# 7 Random samples and sampling distributions

## 7.1 Introduction - random samples

We will use the term *experiment* in a very general way to refer to some process, procedure or natural phenomena that produces a random outcome.
**Examples:**

- roll a die
- flip a coin 100 times
- flip a coin until we get heads
- pick a U of A student at random and measure his or her height
- call a computer random number generator
- measure the time it takes until a radioactive substance undergoes a decay

The set of possible outcomes is called the sample space. A random variable is a function whose domain is the sample space. We will be primarily interested in RV's whose range is the real numbers (or sometimes $\mathbb{R}^n$). We will usually denote RV's by capital roman letters: $X, Y, Z, U, T, ....$

Suppose we have a random variable $Y$ for our experiment, and we repeat the experiment $n$ times. We denote the resulting values of $Y$ by $Y_1, Y_2, \cdots Y_n$. Note that the $Y_i$ are random variables. If we repeat the experiment another $n$ times, then we will typically get different values for $Y_1, \cdots, Y_n$. This $n$-tuple of random variables $Y_1, \cdots Y_n$ is called a random sampling.

If the process of repeating the experiment does not change the experiment, then the RV's $Y_1, \cdots Y_n$ should all have the same distribution. (We say they are identically distributed.) And if the outcome of one performance of the experiment does not influence the outcome of any other performance fo the experiement, then the RV's $Y_i$ will be independent. We assume that both of these things are true. So a random sample will mean random variables $Y_1, Y_2, \cdots Y_n$ which are idependent and indentically distibuted (i.i.d. for short).

Another way to think of generating a random sample is that there is some large population and a random variable $Y$ defined on this population. We pick $n$ members of the population at random and let $Y_1, Y_2, \cdots, Y_n$ be the resulting values of $Y$. We require that the population be large so that as we randomly draw member from the population, it does not change significantly. If $f(y)$ is the density for $Y$ in the population we can describe this as $Y_1, \cdots Y_n$ is a random sample from a population with density $f(y)$.

In probability theory we usually know what $f(y)$ is and then want to *deduce* something about the random variable $Y$, e.g., what is the probability that $Y$ is greater than 10. In statistics we typically do not know $f(y)$ and want to use the random sample to *infer* something about $f(y)$. For example, we may want to estimate the mean of $Y$. The mean of $Y$ is called the *population mean* and typically denoted by $\mu$. Note that it is a constant, although typically an unknown constant. Given a random sample $Y_1, \cdots Y_n$, the *sample mean* is defined to be

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{1}$$

If $n$ is large than there are theorems that say that $\overline{Y}$ is usually close to $\mu$. However, $\overline{Y}$ is a random variable. Generate another random sample and we will get a different value for $\overline{Y}$. Sometime we will be "unlucky" and get a value for $\overline{Y}$ that is not close to $\mu$. So we should not expect to be able make definitive statements like "$\overline{Y}$ is within 0.01 of $\mu$". The most we can hope for is a statement like "the probability that $\overline{Y}$ is within 0.01 of $\mu$ is at least 0.95."

The RV $\overline{Y}$ is an example of a *statistic*. In general, a statistic is a RV that is a function of the random sample. Here are a few other statistics. The sample variance is

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{2}$$

Here are some order statistics:

$$Y_{(1)} = \min\{Y_1, \cdots, Y_n\}$$

$$Y_{(n)} = \max\{Y_1, \cdots, Y_n\}$$

The median of the random sample is another example of a statistic.

We are especially interested in statistics that should "estimate", i.e., be close to, some population parameter of interest. For example, $\overline{Y}$ is an estimator for $\mu$, $S^2$ is an estimator for $\sigma^2$, $[X_{(1)}, X_{(n)}]$ is an estimator for the range of $X$. In order to study how close our estimator is to the parameter we want to estimate, we need to know the distribution of the statistic. This distribution is often called a sampling distibution. So our study of statistics begins with the study of sampling distributions. But first we review some probability concepts.

