## Chapter Three

**Defn**: A random variable whose space R consists of an interval or union of intervals (and not a countable set of points) is said to be a continuous random variable.

**Remark**: If data falls over a continous range or if the size of the data set is very large, the user can group the data values into some equal-sized categories.

- Determine the range : $r = \max(x_k) - \min(x_k)$.
- Select a number of classes m where generally $5 < m < 20$, where each class is approximately of width $r / m$.
- Force each interval to begin and end halfway between two adjacent possible values.
- Terms: Class Intervals, Class Boundaries, Class Limits, Class Marks
- Display using a relative frequency histogram (density histogram). Note, each bar relates area to relative frequency...not height.

**Stem-and-leaf display**: Grouping the data into categories based on properties of the actual values while retaining the original data's values.

**Quantiles**: Measures which divide ordered data into roughly equally-sized portions.

* **Median** - a value which splits the data into 2 equal parts
* **Quartiles** - three values which split the data into 4 equal parts
* **Deciles** - nine values which split the data into 10 equal parts
* **Percentiles** - ninety-nine values which split the data into 100 equal parts

**Computing Percentiles**: For a set of n values $\{x_r\}$, sort these to yield the values $\{y_r\}$. Then, the (100p)th percentile $\pi_p$ for $0<p<1$ is the value for which approximately (100p)% of the data lies less than $\pi_p$ and 100(1-p)% lies above $\pi_p$.

To determine the value of the (100p)th percentile, consider:

- r = integer part of (n+1)p
- s = the fractional part of (n+1)p
- locate the terms yr and yr+1
- Compute percentile location using a weighted average:

$$\pi_p = y_r + s(y_{r+1} - y_r) = (1-s)y_r + s\, y_{r+1}.$$

Several named values correspond to certain percentile values:

* $\pi_{0.10}$ = first decile, p0.20 = second decile, etc
* $\pi_{0.25}$ = Q1 = first quartile, p 0.75 = Q3 = third quartile
* $\pi_{0.50}$ = median or second quartile or 5th decile

**HOMEWORK**: page 136

**Defn**: The function f(x) is a probability density function (pdf) for a continuous random variable X over the space R if f(x) satisfies:

- f(x) > 0 for $x \in R$
- f(x) = 0 for $x \notin R$
- $\int_R f(x)dx = 1$
- P(A) = $\int_A f(x)dx$ f(x) > 0, for $A \subseteq R$.

**Remark**: The pdf for a continuous random variable may be unbounded. Indeed, consider f(x) = $0.5/\sqrt{x}$, on R=(0,1). This satisfies the definition of pdf but as x approaches zero, f(x) becomes unbounded.

**Defn**: If X is a random variable, define the *distribution function* F(x) = P( X $\leq$ x ) = $\int_{-\infty}^{x} f(t)dt$ .

**Remark**: If X is a random variable and F the corresponding distribution function, then

$$P(a < X \leq b) = F(b) - F(a).$$

Hence, we can determine probabilities over arbitrary intervals if we know the distribution function explicitly.

**Result**: F(x) is always continuous, even if f(x) is not.

**Result**: By using the fundamental theorem of calculus, F'(x) = f(x). So, f(x)>0 implies F(x) is strictly increasing over the space R.

**Note**: P( X=a ) = P( X$\leq$a ) - P( X<a ) = F(a) - $\lim_{\varepsilon \to 0}$ F(a-ε) = 0. Hence, for a continuous random variable, the probability of any particular value occurring is exactly zero. Therefore, the only nonzero probabilities occur for intervals of values. In that case, we obtain:

$$P( a \leq x \leq b ) = P( X \leq b ) - P( X < a ) = P( X \leq b ) - P( X \leq a ) = F(b) - F(a).$$

This indicates that, for continuous random variables, the distribution function will play the pivotal role when computing probabilities.

**Remark**: Let B = { u: u = X(s) with $s \in S$ and u$\leq$x }. We call

$$F_n(x) = |B| / |S|$$

the *empirical distribution function*. This allows approximating a distribution function by using a sample.

**Defn**: For X a continuous random variable and f(x) its pdf:

- Expected Value of u(x) = $E[u(x)] = \int_R u(x)f(x)dx$ .

- Mean $\mu = E[x] = \int_R x\, f(x)dx$.
- Variance $\sigma^2 = E[(x-\mu)^2] = \int_R (x-\mu)^2 f(x)dx = \int_R x^2 f(x)dx - \mu^2$.

**Defn**: For a continuous random variable X with distribution function F(x), the 100pth percentile is the number a such that

$$p = \int_{-\infty}^{a} f(x)dx$$

As with discrete data, we define the median with p=0.5 and the quartiles with p=0.25 and p=0.75.

**HOMEWORK**: page 151

**Uniform Distribution**: Let X be the outcome upon selecting any point randomly from an interval or collection of intervals R. If sub-intervals of equal widths are equally likely to be chosen, then we say we have a uniform distribution. Notice, the hypothesis states that intervals in R of equal width must have equal probabilities implies the distribution function must be linear over the space. Since F(x) is linear, then F'(x) = f(x) = c and so a uniform continuous distribution has a variable whose pdf is constant over the space R.

**Special Case of Uniform Distribution**: We will almost always consider the space of X to be R=[a, b]. If so, the for any $x \in [a, b]$, we have

$$f(x) = 1/(b-a)$$
$$F(x) = (x-a)/(b-a)$$
$$M = (a+b)/2$$
$$\sigma^2 = (b-a)^2/12.$$

Such a distribution will be denoted U(a,b).

**Exponential Distribution**: Consider a Poisson process on with $\mu = \lambda T$. Let W = continuous variable measuring the waiting time till the first change. Then,

$$F(w) = P(W \leq w) = 1 - P(W > w) = 1 - P(\text{no changes in the interval } [0,w]) = 1 - e^{-\lambda w},$$

since this last probability is in a discrete Poisson problem.
Since F'(x) = f(x), then we have

$$f(x) = \lambda e^{-\lambda w}.$$
$$\mu = 1/\lambda$$
$$\sigma^2 = 1/\lambda^2 = \mu^2.$$

Often, textbooks write $\theta = 1/\lambda$. Since there is a simple formula available for F(x), calculators generally do not include this distribution in their statistical buttons.

The exponential distribution models a situation in which the variable has no *memory*.

**HOMEWORK**: page 159

**Gamma Distribution**: Uses the Gamma function...general case distribution which leads to the very useful Chi-Square distribution.

**Chi-Square Distribution**: Do this…see text for formulas