WEEK 8

8. Sampling and Sampling Distributions

In order to make any statistical inference about the population (parameters), we repeat the experiment many times (say n times) and based on these experimental observations we make some statistical inference about the population unknowns (usually the mean and variance).

The goal of any field of positive science is to understand the nature (which we will call population). Understanding means that to get some information about the unknowns (which we call parameter/parameters). The parameters are non-measurable real numbers which characterize the population.

Let $X_1, X_2, ..., X_n$ be a set of random variables.

	X_1 is the random variable at the first trial X_2 is the random variable at the second trial X_3 is the random variable at the third trial . . . X_n is the random variable at the nth trial

Definition: If the random variables $X_1, X_2, ..., X_n$ are independent and identically distributed, then it is called a *random sample*.

That is, a random sample is $X_1, X_2, ..., X_n \; iid \; f(x;\theta)$. Here θ is the parameter which characterize the population. Actually, a random sample does not have to be independent and identically distributed random variables but in our class when we say "a random sample" we will understand that $X_1, X_2, ..., X_n$ independent and identically distributed random variables with a probability (or probability density) function $f(x;\theta)$.

<u>Example</u>: Consider an experiment of tossing a coin 5 times and repeat the experiment 5 times. That is, the first person tosses a coin 5 times. Then the second person tosses the same coin 5 times and it continuous until the fifth person. What about the random variables:

X_1 is a random variable which counts the number of tails at the first trial (say 2 tails),

X_2 is a random variable which counts the number of tails at the second trial (say 3 tails),

X_3 is a random variable which counts the number of tails at the third trial (say 3 tails),

X_4 is a random variable which counts the number of tails at the fourth trial (say 4 tails)

X_5 is a random variable which counts the number of tails at the fifth trial (say 3 tails).

Note that X_1, X_2, X_3, X_4, X_5 is a random sample (of size 5) and for each $i = 1,2,3,4,5$, $X_i \sim Binom(5,1/2)$. Since each X_i is a random variable, it is a function from the sample space to real line ($X_i : \Omega \to \mathbb{R}$). As it is given above, we observe the following values as:

$$X_1(w) = x_1 = 2, \quad X_2(w) = x_2 = 3, X_3(w) = x_3 = 3, X_4(w) = x_4 = 4, X_5(w) = x_5 = 3.$$

These values ($x_1, x_2, ..., x_n$) are the sample values.

Note that, $X_i \sim Binom(5,1/2)$,

$$E(X) = np = 5(1/2) = 2.5 \quad \text{and} \quad Var(X) = npq = 5(1/2)(1/2) = 1.25.$$

Remember the normal distribution again. If the random variable X is normally distributed with mean μ and variance σ^2. Note that $E(X) = \mu$ and $Var(X) = \sigma^2$. Define the standard normal random variable $Z = (X - \mu)/\sigma$ and it is obviously, $Z \sim N(0,1)$ and moreover $X = \mu + \sigma Z$.

	$Z \sim N(0,1)$ and $X = \mu + \sigma Z$ Here, μ and σ^2 are the parameters to be estimated. That is, μ is the population mean and σ^2 is the population variance.

Here, μ and σ^2 are the parameters to be estimated. That is, μ is the population mean and σ^2 is the population variance. In order to estimate the population mean and the population variance, we use the sample mean and and the sample variance (the reasons to be used these estimators will be explained later) defined as

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ , sample mean ,} \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \text{ , sample variance.}$$

A) Let $X_1, X_2, ..., X_n \, iid \, N(\mu, \sigma^2)$ and define the sample mean and the sample variance as it is given above:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \ , \ S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \ .$$

Notice that,

$$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{\mu + \mu + ... + \mu}{n} = \frac{n\mu}{n} = \mu$$

and

$$Var(\bar{X}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2 + \sigma^2 + ... + \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \ .$$

Since the sum of independent and normally distributed random variables is also normally distributed random variable we have $\bar{X}_n \sim N(\mu, \sigma^2 / n)$ which implies that

$$\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1)$$

or equivalently

$$\frac{\sum_{i=1}^{n} X_i - E\left(\sum_{i=1}^{n} X_i\right)}{\sqrt{Var\left(\sum_{i=1}^{n} X_i\right)}} \sim N(0,1) \ .$$

Note that if we have a random sample from a $N(\mu, \sigma^2)$ distribution (or population), the random variable $\sqrt{n}(\bar{X}_n - \mu) / \sigma$ is distributed as standard normal and therefore the standard normal distribution ($N(0,1)$ can be taken as a sample distribution. The test statistic (will be explained later) Z can be used to make any statistical inference about the population mean μ when the population variance σ^2 is known.