## PROBABILITY REVIEW

**Definition 1.** *An* **experiment** *is a well defined procedure (possibly multistage) that produces an* **outcome**. *The set of possible outcomes is called the* **sample space**. *We will typically denote an individual outcome by $\omega$ and the sample space by $\Omega$. An* **event** *is a subset of the sample space.*

A probability measure is a function that gives the probabilities of events. Note that it is not a function on $\Omega$ but rather a function on subsets of $\Omega$. It has to satisfy some properties:

**Definition 2.** *A probability measure is a real-valued function* $\mathbf{P}$ *on subsets of $\Omega$, the set of outcomes, with the following properties.*

1. $\mathbf{P}(A) \geq 0$, *for all events $A$.*
2. $\mathbf{P}(\Omega) = 1$, $\mathbf{P}(\emptyset) = 0$.
3. *If $A_n$ is a pair-wise disjoint sequence of events, i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$, then*

$$\mathbf{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbf{P}(A_n) \tag{3}$$

We will call the pair $(\Omega, P)$ a probability space.

**Remark:** This definition has a cheat. It ignores $\sigma$-fields. Usually the probably measure is not defined on *all* subsets of the sample space, but just on some of them. We will not worry about this issue.

From this definition you can prove the following properties

**Theorem 1.** *Let $P$ be a probability measure. Then*

1. *$\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$ for all events $A$.*

2. *$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$ for all events $A, B$.*

3. *$\mathbf{P}(A \setminus B) = \mathbf{P}(A) - \mathbf{P}(A \cap B)$. for all events $A, B$.*

4. *If $A$ and $B$ are events with $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.*

An important concept in probability is independence. For events it means the following.

**Definition 3.** *Two events are independent if*

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) \tag{4}$$

**Caution:** Independent and disjoint are two quite different concepts. If $A$ and $B$ are disjoint, then $A \cap B = \emptyset$ and so $\mathbf{P}(A \cap B) = 0$. So if they are disjoint, then they are also independent only if $\mathbf{P}(A) = 0$ or $\mathbf{P}(B) = 0$.

A *random variable* is a function whose domain is the sample space $\Omega$. The range can be any set, but in this course the range will almost always be either the real numbers or $\mathbb{R}^n$. In the first case we can say we have a real-valued random variable, but we will typically just say we have random variable. In the second case we can call the random variable a random vector.

In this course random variables will either be discrete or continuous. A RV is discrete if its range is finite or countable, and is continuous otherwise. In the continuous case the range is usually an interval, a half-infinite interval like $[0, \infty)$ or the entire real line.

**Very important idea:** The sample space $\Omega$ may be quite large and complicated. But we may only be interested in one or a few RV's. We would like to be able to extract all the information in the probability space $(\Omega, \mathbf{P})$ that is relevant to our random variable(s), and forget about the rest of the information contained in the probability space.

**Definition 4.** *For a discrete RV the probability mass function (pmf) $p(x)$ of $X$ is the function on $\mathbb{R}$ given by*

$$p(x) = \mathbf{P}(X = x) = \mathbf{P}(\{\omega \in \Omega : X(\omega) = x\})$$

**Notation/terminology:** If we have more than one RV, then we have more than one pmf. To distinguish them we use $p_X(x)$ for the pmf for $X$, $p_Y(x)$ for the pmf for $Y$, etc. Sometimes the pmf is called the "density function" and sometimes the "distribution of $X$." The latter is really confusing as the term "distribution function" refers to something else (the cdf).

**Example - binomial RV** This pmf has two parameters: $p \in [0,1]$ and a positive integer $n$. Suppose we have a coin which has probability $p$ of coming up heads, $1-p$ of coming up tails. We flip the coin $n$ times and let $X$ be the number of heads we get. So the range of the random variable $X$ is $0, 1, 2, \cdots, n$. It can be shown that

$$p(k) = \mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

This is not just about coin flips. We can take any simple experiment which has only two possible outcomes and repeat it $n$ times. The outcomes are typically called success (instead of heads) and failure (instead of tails). The RV $X$ is the number of successes. The parameter $n$ is called the "number of trials." The notation $\binom{n}{k}$ is defined by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}$$

The mean of $X$ is $\mu = np$ and the variance is $\sigma^2 = np(1-p)$.

