Appendix 1: Population and Sampling Distributions (HAM pp 289-293)

- In **theoretical statistics** we make statements about a population based on is distribution $f(X)$ of a ***continuous*** random variable X or $\Pr(X = x)$ for a ***discrete*** random variable X, which takes the specific value x, respectively.
- In **applied statistics** we are dealing with ***sampled data*** from the population and aim at estimating properties of the underlying population from which the random sample has been drawn
- The sample is our narrow keyhole, which allows us to look at parts of the unknown population.
- **Conventions:**
  - Parameters characterizing the population are usually denoted by Greek characters, e.g., the expectation μ_X of the random variable *X*. Their estimates are either expressed by Latin characters, e.g., the mean $\bar{X}$, or by a hat symbol that denotes an estimate, e.g., $\hat{\mu}_X$.
  - A random variable from the population is usually denoted by a capital letter, e.g., X, whereas its observed realization in the sample is denoted by small letters, e.g., $x_1, x_2, \ldots, x_n$
- **Distribution function**
  - The **density function** of a ***discrete*** random variable is $\Pr(X = x)$ with $0 \leq \Pr(X = x) \leq 1$ and $\sum_{\text{all } x \text{ in support of } X} \Pr(X = x) = 1$.

- The **cumulative distribution** functions for all $x_i \leq x_{upper}$ is $\sum_{x_i \leq x_{upper}} \Pr(X = x_i)$.
- The density function for a continuous random variable is $f(x)$ with $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x) \cdot dx = 1$.
- Remember: the density function $f(x)$ at x cannot be interpreted as probability. We only can express the probability for a range of value:
  $\boxed{X \in [a, b] \Rightarrow \Pr(a \leq X \leq b) = \int_a^b f(x) \cdot dx.}$
  For infeasible values of x the density will become $f(x) = 0$ because these values are not probable.
- The distribution function is $F(x_{upper}) = \int_{-\infty}^{x_{upper}} f(x) \cdot dx$.

- **Population expectation (central tendency)**
  - The mean of a population is called ***expectation*** and denoted by $E[X] = \mu_X$
  - For ***discrete*** variables the expectation is calculated by
    $E[X] = \sum_{i=1}^{I} x_i \cdot \Pr(X = x_i)$
    where I is the total number of representation, which can be an infinite number as for the Poisson distribution $x_i \in \{0, 1, 2, \ldots, \infty\}$ or a finite set as in the sum of two throws of a dice $x_i \in \{2, 3, \ldots, 12\}$
  - For ***continuous*** random variables the expectation is defined in terms of the density function $f(x)$
    $E[X] = \int_{-\infty}^{\infty} x \cdot f(x) \cdot dx$.

- Some rules for the expectation:
  $E[a] = a$ for a ***deterministic*** constant a
  $E[a \cdot X] = a \cdot E[X]$
  $E[X \pm Y] = E[X] \pm E[Y]$
  $E[a + b \cdot X] = a + b \cdot E[X]$
  $$\begin{aligned} E[a \cdot X + b \cdot Y] &= E[a \cdot X] + E[b \cdot Y] \\ &= a \cdot E[X] + b \cdot E[Y] \end{aligned}$$
- An unbiased sample estimator of the expectation $E[X]$ is the mean
  $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$

- **Variance**
  - The variance is a measure of squared spread around the center of a random variable
    $$\begin{aligned} Var[X] &= \int \left(x - E[X]\right)^2 \cdot f(x) \cdot dx \\ &= E\left[\left(X - E[X]\right)^2\right] \\ &= E\left[X^2\right] - \left(E[X]\right)^2 \end{aligned}$$
  - The unbiased sample variance estimator s_X^2 is of the population variance σ^2 is:
    $s_X^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \bar{X}\right)^2$
  - Basic properties:
    $Var[a] = 0$ because a is a constant (i.e., not random)
    $Var[b \cdot X] = b^2 \cdot Var[X]$

$$Var[X+Y] = Var[X] + Var[Y] + 2 \cdot Cov[X,Y]$$
$$Var[X-Y] = Var[X] + Var[Y] - 2 \cdot Cov[X,Y]$$
$$\begin{aligned} Var[a + b \cdot X] &= Var[a] + Var[b \cdot X] \\ &= b^2 \cdot Var[X] \end{aligned}$$
$$Var[a \cdot X + b \cdot Y] = a^2 \cdot Var[X] + b^2 \cdot Var[Y] + 2 \cdot a \cdot b \cdot Cov[X,Y]$$