B) Now let us consider the sample variance, S_n^2. Remember that $Var(X) = E(X^2) - (E(X))^2$. Since, $E(X) = \mu$ and $Var(X) = \sigma^2$ we have $E(X^2) = \sigma^2 + \mu^2$. Note that the sample variance S_n^2 can also be wirtten as

$$(n-1)S_n^2 = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \sum_{i=1}^{n} X_i - n\bar{X}_n^2$$

and therefore,

$$E(S_n^2) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n}X_i^2 - n\bar{X}_n^2\right)$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}E(X_i^2) - nE(\bar{X}_n^2)\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n((\sigma^2 / n) + \mu^2)\right)$$
$$= \frac{1}{n-1}\left(n\sigma^2 + \cancel{n\mu^2} - \sigma^2 - \cancel{n\mu^2}\right) = \frac{(n-1)\sigma^2}{n-1} = \sigma^2.$$

The result is also true for non-normal sample. That is, $E(S_n^2) = \sigma^2$ and therefore the sample variance S_n^2 can be used to estimate.

<u>Theorem</u> (without proof): Let $X_1, X_2, ..., X_n$ be a random sample from normal population with mean μ and variance σ^2. That is, $X_1, X_2, ..., X_n$ *iid* $N(\mu, \sigma^2)$ random variables. The sample mean and variance are

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \ , \ S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \ .$$

Then

a) $\bar{X}_n \sim N(\mu, \sigma^2 / n)$,

b) $\bar{X}_n$ and S_n^2 are independent,

c) $\dfrac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$.

Using this theorem, we can calculate the expected value of the sample mean much easily. Remember that if $X \sim \chi_p^2$ then $E(X) = p$ and $Var(X) = 2p$. Since, $(n-1)S_n^2 / \sigma^2 \sim \chi_{n-1}^2$ the mean of the sample variance, we have

$$E((n-1)S_n^2 / \sigma^2) = (n-1) \text{ and } Var((n-1)S_n^2 / \sigma^2) = 2(n-1)$$

and therefore, $E((n-1)S_n^2 / \sigma^2) = (n-1) \Rightarrow E(S_n^2) = \sigma^2$. Moreover, we can also calculate the variance of the sample variance by using the theorem (c)

$$Var\left(\frac{(n-1)S_n^2}{\sigma^2}\right) = Var(\chi_{n-1}^2) = 2(n-1) \Rightarrow \frac{(n-1)^2}{\sigma^4}Var(S_n^2) = 2(n-1) \Rightarrow Var(S_n^2) = \frac{\sigma^4}{n-1}.$$

Since, S_n^2 can be taken as an estimator of the population variance and we have the distrtibution of S_n^2, it can be used to make statistical inference about the population variance. The distribution of S_n^2 is the chi-square and therefore, the chi-square distribution can be considerd as a sample distribution. When we were discussing the Gamma distribution, we have seen that the chi-square distribution is a special case of the Gamma distrtibution. The chi-square distribution can also be obtaine from the normal distribution. That is, if $Z \sim N(0,1)$ then $Z^2 \sim \chi_1^2$ and moreover if $Z_1, Z_2 ..., Z_k$ are independent standard normal random variables, then $Z_1^2 + Z_2^2 + .. + Z_k^2 \sim \chi_k^2$.

