Set 17

The **gamma function**, $\Gamma(\alpha)$ is defined for $\alpha > 0$ by:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

It can be shown through integration by parts that the gamma function satisfies the relation $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ for all $\alpha > 0$. It can also be shown that $\Gamma(1) = 1$.

Putting these two facts together yields the property that $\Gamma(n) = (n-1)!$ for any positive integer n.

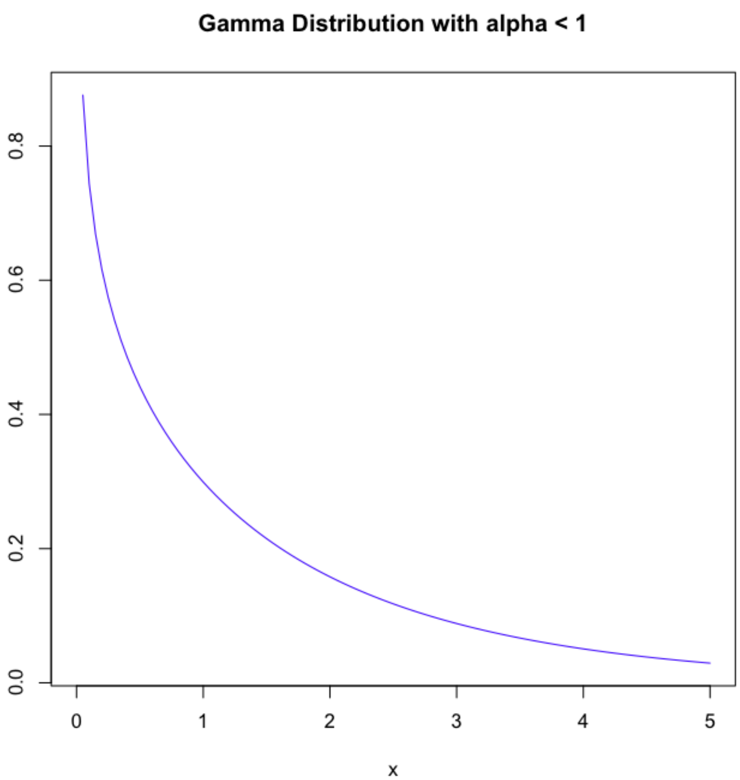
A continuous random variable X has **gamma distribution** with parameters $\alpha > 0$ and $\beta > 0$ if the pdf is

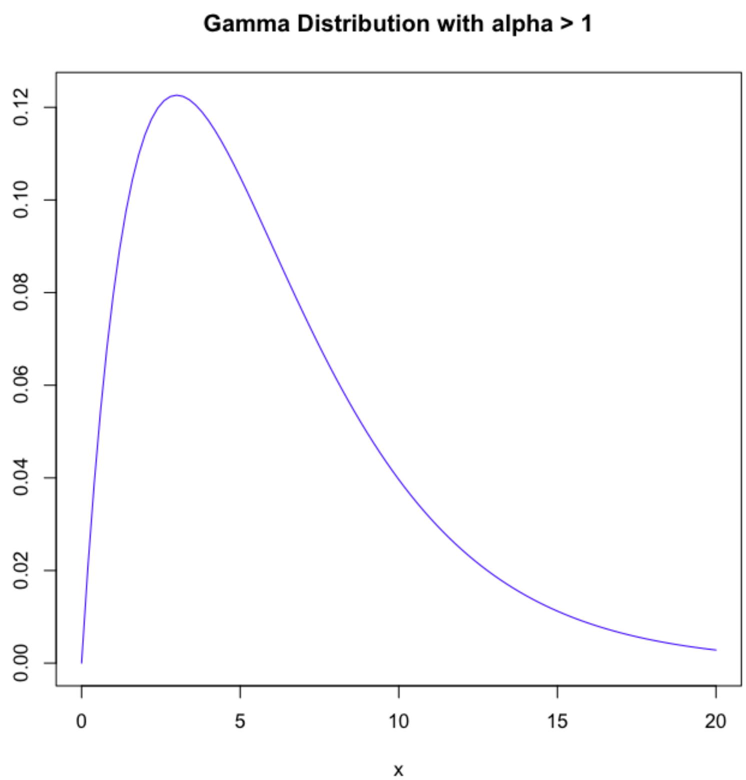
$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The gamma distribution is often used as a probability model for waiting times (e.g. time until death, time until failure).

We call β the **scale parameter** (since it stretches/compresses the pdf) and α the **shape parameter** (since it determines the shape of the pdf).