Example: Explanation of Integration using The Exponential Distribution

- Background information on the exponential distribution:
  - Example: the ***waiting times*** x between two independent random events (earth quakes, customers lining up in-front of a cashier, spacing between random points) may be exponential distributed.
  - The exponential distribution only has the parameter λ, with $E[X] = 1/\lambda$ being the average waiting time.
  - The exponential distribution is related to the ***Poisson*** distribution:
    - The Poisson distribution provides a stochastic model for the number of independent random events x within a fixed time-interval.
    - The expected number of random events within a fixed time-interval is $E[X] = \lambda$.
    - Inverse proportionality: If the expected number of events is large the average waiting time between the events will be small.
  - The density function of the exponential distribution is

$$\boxed{f(x \mid \lambda) = \begin{cases} \lambda \cdot \exp(-\lambda \cdot x) & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases}.}$$

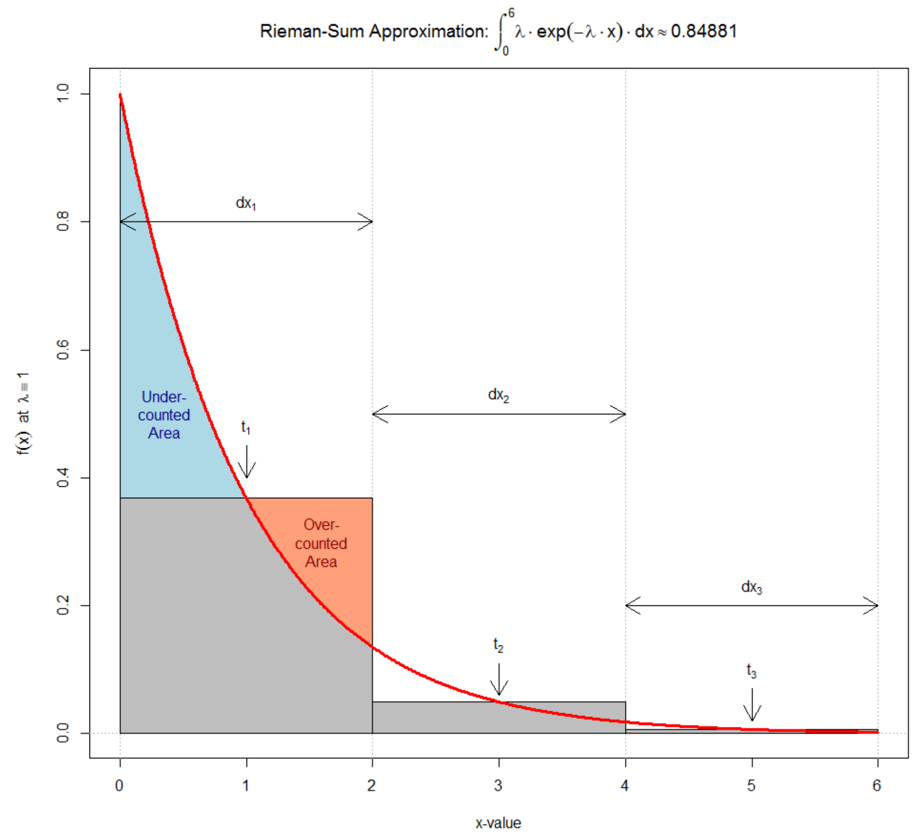
- Its distribution function (cumulative density) is

$$\boxed{F(x\,|\,\lambda)=\int_0^x \lambda\cdot\exp(-\lambda\cdot x)\cdot dx=\begin{cases}1-\exp(-\lambda\cdot x) & \text{for } x\geq 0\\ 0 & \text{otherwise}\end{cases}}$$

- Its moments are known analytical:
  - The expectation is $E[X]=\int_0^\infty x\cdot\lambda\cdot\exp(-\lambda\cdot x)\cdot dx=\frac{1}{\lambda}$ and
  - The variance is $Var[X]=\int_0^\infty \left(x-\underbrace{1/\lambda}_{=E[X]}\right)^2\cdot\lambda\cdot\exp(-\lambda\cdot x)\cdot dx=\frac{1}{\lambda^2}$
- The parameter λ can be estimated from sample observations by $\hat{\lambda}=1/\bar{x}$.