C) If $X_1, X_2, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$ we know that

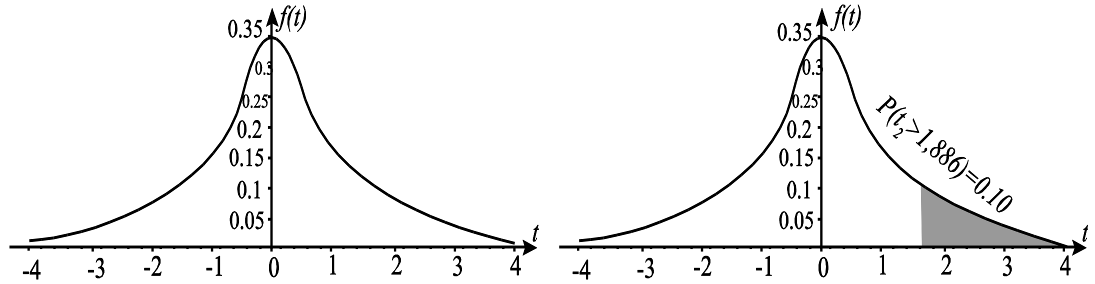
$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1) \quad \text{and} \quad \frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2 .$$

That is, the normal and chi-square distributions are sample distribution. We also know that the sample mean and the sample variance are independently distributed random variables.

t **distribution:** Consider two independent random variables X and Y such that $X \sim N(0,1)$ and $Y \sim \chi_p^2$. Then the probability density function of $T = X / \sqrt{Y / p}$ is

$$f_T(t) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p+1}{2}\right)(p\pi)^{1/2}} \frac{1}{\left(1+\frac{t^2}{p}\right)^{(p+1)/2}} \ , t \in \mathbb{R} \ .$$

If a random variable T has the probability density function as given above we say that T is distributed as t with p degrees of freedom (Student's t distribution) and denoted by $T \sim t_p$. The graph of the probability density function of the t distribution is given below.



The graph of the probability density function of t distribution with 2 degrees of freedom

As we are going to see later, if we want to make any statistical inference about the mean of a normal distribution, we use the $Z = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ statistic. If the variance is a parameter, then it should be estimated. Consider any statistical inference about the normal mean μ when σ^2 is unknown. Since σ^2 is a parameter (unknown) we use its estimator S_n^2. That is,

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n}$$

to make any statistical inference about mean.

Notice that if $X_1, X_2, ..., X_n$ is a random sample from a $N(\mu, \sigma^2)$ population, we have

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1) \text{ and } \frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2 .$$

Moreover the sample mean and the sample variance are independent ($\bar{X}_n$ and S_n^2 are independent). Therefore,

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)/\sigma}{\sqrt{S_n^2/\sigma^2}} = \frac{\sqrt{n}(\bar{X}_n - \mu)/\sigma}{\sqrt{[(n-1)S_n^2/\sigma^2]/(n-1)}} \sim t_{n-1} .$$

That is,

$$T = \sqrt{n}(\bar{X}_n - \mu)/S_n \sim t_{n-1}$$

which is another sample distribution which can be used to make any statistical inference about the normal mean μ when the variance σ^2 is unknown.

The t distribution is commonly used in many statistical problems (hypothesis testing, confidence intervals and regression analysis) that we are going to discuss some of the applications. Since it is very useful distribution the probabilities of the distribution for various degrees of freedom have been tabulated and they can be found in any basic statistical textbooks. You can even find these probabilities by using your mobile phones (download the application "probability distributions", for example if $t \sim t_{10}$ then $P(t > 2) = 0.0367$ and if $X \sim \chi_{10}^2$ then $P(X > 12) = 0.285$ and more many distributions you can find the application).

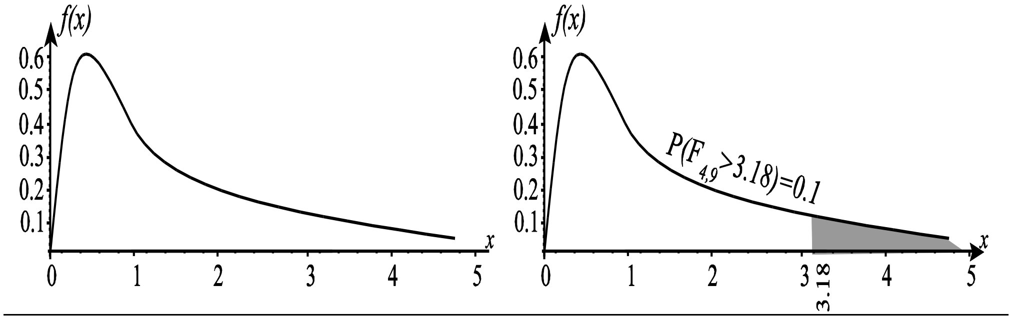
D) Another sample distribution is the F distribution. Suppose we want to compare two population mean. In order to make any statistical inference about two population means, we need to assume that the variances are the same. In order to test (check) the equality of the variances, we use F statistic. We use F statistic to check the model adequacy in regression analysis. We will also discuss model fitting and the use of F distribution later.