- $E(X) = \alpha\beta$
- $V(X) = \alpha\beta^2$
- There are two basic shapes for the gamma distribution. The left image is the shape for $\alpha \leq 1$, and the right image is for $\alpha > 1$





- For most values of α, β a closed-form expression for the cdf does not exist; tables or software packages are used. In cases where α is an integer, however, we can calculate probabilities by integrating.

Example: Suppose $X \sim Gamma(\alpha = 2, \beta = 3)$. Calculate $P(X \leq 5)$.

The **exponential distribution** is a member of the gamma family when $\alpha = 1$. The random variable X has exponential distribution with parameter λ ($\lambda > 0$) if the pdf is:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Note: Be aware that a second definition exists, with a parameter θ where $\theta = 1/\lambda$. We will not be using this alternate definition.

We find $E(X)$ and $V(X)$ either by integrating, or by recognizing that if $X \sim Exp(\lambda)$ then $X \sim Gamma(\alpha = 1, \beta = 1/\lambda)$. Either way gives us:

$$E(X) = \frac{1}{\lambda}\ ,\ V(X) = \frac{1}{\lambda^2}$$

Unlike other gamma distributions, the pdf of the exponential distribution can be easily integrated, giving us

$$P(X \leq x) = F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

As with other gamma distributions, the exponential distribution is usually used to model waiting times.

Example: During the lunch hour, the average waiting time to use an automatic bank machine is 6 minutes. Let the random variable X measure the time (in minutes) that a customer waits before service. It is known that X has exponential distribution. What is the probability that a customer will need to wait at least 9 minutes?

Relationship between Poisson and Exponential Distributions:

Suppose we have a Poisson process, where events occur at a rate of λ occurrences per unit of time/space.

If random variable X denotes the number of occurrences of an event in a unit of time/space then $X \sim Poisson(\lambda)$.

If we now let the random variable Y measure the units of time/space until the next occurrence then $Y \sim Exp(\lambda)$.

Example: In our last example, the time (in minutes) between customers had exponential distribution with $\lambda = 1/6$.

If we now count the number of customers per minute, then this would have Poisson distribution with $\lambda = 1/6$. There is an average rate of is $1/6$ customers per minute (or 1 customer per 6 minutes) for the machine.

Note: More generally, there is also a relationship between Poisson and Gamma distributions. Suppose again that we have a Poisson process, where events occur at a rate of λ occurrences per unit of time/space

If we let the random variable Y measure the units of time/space until the k^{th} occurrence, then $Y \sim Gamma(\alpha = k, \beta = 1/\lambda)$.

Example: It is known that accidents in a factory follow a Poisson process, with an average rate of 1 accident per week. What is the probability that the next accident at the factory will occur within the next two weeks?

The **memoryless property:** Suppose $X \sim Exp(\lambda)$. Then for any $a, b \geq 0$:

$$P(X \geq a + b | X \geq b) = P(X \geq a)$$

This means that the probability of a person needing to wait at least a minutes more if they've already waited b minutes, is **the same** as the probability of a newly-arrived person needing to wait a minutes.

Example: Suppose I've already been waiting to use the bank machine for six minutes. What is the probability my total waiting time will be at least 10 minutes?

Let X and Y be discrete random variables defined on some sample space S. The **joint probability function** $f(x, y)$ is defined as:

$$f(x, y) = P(X = x \text{ and } Y = y)$$

Let A be any set of (x, y) pairs. Then;

$$P((X, Y) \in A) = \sum_{(x,y)\in A} \sum f(x, y)$$

Example: Suppose that we consider the manufacture of wind turbines. Before the turbines are shipped, they are checked for flaws and repaired (if necessary). Let X denote the number of manufacturing flaws in a randomly selected turbine. Let Y denote the maximum number of days it takes to repair the flaws.

The following the the **joint probability table** for the probability function $f(x, y)$;

		y		
$f(x,y)$		0	1	2
	0	0.512	0.000	0.000
x	1	0.000	0.102	0.008
	2	0.000	0.175	0.089
	3	0.000	0.015	0.099

Note that $\displaystyle\sum_{\text{all } (x,y)}\sum f(x, y) = 1$

Example: Calculate $P(X \geq 2 \cap Y = 2)$.

The **marginal probability function** of X and Y, denoted $f_X(x)$ and $f_Y(y)$ are:

$$f_X(x) = \sum_y f(x,y) \quad f_Y(y) = \sum_x f(x,y)$$

Example: Find $f_X(x)$ and $f_Y(y)$ for the previous example.

If X and Y are independent random variables, then $f(x, y) = f_X(x) f_Y(y)$ for every (x, y) pair.

We can show without too much difficulty that X and Y are not independent in our turbine example.

We can extend our definitions quite naturally to any sequence $X_1, X_2, \ldots, X_n$ of random variables.