**Theorem 2.** *Let $X$ be a discrete RV with pmf $p(x)$. Let $A \subset \mathbb{R}$. (Note that $A$ is not an event, but $X \in A$ is.) Then*

$$\mathbf{P}(X \in A) = \sum_{x \in A} p(x)$$

**Definition 5.** *Let $X$ be a discrete RV with probability mass function $p_X(x)$. The expected value of $X$ is denoted $\mathbf{E}[X]$ and is given by*

$$\mathbf{E}[X] = \sum_x x \, p_X(x)$$

*provided that*

$$\sum_x |x| \, p_X(x) < \infty$$

**Terminology/notation** The expected value is also called the mean of $X$ and is often denoted by $\mu$. When the above does not converge absolutely, we say the mean is not defined.

For continuous random variables, we will have integrals instead of sums.

**Definition 6.** *A random variable $X$ is continuous if there is a non-negative function $f_X(x)$, called the probability density function (pdf) or just the density, such that*

$$\mathbf{P}(X \leq t) = \int_{-\infty}^{t} f_X(x)\, dx$$

**Proposition 1.** *If $X$ is a continuous random variable with pdf $f(x)$, then*

1. $\mathbf{P}(a \leq X \leq b) = \int_a^b f(x)\, dx$
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$
3. $\mathbf{P}(X = x) = 0$ *for any* $x \in \mathbb{R}$.

**Remark:** Many books use $f(x)$ or $f_X(x)$ to denote the pmf of a discrete RV instead of $p(x), p_X(x)$. In this course we will always use $p$.

**Caution** Often the range of $X$ is not the entire real line. Outside of the range of $X$ the density $f_X(x)$ is zero. So the definition of $f_X(x)$ will typically involves cases: in one region it is given by some formula, elsewhere it is simply 0. So integrals over all of $\mathbb{R}$ which contain $f_X(x)$ will reduce to intervals over a subset of $\mathbb{R}$. If you mistakenly integrate the formula over the entire real line you will get nonsense.

**Example - Normal distribution:** This pdf has two parameters $\sigma > 0$, $\mu \in \mathbb{R}$. The pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right)$$

The range of a normal RV is the entire real line. It is not obvious that the integral of this function is 1, but it is. The mean of this pdf is $\mu$ and the variance is $\sigma^2$.

**Definition 7.** *Let $X$ be a continuous RV with pdf $f_X(x)$. The expected value of $X$, denoted $\mathbf{E}[X]$ is*

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

*provided that*

$$\int_{-\infty}^{\infty} |x|\, f_X(x)\, dx < \infty$$

If the last integral is infinite, then we say the mean is not defined.

**Another very important idea:** Suppose we have two possible different probability spaces $(\Omega_1, \mathbf{P}_1)$ and $(\Omega_2, \mathbf{P}_2)$, and RV's $X_1$ on the first and $X_2$ on the second. Then it is possible that $X_1$ and $X_2$ have the same range and identical pmf's (if they are discrete) or identical pdf's (if they are continuous). When this happens, if we only look at $X_1$ and $X_2$ then when we do the two experiments, then we won't be able to tell the experiments apart.

**Definition 8.** *Let $X_1$ and $X_2$ be discrete random variables which are not necessarily defined on the same probability space. If $p_{X_1}(x) = p_{X_2}(x)$ for all $x$, then we say $X_1$ and $X_2$ are identically distributed.*

**Definition 9.** *Let $X_1$ and $X_2$ be continuous random variables which are not necessarily defined on the same probability space. If $f_{X_1}(x) = f_{X_2}(x)$ for all $x$, then we say $X_1$ and $X_2$ are identically distributed.*

Suppose $X$ is a RV and $g : \mathbb{R} \to \mathbb{R}$. We can define a new random variable by $Y = g(X)$. Suppose we know the pmf of $X$ (if $X$ is discrete) or the pdf of $X$ (if $X$ is continuous), and we want to compute the mean of $Y$. The long way to do this is to first work out the pmf or pdf of $Y$ and then compute the mean of $Y$. However, there is a shortcut.