Note that if two independently distributed random variables X and Y are distributed as chi-square with p and q degrees of freedom ($X \sim \chi_p^2$ and $Y \sim \chi_q^2$) respectively, the probability density function of

$$F = \frac{X / p}{Y / q}$$

is

$$f(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} \left(\frac{p}{q}\right)^{p/2} \frac{x^{(p-2)/2}}{\left(1+\frac{p}{q}x\right)^{(p+q)/2}} \quad , \quad x \in \mathbb{R}^+ .$$



The probability density function of F distribution (for $p = 4$ and $q = 6$)

If a random variable F has such a probability density function, we say that F is distributed as F with p and q degrees of freedeoms and denoted by $F \sim F(p,q)$. The graph of the probability density function of the F distribution is given above.

Let $F \sim F(p,q)$ the the mean of the F disatribution can be calculated from tye mean of chi-square distributions as shown below. Note that if $F \sim F(p,q)$ the the random varaibles X and Y are independently distributed ($X \sim \chi_p^2$, $Y \sim \chi_q^2$) such that

$$F = \frac{X / p}{Y / q}$$

and therefore the mean of the F disytribution is

$$E(F) = E\left(\frac{X / p}{Y / q}\right) = E\left(\frac{X}{p}\right) E\left(\frac{q}{Y}\right) = \frac{q}{q-2} .$$

<u>Example</u>: Let $X_1, X_2 \ldots, X_n \sim N(\mu_x, \sigma_x^2)$ and $Y_1, Y_2 \ldots, Y_m \sim N(\mu_y, \sigma_y^2)$ be two independent samples. In order to estimate the ratio σ_x^2 / σ_y^2, a reasonable estimator will be $S_{n,X}^2 / S_{m,Y}^2$. ıf we want to find the distribution of the ratio of two sample means, the ratio can be rewritten as

$$F = \frac{S_{n,X}^2 \,/\, S_{m,Y}^2}{\sigma_x^2 \,/\, \sigma_y^2} = \frac{S_{n,X}^2 \,/\, \sigma_x^2}{S_{m,Y}^2 \,/\, \sigma_y^2} = \frac{\dfrac{(n-1)S_{n,X}^2}{\sigma_x^2} \,/\, (n-1)}{\dfrac{(m-1)S_{m,Y}^2}{\sigma_y^2} \,/\, (m-1)}.$$

Since $\dfrac{(n-1)S_{n,X}^2}{\sigma_x^2} \sim \chi_{n-1}^2$ and $\dfrac{(m-1)S_{m,Y}^2}{\sigma_y^2} \sim \chi_{m-1}^2$

and they are independently distributed random variables, according to the definition given above it is obvious that the ratio is distributed as F. That is,

$$F = \frac{S_{n,X}^2 \,/\, S_{m,Y}^2}{\sigma_x^2 \,/\, \sigma_y^2} \sim F(n-1, m-1) \quad \text{or} \quad F = S_{n,X}^2 \,/\, S_{m,Y}^2 \sim F(n-1, m-1).$$

Thus, the F statistic can be used to make any statistical inference about the ratio σ_x^2 / σ_y^2 (or to test whether $\sigma_x^2 = \sigma_y^2$). And the F statistic can be considered another sample distribution. The probabilities of the distribution have been tabulated for various degrees of freedoms p and q. these probabilities are available in any textbook. For example, if $X \sim F(4,5)$, then $P(0 < X < 5.2) = 0.95$.

<u>The Central Limit Theorem</u> (VERY IMPORTANT)

Note that if $X_1, X_2 \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$ population, we know that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1), \quad \frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2 \text{ and } \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \sim t_{n-1}.$$

In general, the distribution of the population is unknown.

