# Variance of a distribution: VarX or $\sigma_X^2$.

- Definition of variance: $VarX = E(X - EX)^2$.
- For continuous X, we have

$$VarX = EX^2 - (EX)^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \left[ \int_{-\infty}^{\infty} (xf(x) dx \right]^2,$$

  where $f(x)$ is the pdf of X.
- For discrete X, we have

$$VarX = EX^2 - (EX)^2 = \sum_{allx} x^2 P(X = x) - \left[ \sum_{allx} xP(X = x) \right]^2.$$

- Note, that VarX or EX do not always exist.

# Percentiles

- We will only work with percentiles for continuous distributions in this course.
- Let $X$ be a continuous random variable with pdf $f$. A number $a$ is the $p^{th}$ percentile of $X$ if
$$P(X \leq a) = p.$$
- **Quantile function $Q$ of a rv X.** Quantile function of a rv $X$ is the inverse of its cdf $F$, if the inverse exists.
$$Q_X(x) = F_X^{-1}(x), \text{ if } F^{-1} \text{ exists.}$$
- We can also say
$$F_X(x) = y \text{ iff } Q_X(y) = x,$$
or analytically: $Q_X(x) = y = F^{-1}(x) \Longleftrightarrow F_X(y) = x$.

# Percentiles contd.

- What if cdf is not $1-1$, as for discrete rvs? We can use a more general definition of an inverse function as follows:

$$Q_X(y) = \min\{x : F_X(x) \geq y\}.$$

- The 50th percentile of a distribution is called median. Median divides the distribution into two halfs.
- Cumulative distribution function provides probabilities for sets of values of a random variable. The quantile function provides quantiles (or percentiles) of a random variable.
- The domain of a cdf is all real numbers, the range of a cdf is the interval $[0, 1]$. The domain of the quantile function is the interval $[0, 1]$, its range is all real numbers.

# Symmetry of a distribution.

- Let rv $X$ have pdf $f$. We say that $X$ is a "symmetric rv" or "has a symmetric distribution" if its pdf is symmetric around some value.
- If a rv is symmetric, and if its mean exists, then the random variable is symmetric around its mean.
- For symmetric rv's mean=median.
- If a distribution is not symmetric, it is called "skewed". A distribution is "skewed to the right" if its median is smaller than its mean. A distribution is skewed to the left, if its median is larger than its mean.

# "Tails" of a distribution

- The values of a rv far to the right (or far to the left) of its center are called "tails of a distribution".
- For a continuous random variable $X$ with pdf $f$, we say "the weight of the tail" is the area under the pdf in the far right (or left) of the center. Thus weight of a tail of a distribution is the probability of values from this distribution falling far to the right (or left) of the center.
- We call events "far in the tails" of a distribution "extreme" events. they can be extremely large or extremely small.
- "Light tailed" distributions are those with small probability of extreme events.
- "Heavy tailed" distributions are those with large probability of extreme events.
- Often in practical applications we are interested in the extreme events, not in the "usual" or "central" events.

## Discrete: Bernoulli

- **Bernoulli distribution with parameter $p$.** We say that a rv $X$ has a Bernoulli distribution with parameter $p$, denoted $X \sim Bern(p)$ if its probability mass function is

| x | 0 | 1 |
|---|---|---|
| P(X=x) | 1 - p | p |

which can be written as

$$P(X = x) = \begin{cases} p^x(1-p)^{1-x} & \text{if } x = 0, 1, \\ 0 & \text{otherwise.} \end{cases}$$

- Mean and variance of a Bernoulli rv. $EX = p$, and $Var(X) = p(1-p)$.
- Characteristic experiment resulting in a Bernoulli rv is a toss of a coin and assignment of, say, 1 to H and 0 to T. Such experiment is called "Bernoulli trial with probability of success equal to p". We say that Bernoulli rv is an indicator of success in a Bernoulli trial.

## Discrete: Binomial

- **Binomial distribution with parameters n and p.** We say that a rv $X$ has a Binomial distribution with parameter $p$, denoted $X \sim Bin(n, p)$ where $x = 0, 1, 2, \ldots, n$, if its probability mass function is

$$P(X = x) = \begin{cases} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} & \text{if } x = 0, 1, 2, \ldots, n, \\ 0 & \text{otherwise.} \end{cases}$$

- Characteristic experiment that yields a binomial rv is tossing a coin n times and counting the number of, say, H in the n tosses.
- We can think of $X$ as the number of H in n Bernoulli trials with probability of H is p. Possible values of $X$ are $x = 0, 1, 2, \ldots, n$.
- We say that a Binomial rv counts the number of successes in n independent and identical Bernoulli trials with probability of success equal to p.