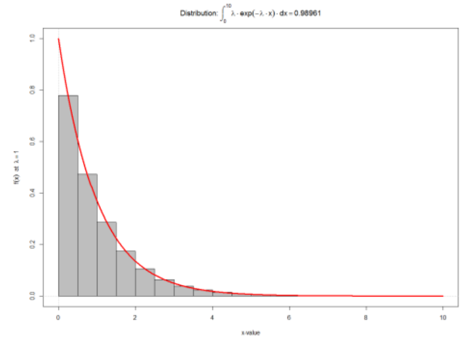
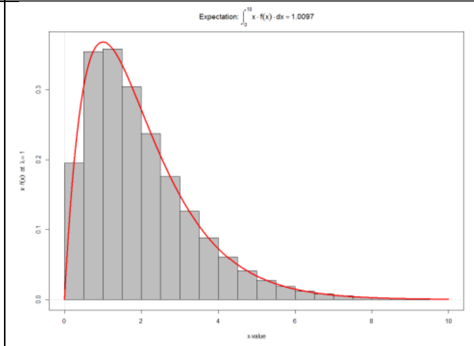
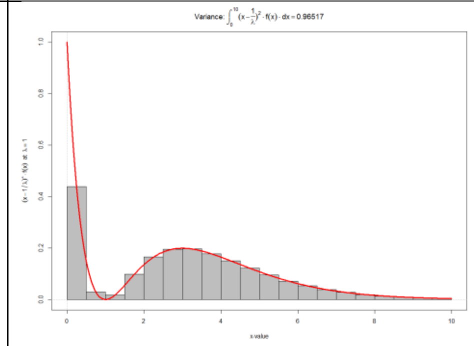
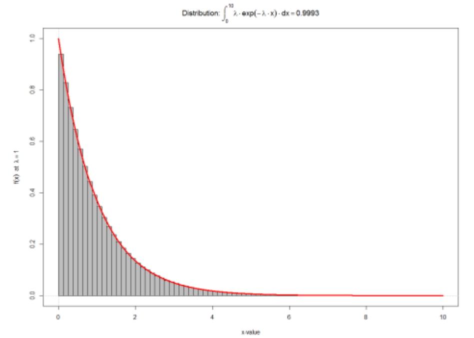
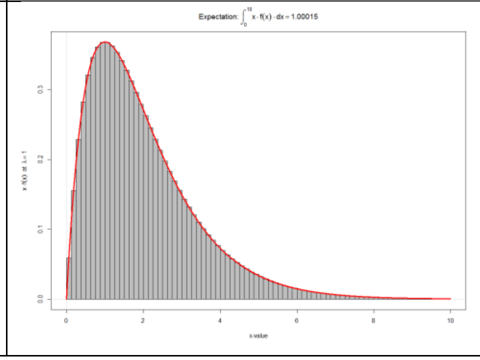
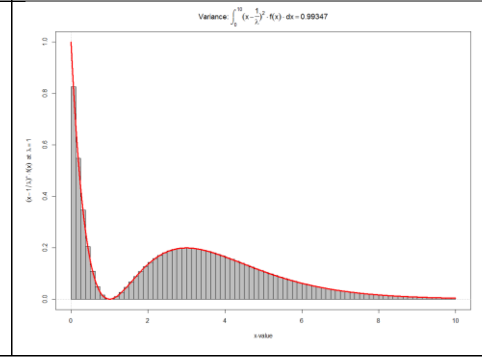
- Evaluation of the moments by numerical integration (see script **RIEMANNSUM.R**):
  - The Riemann sum approximates a continuous integral by $\int_a^b f(x)\cdot dx\approx\sum_{i=1}^n f(t_i)\cdot dx_i$ by discrete evaluations with $a=x_0<x_1<x_2<\cdots<x_{n-1}<x_n=b$, with the bin-width $dx_i=x_i-x_{i-1}$ and $t_i\in[x_{i-1},x_i]$, which usually is set to the halfway point $t_i=\frac{x_{i-1}+x_i}{2}$