Note: If a random sample $X_1, X_2 \ldots, X_n$ we will usually denote this as $\underset{\sim}{X} = (X_1, X_2 \ldots, X_n)'$. Any function of the sample will be called an *<u>estimator</u>*. That is, $T_n = T_n(\underset{\sim}{X}) = T_n(X_1, X_2, \ldots, X_n)$.

Let $X_1, X_2 \ldots, X_n$ be a random sample from a population with probability (or probability density) function $f(x;\theta)$, $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$.

If $T_n = T_n(\underset{\sim}{X}) = \bar{X}_n$ then

$$T_1 = X_1,\ T_2 = \frac{X_1 + X_2}{2},\ T_3 = \frac{X_1 + X_2 + X_3}{3},\ \ldots,\ T_n = \frac{X_1 + X_2 + X_2 + \ldots + X_n}{n}$$

Definition: Let $X_1, X_2 \ldots, X_n$ be a random sample from a population with probability (or probability density) function $f(x;\theta)$, $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Let T_n be any estimator for the parameter θ. We say that T_n converges to the parameter θ in probability and denoted by $T_n \xrightarrow{P} \theta$ as $n \to \infty$ if for any $\varepsilon > 0$

$$\lim_{n\to\infty} P(|T_n - \theta| > \varepsilon = 0 \ .$$

Chebyshev's Inequality:

Let X be any random variable such that $E(X) = \mu$ and $Var(X) = \sigma^2$ then

$$P(|X - \mu| > \varepsilon) \le \frac{E(X-\mu)^2}{\varepsilon^2}.$$

Example (Weak Law of Large Numbers, WLLN):

Let $X_1, X_2 \ldots, X_n$ be a random sample from a population with probability (or probability density) function $f(x;\theta)$, $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ then the sample mean converges to the population mean in probability, that is $\bar{X}_n \xrightarrow{P} \mu$ as $n \to \infty$. Similarly, the sample variance converges to the population variance in probability, namely $S_n^2 \xrightarrow{P} \sigma^2$ as $n \to \infty$.

Before we state the central limit theorem, let us introduce another type of convergence known as the convergence in distribution.

Definition: Let X_n be any sequence of random variables with distribution function $F_n(x)$ and X be a random variable with cumulative distribution function $F(x)$. We say that X_n converges to the random variable X in distribution and denoted by $X_n \xrightarrow{D} X$ as $n \to \infty$ if for every x such that $F(x)$ is continuous at the points x, $F_n(x) \to F(x)$ as $n \to \infty$.

Theorem (The central Limit Theorem): Let $X_1, X_2 \ldots, X_n$ be a sequence of independent and identically distributed random variables with probability (or probability density) function $f(x;\theta)$, $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ such that $\sigma^2 < \infty$. Then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0,1) \text{ as } n \to \infty.$$

Note that since

$$\frac{\sum_{i=1}^{n} X_i - E\left(\sum_{i=1}^{n} X_i\right)}{\sqrt{Var\left(\sum_{i=1}^{n} X_i\right)}} = \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} E(X_i)}{\sqrt{\sum_{i=1}^{n} Var(X_i)}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} = \frac{n\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)}{\sqrt{n}\,\sigma} = \frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma}$$

the central limit theorem can also be stated as

$$\frac{\sum_{i=1}^{n} X_i - E\left(\sum_{i=1}^{n} X_i\right)}{\sqrt{Var\left(\sum_{i=1}^{n} X_i\right)}} \xrightarrow{D} N(0,1) \text{ as } n \to \infty .$$

The theorem says that whatever the population is, the sample mean approaches to the standard normal random variable when the sample size (here n) is large enough. Usually, the distribution of the population is unknown and in order to make any statistical inference we need the normalility assumption. The CLT provides such assumption when the data do not obey the normality. Using the central limit theorem (CLT), we can do any statistical inference if the data do not come from a normal population. Also we can calculate many probabilities by using the CLT.

Example: Consider an experiment of tossing a coin 100 times. Find the probability of observing more than 60 tails.