## Discrete: Binomial, contd.

- Mean and variance of a binomial rv. If $X \sim Bin(n, p)$, then $EX = np$, and $Var(X) = np(1-p)$.
- **Connection to Bernoulli rv.** $X \sim Bin(n, p)$ is a sum of n iid (independent and identically distributed) Bernoulli random variables with probability of success p.

- **Geometric distribution with parameter p.** We say that a rv $X$ has a geometric distribution with parameter $p$, denoted $X \sim Geo(p)$, if its probability mass function is

$$P(X = x) = \begin{cases} p(1-p)^{x-1} & \text{if } x = 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

- Characteristic experiment that yields a geometric rv is tossing a coin UNTIL the first H comes up, and counting the number of tosses required. We can think of $X$ as the number of tosses until we get the first H (success) in iid tosses of a coin with probability of H equal to p. Possible values of $X$: $x = 1, 2, \ldots$.
- We say that a geometric rv counts the number of independent and identical Bernoulli(p) trials needed UNTIL the first success happens.
- Mean and variance of a geometric rv. If $X \sim Geo(p)$, then $EX = 1/p$, and $Var(X) = (1-p)/p^2$.

## Discrete: Poisson

- **Poisson distribution**. Discrete random variable X has a Poisson distribution with parameter $\lambda$ if
$$P(X=k)=\frac{e^{-\lambda}(\lambda^k)}{k!} \text{ for } k=0,1,2,\ldots.$$
- The mean and variance of the Poisson r.v. are the same: $EX=Var(X)=\lambda$.
- **Poisson Model.** Suppose events can occur in space or time in such a way that:
    1. The probability that two events occur in the same *small* area or time interval is zero.
    2. The events in disjoint areas or time intervals occur independently.
    3. The probability than an event occurs in a given area or time interval T depends only on the size of the area or length of the time interval, and not on their location.

## Discrete: Poisson model and process

- Suppose that events satisfying the Poisson model occur at the rate $\lambda$ per unit time. Let X(t) denote the number of events occurring in time interval of length t. Then
$$P(X(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}.$$
- X(t) is called *Poisson process* with rate $\lambda$.
- Further, the waiting time Y between consecutive events has an exponential distribution with parameter $\lambda$ (that is with mean $1/\lambda$), that is
$$P(Y > y) = e^{-\lambda y}, \quad y > 0.$$

## Continuous: Uniform

- **Uniform distribution on an interval (a, b).** We say that a rv $X$ has a uniform distribution on an interval $(a, b)$, or $[a, b]$, denoted as $X \sim U(a, b)$, if its pdf is constant over that interval:

$$f(x) = \begin{cases} 1/(b-a) & \text{if } x \in (a, b), \\ 0 & \text{otherwise.} \end{cases}$$

- Mean and variance of a uniform rv. If $X \sim U(a, b)$, then $EX = (a + b)/2$, and $Var(X) = (b - a)/12$.

## Continuous: Normal/Gaussian

- **Normal (Gaussian) distribution with parameters $\mu$ and $\sigma$.** Continuous random variable X has a normal distribution with mean $\mu$ and variance $\sigma^2$ if its pdf is of the form:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where $\mu$ and $\sigma^2$ are real valued constants. If X has pdf as above, we denote it: $X \sim N(\mu, \sigma^2)$.

- The normal pdf is bell shaped and centered around the mean $\mu$.
- There is a special Normal distribution with mean 0 and variance 1, called standard normal distribution, and denoted by $Z \sim N(0, 1)$. The standard normal pdf is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

## Continuous: Normal/Gaussian, contd.

- The values of the standard normal cdf are tabulated. To find probabilities related to general normal random variables, use the following fact:
- **Theorem**. If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.
- **Properties of a normal distribution** Let $X \sim N(\mu, \sigma^2)$, then
  1. its pdf is symmetric around $\mu$,
  2. change in $\mu$ causes horizontal shift of the pdf;
  3. change in $\sigma$ causes change in shape of the cdf: the larger the $\sigma$ the "flatter" the pdf with heavier tails.