Example: Suppose that in a copy shop, three photocopiers work continually. Let X_i be the number of paper jams that copier i experiences in a day, where $i = 1, 2, 3$. Suppose that X_1, X_2, X_3 are independent, $X_1 \sim Poisson(\lambda = 4)$, $X_2 \sim Poisson(\lambda = 3)$, $X_3 \sim Poisson(\lambda = 10)$. Find the joint pmf $f(x_1, x_2, x_3)$.

Let $X_1, X_2, \ldots, X_n$ be jointly distributed random variables with joint pmf $f(x_1, x_2, \ldots, x_n)$.

For any function $h(X_1, X_2, \ldots, X_n)$ the expected value $E[h(X_1, X_2, \ldots, X_n)]$ is:

$$\sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} h(X_1, X_2, \ldots, X_n) f(x_1, x_2, \ldots, x_n)$$

Example: Let X and Y be jointly distributed random variables with the following joint pmf

$f(x,y)$		y = 1	2
	5	0.3	0.15
x	10	0.2	0.35

Calculate $E(X + Y^2)$.

Example: Recall the wind turbine experiment from the other day. The random variable X denotes the number of manufacturing flaws in a randomly selected turbine and the random variable Y denotes the number of days it takes to repair the flaws.

$f(x,y)$		y = 0	1	2
	0	0.512	0.000	0.000
x	1	0.000	0.102	0.008
	2	0.000	0.175	0.089
	3	0.000	0.015	0.099

The company pays $\$18\ 000$ for each turbine. For each flaw present in the manufacturing process, the company will subtract $\$500$ from the payment amount. For each extra day it takes to complete a turbine due to repairs, the company will subtract $\$1000$ from the payment amount.

What is the expected amount per turbine that the company will pay?

For each turbine, the payment amount is $h(X,Y) = 18000 - 500X - 1000Y$. We wish to find $E(h(X,Y))$

Let X and Y be jointly distributed random variables. The **covariance** between X and Y, written $Cov(X,Y)$ is:

$$\sum_x \sum_y (x-\mu_x)(y-\mu_y)f(x,y)$$

Computational Shortcut: $Cov(X,Y) = E(XY) - \mu_x \mu_y$

Interpretation: Unlike variance, covariance may be positive or negative.

A positive covariance implies that large values of X correspond with large values of Y. A negative covariance implies that large values of X correspond with small values of Y.

Much like variance, the magnitude is difficult to interpret directly.

Example: Find $Cov(X, Y)$, where X and Y have the following joint pmf.

$f(x,y)$		y = 1	2
	5	0.3	0.15
x	10	0.2	0.35

The **correlation coefficient** of X and Y, written ρ or $Corr(X, Y)$, is defined by:

$$\rho = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

Example: Find ρ for the previous example.

1. For any X, Y, we have $-1 \leq \rho \leq 1$.
2. Provided that a and c have the same sign: $Corr(aX+b, cY+d) = Corr(X,Y)$
3. If X and Y are independent then $\rho = 0$.
4. If $\rho = 0$ then X and Y are *uncorrelated* (but still possibly dependent).
5. If $\rho = 1$ or $\rho = -1$ then X and Y have a strictly linear relationship.

Example: Let X take on values $-2, -1, 1, 2$ each with probability 0.25. Let $Y = X^2$, so that Y takes on values $1, 4$ each with probability 0.5.

Clearly, since Y is defined in terms of X, then Y is dependent on X. You can show quickly that $Cov(X,Y) = 0$, and as a consequence, that $\rho = 0$. These random variables are uncorrelated, yet dependent.

Suppose that $X_1, \ldots, X_n$ is a collection of random variables and $a_1, a_2, \ldots, a_n$ are any collection of constants. Then the following hold:

$$E(a_1X_1 + a_2X_2 + \cdots + a_nX_n + b) = a_1E(X_1) + \cdots + a_nE(X_n) + b$$

$$V(a_1X_1 + \cdots + a_nX_n + b) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_jCov(X_i, X_j)$$

It is easy enough to show that $Cov(aX, aX) = a^2V(X)$. Also, $Cov(X, Y) = Cov(Y, X)$.

This gives us:

$$V(a_1X_1 + \cdots + a_nX_n + b)$$

$$= a_1^2V(X_1) + \cdots + a_n^2V(X_n) + 2\sum_{i<j}\sum a_ia_jCov(X_i, X_j)$$

Also, recall that for independent random variables, their covariance is 0.

Then if our random variables are independent of each other then:

$$V(a_1X_1 + \cdots + a_nX_n + b) = a_1^2V(X_1) + \cdots + a_n^2V(X_n)$$

An important consequence: Suppose $X_1, \ldots, X_n$ are independent, each with the same mean and variance μ and σ^2.