- The parameters dx_i and n determine the resolution and therefore the accuracy of the Riemann sum integral approximation.
- The differences $dx_i = x_i - x_{i-1}$ can be adaptive in length to capture the variability of $f(x)$:
  - If the underlying function $f(x)$ varies heavily, then the differences dx_i should be small.
  - On the other hand, if the underlying function is fairly smooth the differences dx_i can be larger.



The underlying idea is similar to an adaptive kernel density estimator.

- Evaluation of the exponential expectation and variance for $\lambda = 1$ in the range $x \in [0,10]$:

Number of bins n	$f(x)$	$x \cdot f(x)$	$(x - E[X])^2 \cdot f(x)$
$n = 20$			
$n = 80$			

- Notes:
  - The integral $\int_{-\infty}^{\infty} f(x) \cdot dx = 1$ over any ***density functions*** $f(x)$ always is one.
  - ***Theoretically*** all integrals in the above example for the exponential distribution (first column) with $\lambda = 1$ should be one, because $\int_{-\infty}^{\infty} f(x) \cdot dx = 1$, $E(X) = 1/\lambda$ and $Var(X) = 1/\lambda^2$ for an exponential distribution with $\lambda = 1$.

    - Even if we increase the number of bins, these integrals will ***not*** approach 1 because we are ***truncating*** the infinitive upper integration bound at the value $b = 10 < \infty$.

- **Covariance**
  - The covariance is a basic measure of the **linear relationship** between pairs of random variables.
  - The covariance is the numerator of the correlation coefficient. That is $\rho = \frac{Cov[X,Y]}{\sqrt{Var[X] \cdot Var[Y]}}$
  - The covariance of a variable with itself is called the variance: $Cov[X,X] = Var[X]$

$$Cov[X,Y] = \iint (x - E[X]) \cdot (y - E[Y]) \cdot f(x,y) \cdot dx \cdot dy$$

  - $$= E\left[(X - E[X]) \cdot (Y - E[Y])\right]$$

$$= E[X \cdot Y] - E[X] \cdot E[Y]$$

  - The function $f(x,y) \geq 0$ is the joint-density of the random variables X and Y with $\iint f(x,y) \cdot dx \cdot dy = 1$
  - An unbiased estimator for the population covariance is $s_{XY} = \frac{\sum_{i=1}^{n}\left[(x_i - \bar{X}) \cdot (y_i - \bar{Y})\right]}{n-1}$
  - Some rules:
    $Cov[a,Y] = 0$
    $Cov[b \cdot X, Y] = b \cdot Cov[X,Y]$
    $Cov[X + W, Y] = Cov[X,Y] + Cov[W,Y]$

- The covariance is unaffected by the addition of a constant to either random variable:
$$\begin{aligned} Cov[a+X,Y] &= \underbrace{Cov[a,Y]}_{=0} + Cov[X,Y] \\ &= Cov[X,Y] \end{aligned}$$
- The covariance between sums of variables reduces to sums of covariances between their components
$$\begin{aligned} Cov[X+W,Y+Z] &= Cov[X+W,Y] + Cov[X+W,Z] \\ &= Cov[X,Y] + Cov[W,Y] + Cov[X,Z] + Cov[W,Z] \end{aligned}$$
$$\begin{aligned} Cov[X,Y-X] &= Cov[X,Y] - Cov[X,X] \\ &= Cov[X,Y] - Var[X] \end{aligned}$$

- The **O**rdinary **L**east **S**quares (OLS) slope estimator in terms of covariances becomes
  - The slope regression coefficient for a regression of Y onto X becomes
$$\boxed{\beta_{1,Y|X} = \frac{Cov[X,Y]}{Var[X]}}$$
  - Vice versa, for a regression of X onto Y one gets $\beta_{1,X|Y} = \dfrac{Cov[X,Y]}{Var[Y]}$
  - The regression intercept for a regression of Y onto X becomes $\boxed{\beta_{0,Y|X} = E[Y] - \beta_{1,Y|X} \cdot E[X]}$, because the expectations $E[Y]$ and $E[X]$ lie on the regression line.