The probability can be calculated directly by using the binomial distribution. Note that if X is a random variable which counts the number of tails in the experiment it is distributed as binomial with $p = 1/2$ and $n = 100$. That is $X \sim Binom(1/2, 100)$ and we want to calculate $P(X > 60)$. Note that the probability function of X is

$$P(X = x) = \binom{100}{x}\left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{100-x}, \; x = 0,1,2,...,100$$

and the exact probability is calculated as (by computer) as

$$P(X > 60) = \sum_{x=61}^{100} P(X = x) = \sum_{x=61}^{100} \binom{100}{x}\left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{100-x}$$

$$= \left(\frac{1}{2}\right)^{100} \sum_{x=61}^{100} \frac{100!}{x!\,(100-x)!} = 0.02844.$$

This probability can easily calculated by using the CLT approximately. Let X_i be a random variable which counts the number of tails at the $i.th$ trial (which is either 0 or 1). As it is obviously seen that each random variable $X_i \sim Bern(1/2)$ and the sum of these random variables gives the total number of tails in 100 trial. That is

$$X = \sum_{i=1}^{100} X_i \ , \quad E\left(\sum_{i=1}^{100} X_i\right) = 100\left(\frac{1}{2}\right) = 50 \ \text{ and } \ Var\left(\sum_{i=1}^{100} X_i\right) = 100\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = 25$$

and therefore by the central limit theorem we calculate the probability as

$$P(X > 60) = P\left(\sum_{i=1}^{100} X_i > 60\right) = P\left(\frac{\sum_{i=1}^{100} X_i - E\left(\sum_{i=1}^{100} X_i\right)}{Var\left(\sum_{i=1}^{100} X_i\right)} > \frac{60-50}{5}\right) \cong P(Z > 2) = 0.0228 \, .$$

As you notice that this probability is very close the exact probability calculated by computer. If we had more experiment we get much closer number the the exact probability.

<u>Example</u>:

Let $X_1, X_2 \ldots, X_n$ be a sequence of independent and identically distributed random variables with probability (or probability density) function $f(x;\theta)$ and $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Then we know that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0,1) \ \text{ as } n \to \infty \ \text{ and } S_n^2 \xrightarrow{P} \sigma^2 \ \text{ as } n \to \infty$$

where

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \ \text{ and } S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \ .$$

Consider the t statistic to make any inference about the population mean:

$$t_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)/\sigma}{S_n/\sigma} \xrightarrow{D} N(0,1) \ \text{ as } n \to \infty \, .$$

That is, for large n the t statistic also congerges to normal distribution, $t_n \xrightarrow{D} N(0,1)$ as $n \to \infty$. That is, if we have large number of observations, we can still use the normal approximation. However if we do not have large number of observations, we should prefer to use t distribution. In summary,

a) For larger number of observations, we use Z distribution

b) For small number of observation, we use t distribution.

Order Statistics:

In general, the distribution of the sample is unknown. The order statistics are very helpful to get an intuitive information about the sample. Let $X_1, X_2 \ldots, X_n$ be a random sample from a population whit a probability (or probability density) function $f(x;\theta)$ and cumulative distribution function $F(x;\theta)$. Using the values of the order statistics we produce some plots (Box-Cox plot, normal probability plot, histogram etc.) to get some distributional properties of the sample. First of all, we need to define the ordering the sample. Consider two random variables X_1 and X_2 defined on the same sample space. We say that X_1 is smaller than X_2 if for any $w \in \Omega$, $X_1(w) \le X_2(w)$ then we define the order statistics as

$$
\begin{aligned}
X_{(1)} &= \min\{X_1, X_2, \ldots, X_n\}, \\
X_{(2)} &= \text{second smallest}\{X_1, X_2, \ldots, X_n\} \\
X_{(3)} &= \text{third smallest}\{X_1, X_2, \ldots, X_n\} \\
&\;. \\
&\;. \\
X_{(n)} &= \max\{X_1, X_2, \ldots, X_n\}.
\end{aligned}
$$