If we let $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$, you can show that:

- $E(\overline{X}) = \mu$
- $V(\overline{X}) = \dfrac{\sigma^2}{n}$
- $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$

Similarly, if we let $T = X_1 + \cdots + X_n$, you can show that

- $E(T) = n\mu$
- $V(T) = n\sigma^2$
- $\sigma_T = \sigma\sqrt{n}$

A **statistic** is any function of random variables.

Example: Let $X_1, X_2, \ldots, X_n$ be n random variables. Some common statistics include:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \text{ (the sample mean)}$$

$$T = X_1 + X_2 + \cdots + X_n \text{ (the sample sum)}$$

$$S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1} \text{ (the sample variance)}$$

$$\widetilde{X} = median(X_1, X_2, \ldots, X_n) \text{ (the sample median)}$$

For the observed value of these statistics, we use lower-case letters (e.g. $\overline{x}, t, s^2, \widetilde{x}$)

The probability distribution of a statistic is called a **sampling distribution**.

Typically, we will be interested in statistics defined on a **random sample**, a sequence of independent, identically distributed (**iid**) random variables.

Example: Suppose two people flip a coin three times. Let X_1, X_2 denote the number of tails flipped by the first and second person. Find the sampling distribution of the sample mean.

Since statistics are random variables, they have an expected value and variance.

Example: Find $E(\overline{X})$

Proposition: Any linear combination of independent normal random variables is normally distributed.

Example: The mass of an adult male greater flamingo is normally distributed, with $\mu = 4.5\ kg$, $\sigma = 0.3\ kg$. The mass of an adult female greater flamingo is normally distributed with $\mu = 2\ kg$, $\sigma = 0.1\ kg$. Suppose that a male and female greater flamingo are randomly selected and weighed together. What is the probability that the total mass exceeds $7kg$?

Example: Manufacture of a certain component requires three different machining operations. The amount of time each operation requires (operation time) is normally distributed with mean 10 and variance 4. The three operation times are independent.

For a randomly selected component, what is the probability that the *average* of the three operation times is less than 9 minutes?

Example: Referring to the previous manufacturing example, now suppose that the cost for the first machining operation is \$1 per minute. That for the second and third operations are \$2 per minute and \$3 per minute, respectively.

What is the probability that total cost for making the next component is more than \$70?

Set 21

Recall that using the laws of variance and expectation, we had that if $X_1, X_2, \ldots, X_n$ are *iid* random variables with mean μ and standard deviation σ. Then:

1. The sample mean, $\overline{X}$, has mean μ and standard deviation $\sigma/\sqrt{n}$.
2. The sample sum, T, has mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.

We also had that linear combinations of independent normal random variables were normal. This gave us:

Let $X_1, X_2, \ldots, X_n$ be *iid* random variables, *normally distributed*, with mean μ and standard deviation σ. Then:

1. $\overline{X}$, has normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.
2. T, has normal distribution with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.

Note: The above is true, regardless of sample size.

Example: Suppose it is known that the levels of fluid in soda bottles is normally distributed, with a mean of 355 mL, and standard deviation of 2 mL. Let X_1, X_2, X_3, X_4 denote liquid content of four randomly selected bottles. Find the probability that the average liquid content will be less than 356 mL.

Central Limit Theorem: Let $X_1, X_2, \ldots, X_n$ be *iid* random variables, each with mean μ and standard deviation σ. Provided that n is large enough:

1. $\overline{X}$, has *approximately* normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.
2. T, has *approximately* normal distribution with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.

Note: The above is true, regardless the underlying distribution. The larger the sample size is, the closer $\overline{X}$ and T will be to a normal distribution.

Typically we use $n = 30$ as the cut-off for being "large enough".

Example: The number of bacteria per mL sample of water has a Poisson distribution, with an average of 50 bacteria per sample. Suppose that 100 samples are tested. What is the probability that the average number of bacteria per sample is at least 52?

Example: In a particular lake, the amount of pollutant in a $1\ L$ sample is has a mean of $6\ mg$ with a standard deviation of $1\ mg$. Suppose we take 50 randomly selected samples, each of $1\ L$ of lake water. What is the probability that the total amount of pollutant will be between $295\ mg$ and $305\ mg$?

Example: Pheasants in a particular region were found to have an appreciable mercury contamination. The mercury level in parts per million for these birds is normally distributed with mean 0.25 and standard deviation 0.08. If I select 4 pheasants at random, what is the probability that the mean mercury level will be greater than $0.3\ ppm$?

Example: Suppose again that we select 4 pheasants at random. What is the probability that all of the pheasants will have a mercury level which is less than 0.2?