Normal Distribution and its Relatives

- Definition: Let z and the sets $\{z_1, z_2, \dots, z_n\}$ with n elements and $\{\tilde{z}_1, \tilde{z}_2, \dots, \tilde{z}_m\}$ with m elements be ***standard normal*** distributed random variables, which are all ***mutually independent***.
- The χ^2-distribution: The random variables

  $$\boxed{s_n^2 = \sum_{i=1}^{n} z_i^2 \text{ and } \tilde{s}_m^2 = \sum_{i=1}^{m} \tilde{z}_i^2}$$

  of the ***sums of squared*** independent standard normal distributed variables are χ^2-distributed

  $$s_n^2 \sim \chi^2_{df=n} \text{ and } \tilde{s}_m^2 \sim \chi^2_{df=m}$$

  with n and m degrees of freedom, respectively.

  The expected value of a χ^2-distributed variable is equal to its degrees of freedom, that is, $E[s_n^2] = n$.
- The t-distribution: Let $t_n = \frac{z}{\sqrt{s_n^2/n}}$ and $\tilde{t}_m = \frac{z}{\sqrt{\tilde{s}_m^2/m}}$. Then t_n and $\tilde{t}_m$ are t-distributed with with n and m degrees of freedom, respectively.
- The F-distribution: Let $F_m^n = \frac{s_n^2/n}{\tilde{s}_m^2/m}$. Then F_m^n is F-distributed with n and m degrees of freedom.

Bias and Mean Square Error (Skipped)

- The sampling distribution of a statistic $\hat{\theta}$ is its theoretical distribution over all possible random samples of a given size n.
- A statistic is ***unbiased*** if $E[\hat{\theta}] = \theta$, that is, $E[\hat{\theta}] - \theta = 0$. It is biased if $E[\hat{\theta}] \neq \theta$, that is, it's expected value differs from the true population parameter θ.
- The variance $Var[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$ expresses the precision of a sample statistics. The square root of this variance is called ***standard error*** of a sample statistics.
- The mean square error is

$$\begin{aligned} MSE &= E\left[\left(\hat{\theta} - \theta\right)^2\right] \\ &= E\left[\left(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta\right)^2\right] \\ &= E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2 + 2 \cdot \left(\hat{\theta} - E[\hat{\theta}]\right) \cdot \left(E[\hat{\theta}] - \theta\right) + \left(E[\hat{\theta}] - \theta\right)^2\right] \\ &= Var[\hat{\theta}] + bias^2 \end{aligned}$$

with the term $E[2 \cdot (\hat{\theta} - E[\hat{\theta}]) \cdot (E[\hat{\theta}] - \theta)] = 0$ because

$$\begin{aligned} E[(\hat{\theta} - E[\hat{\theta}]) \cdot (E[\hat{\theta}] - \theta)] &= E[\hat{\theta} \cdot E[\hat{\theta}] - \hat{\theta} \cdot \theta - E[\hat{\theta}] \cdot E[\hat{\theta}] + E[\hat{\theta}] \cdot \theta] \\ &= E[\hat{\theta}] \cdot E[\hat{\theta}] - E[\hat{\theta}] \cdot \theta - E[\hat{\theta}] \cdot E[\hat{\theta}] + E[\hat{\theta}] \cdot \theta \\ &= 0 \end{aligned}$$

Only $\hat{\theta}$ is a random variable, whereas $E[\hat{\theta}]$ and θ are constants and therefore, $E\left[E[\hat{\theta}]\right] = E[\hat{\theta}]$ and $E[\theta] = \theta$.