All these order statistics are function of the sample and therefore these can be considered as estimators. Moreover, even the random sample $X_1, X_2 \ldots, X_n$ is independent and identically distributed random variables, the order statistics defined as a function of the same sample are not independent and as it is clearly seen we have $X_{(1)} \le X_{(2)} \le X_{(3)} \le \ldots \le X_{(n)}$ such an ordering. Based on the order statistics we define some measure of tendencies (median, range, percentiles, etc.) as given below. First of all the sample range is given by $\mathcal{R} = X_{(n)} - X_{(1)}$ which is also an estimator. The sample median M and the midrange V are defined based on the order statistics as

$$
M = \begin{cases} X_{((n+1)/2)} & , \quad \text{n is odd} \\ \frac{1}{2}[X_{(n/2)} + X_{((n/2)+1)}] & , \quad \text{n is even} \end{cases}, \qquad V = (X_{(1)} + X_{(n)})/2.
$$

Example: In the following table, the test scores for 50 students and in the second table below the ordered values of the tests scores are given.

66	71	67	69	75	66	64	70	62	83
70	79	74	74	79	94	76	69	88	72
84	76	63	70	77	80	77	72	78	73
75	78	90	76	62	78	78	72	77	72

72	59	73	75	76	80	56	67	69	80
50 Student's test scores									

The mean and standard deviation are calculated from the sample as

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = 73.66 \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = 55.78\,.$$

The ordered values are given below.

56	59	62	62	63	64	66	66	67	67
69	69	69	70	70	70	71	72	72	72
72	72	73	73	74	74	75	75	75	76
76	76	76	77	77	77	78	78	78	78
79	79	80	80	80	83	84	88	90	94
50 Student's test scores in order									

Using these ordered values, we observe that

$$x_{(1)} = 56 \quad , \quad x_{(50)} = 94 \quad , \quad x_{(25)} = 74 \quad , \quad x_{(26)} = 74 \quad , \quad x_{(48)} = 88$$

and the median, range and midrange are found to be

$$m = 0.5[x_{(25)} + x_{(26)}] = 0.5(74+74) = 74\,, \;\; v = (x_{(1)} + x_{(n)})/2 = (56+94)/2 = 75$$

$$\text{and } r = x_{(50)} - x_{(1)} = 94 - 56 = 38\,.$$

When the data is ordered from the smallest to the largest 50% of all observations are smaller than the median (here the median is found to be 74). If 25% of all observations are smaller than or equal to a number (say Q_L) this number is called the first quartile (here $Q_L = 69$) and if 75% of all observations are smaller than or equal to a number (say Q_U) this number is called the third quartile (here $Q_U = 78$) and finally, the second quartile is the median. We can also calculate some percentiles of the sample. If 60% of all observations are smaller than or equal to a number (say a_{60}) then the number is called the 60$^{\text{th}}$ percentile and similarly if 90% of all observations are smaller than or equal to a number (say a_{90}) then the number a_{90} is called the 90$^{\text{th}}$ percentile of the sample. Here, these numbers (known as critical values) are calculated as

$$a_{60} = \frac{x_{(30)} + x_{(30)}}{2} = \frac{76+76}{2} = 76 \text{ and } a_{90} = \frac{x_{(45)} + x_{(46)}}{2} = \frac{80+83}{2} = 81.5\,.$$

Some other values of the percentiles are calculated as follows:

$$a_{99} = x_{(50)} = 94\,,\; a_{95} = x_{(48)} = 88\,,\; a_5 = x_{(3)} = 62\,,\; a_1 = x_{(1)} = 56$$

and

$$a_{10} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{63+64}{2} = 63.5\,.$$

Using some probabilistic calculations, we can also find the probability distributions of the order statistics. The following theorem summarizes the distributional properties of order statistics.

Theorem: Let $X_1, X_2 \ldots, X_n$ be a random sample from a population whit a probability (or probability density) function $f(x;\theta)$ and cumulative distribution function $F(x;\theta)$. Moreover, assume that $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are the order statistics as defined above. Then

a) The probability (or probability density) function of j^{th} order statistics $X_{(j)}$ is

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f(x)[F(x)^{j-1}[1-F(x)]^{n-j} \text{ , } x \in D_X$$

b) The joint probability (or probability density) function of i^{th} and j^{th} order statistics $X_{(i)}$ and $X_{(j)}$ is

$$f_{X_{(i)},X_{(j)}}(x,y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} f(x)f(y)[F(x)^{i-1} *[F(y)-F(x)]^{j-i-1}[1-F(y)]^{n-j} \quad , \quad x<y$$

c) And finally, the joint probability (or probability density) function of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_1,x_2,\ldots,x_n) = \begin{cases} n!\prod_{i=1}^{n} f(x_i) & , \quad x_1 < x_2 < \ldots < x_n \\ 0 & , \quad d.y. \end{cases}$$

(Casella and Berger, 2002, page 229-230).