## Continuous: Lognormal

- **Lognormal distribution with parameters $\mu$ and $\sigma$.** Continuous random variable X has a lognormal distribution with parameters $\mu$ and $\sigma$, denoted $X \sim LN(\mu, \sigma)$, if its pdf is of the form:

$$f(x) = \frac{1}{\sqrt{2\pi}x\sigma} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, x > 0,$$

where $\mu$ and $\sigma^2$ are real valued constants.

- Mean and variance of a lognormal rv. If $X \sim LN(\mu, \sigma)$, then $EX = e^{\mu+\sigma^2/2}$ and $Var(X) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$.
- **Theorem**. If $Y \sim N(\mu, \sigma^2)$, then $e^Y = X \sim LN(\mu, \sigma)$. If $X \sim LN(\mu, \sigma)$, then $Y = \ln X \sim N(\mu, \sigma^2)$.

## Continuous: Exponential

- **Exponential distribution with parameter $\beta > 0$.** Continuous random variable X has an exponential distribution with parameter $\beta$, $X \sim exp(\beta)$, if its pdf is of the form:

$$f(x) = \beta e^{-\beta x}, x \geq 0,$$

where $\beta > 0$ is a real valued constant.

- Mean and variance of an exponential rv. If $X \sim exp(\beta)$, then $EX = 1/\beta$ and $Var(X) = 1/\beta^2$.

- **The Gamma distribution.**

  **The Gamma function.** For any positive real number $r > 0$, the *gamma function* of $r$ is denoted $\Gamma(r)$ and equal to

$$\Gamma(r) = \int_0^{\infty} y^{r-1} e^{-y} dy.$$

- **Theorem. Properties of Gamma function**. The Gamma(r) function satisfies the following properties:

  1. $\Gamma(1) = 1$.
  2. $\Gamma(r) = (r-1)\Gamma(r-1)$.
  3. For r integer, we have $\Gamma(r) = (r-1)!$

## Continuous: Gamma, contd.

- **Definition of the $\Gamma(r, \lambda)$ random variable.** For any real positive numbers $r > 0$ and $\lambda > 0$, a random variable with pdf

$$f_X(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, \quad x > 0,$$

  is said to have a Gamma distr. with parameters r and $\lambda$, denoted $X \sim \Gamma(r, \lambda)$.

- **Mean and var.** If $X \sim \Gamma(r, \lambda)$ then EX= $r/\lambda$, and Var(X)= $r/\lambda^2$.

- **Theorem**. Let $X_1, X_2, \ldots, X_n$ be iid exponential r.v.'s with parameter $\lambda$, that is with mean $1/\lambda$. The the sum of $X_i$'s has a gamma distribution with parameters $n$ and $\lambda$. More precisely, $\sum_{i=1}^n X_i \sim \Gamma(r, \lambda)$.

- **Theorem**. A sum of independent gamma r.v.'s $X \sim \Gamma(r, \lambda)$ and $Y \sim \Gamma(s, \lambda)$ with the same $\lambda$ has a gamma distr. with $r' = r + s$ and the same $\lambda$. That is $X + Y \sim \Gamma(r + s, \lambda)$.

# Sample Mean

- **Sample mean.** Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample from a distribution with mean $\mu$ and standard deviation $\sigma$. The sample mean is rv $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- The mean and variance of the sample mean are: $E\bar{X} = \mu$ and $Var\bar{X} = \sigma^2/n$.
- **Convergence in distribution.** Suppose that $(X_1, X_2, \ldots)$ and $X$ are real-valued random variables with distribution functions $(F_1, F_2, \ldots)$ and $F$, respectively. We say that the distribution of $X_n$ converges to the distribution of $X$ as $n \to \infty$ if

$$lim_{n\to\infty} F_n(x) = F(x),$$

for all x at which F is continuous.

# Distribution of $\bar{X}$ and the Central Limit Theorem (CLT).

- **Sample from a Normal Distribution**. Let $X_i \sim N(\mu, \sigma^2)$ iid for $i = 1, \ldots, n$. Then $\bar{X} \sim N(\mu, \sigma^2/n)$.
- **The Central Limit Theorem**. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean $\mu$ and variance $\sigma^2$. Then,

  $$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z, \text{ as } n \to \infty,$$

  or equivalently

  $$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z, n \to \infty,$$

  where *indistributionZ* $\sim N(0, 1)$.
- The CLT provides an approximation of the distribution of a sample mean and of a sum of iid random variables with finite variance.