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from $U(0,\theta)$ distribution. The probability density function and the cumulative distribution of the uniform distribution is given by

$$f(x;\theta) = \begin{cases} 1/\theta & , \quad 0<x<\theta \\ 0 & , \quad d.y. \end{cases} \quad , \quad F(x;\theta) = \begin{cases} 0 & , \quad x<0 \\ x/\theta & , \quad 0 \le x \le \theta \\ 1 & , \quad x>\theta. \end{cases}$$

a) Let us try to find the probability density function of the n^{th} order statistic $X_{(n)}$. By using the theorem given above, the probability density function of the n^{th} order statistics is written for $0<x<\theta$ by

$$f_{X_{(n)}}(x;\theta) = \frac{n!}{(n-1)!(n-n)!}\frac{1}{\theta}\left(\frac{x}{\theta}\right)^{n-1}\left(1-\frac{x}{\theta}\right)^{n-n} = \frac{n}{\theta^n}x^{n-1} \quad , \quad 0<x<\theta$$

or

$$f_{X_{(n)}}(x;\theta)=\begin{cases}\dfrac{n}{\theta^n}x^{n-1} & , \quad 0<x<\theta \\ 0 & , \quad d.y.\end{cases}$$

If we do not remember the statement of the theorem, we can still find the probability density function of the n^th^ order statistic $X_{(n)}$ by using the distribution function of $X_{(n)}$. Remember that the probability density function is the derivative of the cumulative distribution function. First of all, let us try to find the distribution function of $X_{(n)}$. Let $F_{X_{(n)}}(x)$ denote the distribution function of $X_{(n)}$. Obviously $F_{X_{(n)}}(x)=0$ for $x<0$ and $F_{X_{(n)}}(x)=1$ for $x\geq\theta$. Finally, for $0<x<\theta$ we have

$$F_{X_{(n)}}(x)=P(X_{(n)}\leq x)=P(\max\{X_1,\ldots,X_n\}\leq x)=P(X_1\leq x,X_2\leq x,\ldots,X_n\leq x)$$
$$=P(X_1\leq x)P(X_2\leq x)...P(X_n\leq x)=[P(X_1\leq x)]^n=[x/\theta]^n=\theta^{-n}x^n.$$

therefore, the cumulative distribution function and the probability density function of the n$^{\text{th}}$ order statistics $X_{(n)}$ are given below.

$$F(x;\theta)=\begin{cases}0 & , \quad x<0 \\ \theta^{-n}x^n & , \quad 0\leq x<\theta \\ 1 & , \quad x\geq\theta\end{cases} \quad \text{and} \quad f_{X_{(n)}}(x)=\begin{cases}\dfrac{n}{\theta^n}x^{n-1} & , \quad 0<x<\theta \\ 0 & , \quad d.y.\end{cases}$$

The mean and the variance of the n$^{\text{th}}$ order statistic are calculated. First two moments are

$$E(X_{(n)})=\frac{n}{\theta^n}\int_0^\theta x^n\,dx=\frac{n}{n+1}\theta, \qquad E(X_{(n)}^2)=\frac{n}{\theta^n}\int_0^\theta x^{n+1}\,dx=\frac{n}{n+2}\theta^2$$

and therefore, the variance of $X_{(n)}$.

$$Var(X_{(n)})=E(X_{(n)}^2)-[E(X_{(n)})]^2=\frac{n\ \theta^2}{(n+1)^2(n+2)}$$

Note that if we set $T=(n+1)n^{-1}X_{(n)}$ the mean and the variance of T are calculated as

$$E(T)=E\left(\frac{n+1}{n}X_{(n)}\right)=\theta \quad \text{and} \quad Var(T)=Var\left(\frac{n+1}{n}X_{(n)}\right)=\frac{\theta^2}{n(n+2)}.$$