**Theorem 3.** *Let $X$ be a discrete RV, $g$ a function from $\mathbb{R}$ to $\mathbb{R}$. Define a new RV by $Y = g(X)$. Let $p_X(x)$ be the pmf of $X$. Then*

$$\mathbf{E}[Y] = \mathbf{E}[g(X)] = \sum_x g(x) p_X(x)$$

**Theorem 4.** *Let $X$ be a continuous RV, $g$ a function from $\mathbb{R}$ to $\mathbb{R}$. Define a new RV by $Y = g(X)$. Let $f_X(x)$ be the pdf of $X$. Then*

$$\mathbf{E}[Y] = \mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$$

**Definition 10.** *The variance of a RV $X$ is*

$$\sigma^2 = \mathbf{E}[(X-\mu)^2],$$

*where $\mu = \mathbf{E}[X]$. If $\mathbf{E}[(X-\mu)^2]$ is infinite we say the variance is infinite.*

This definition works for both discrete and continuous RV's. How we compute $\mathbf{E}[(X-\mu)^2]$ is different in the two cases. If $X$ is discrete,

$$\mathbf{E}[(X-\mu)^2] = \sum_x (x-\mu)^2 \, p_X(x) \tag{5}$$

If $X$ is continuous,

$$\mathbf{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 \, f_X(x) \, dx \tag{6}$$

The standard deviation, $\sigma$, is just the square root of the variance, $\sigma^2$, as the notation suggests.

There is another way to compute the variance.

**Proposition 2.**

$$\sigma^2 = \mathbf{E}[X^2] - \mu^2$$

The quantity $\mathbf{E}[X^2]$ is called the second moment of $X$. For discrete $X$ it is given by

$$\mathbf{E}[X^2] = \sum_x x^2 \, p_X(x) \tag{7}$$

For continuous $X$ it is given by

$$\mathbf{E}[X^2] = \int_{-\infty}^{\infty} x^2 \, f_X(x) \, dx \tag{8}$$

---

**End of lecture on Thurs, 1/11**

---

Next we define the cdf of a RV. The definition is the same for both discrete and continuous RV's.

**Definition 11.** *The cumulative distribution function (cdf) of the random variable $X$ is the function*

$$F_X(x) = \mathbf{P}(X \leq x)$$

Why introduce this function? It will be a powerful tool when we look at functions of random variables and compute their density.

**Theorem 5.** *Let $X$ be a continuous RV with pdf $f(x)$ and cdf $F(x)$. Then they are related by*

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)\, dt, \\ f(x) &= F'(x) \end{aligned}$$

Now suppose we have $n$ random variables instead of just one. Knowing their individual pmf's or pdf's is not enough. We must introduce joint pmf's and pdf's. For discrete RV's we have

**Definition 12.** *Let $X_1, X_2, \cdots, X_n$ be discrete RV's. Their joint pmf is*

$$p_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \tag{9}$$

For continuous RV's we have

**Definition 13.** *Random variables $X_1, X_2, \cdots, X_n$ are jointly continuous if there is a function $f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n)$ on $\mathbb{R}^n$, called the joint probability density function, such that*

$$\begin{aligned} &\mathbf{P}(X_1 \leq s_1, X_2 \leq x_2 \cdots, X_n \leq x_n) \\ = &\int\int_{x_1 \leq s_1, x_2 \leq s_2 \cdots, x_n \leq s_n} f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n)\, dx_1\, dx_2 \cdots dx_n \end{aligned}$$

In order for a function $f(x_1, x_2, \cdots, x_n)$ to be a joint density it must satisfy

$$\begin{aligned} f(x_1, x_2 \cdots x_n) &\geq 0 \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_n)\, dx_1 \cdots dx_n &= 1 \end{aligned}$$

The definition of the joint cdf is the same for discrete and continuous RV's.

**Definition 14.** *Let $X_1, X_2, \cdots X_n$ be RV's. Their joint cdf is*

$$F(x_1, x_2, \cdots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \cdots X_n \leq x_n) \tag{10}$$

Next we review what it means for RV's to be independent.

**Theorem 6.** *Let $Y_1, Y_2, \cdots Y_n$ be RV's (discrete or continuous). The following statements are equivalent, i.e., if one of them is true then all of them are true.*
*(a) If they are discrete*

$$p(y_1, y_2, \cdots, y_n) = p_{Y_1}(y_1), p_{Y_2}(y_2), \cdots, p_{Y_n}(y_n) \tag{11}$$

*If they are continuous*

$$f(y_1, y_2, \cdots, y_n) = f_{Y_1}(y_1) f_{Y_2}(y_2) \cdots f_{Y_n}(y_n) \tag{12}$$

*(b) For all real numbers $a_i, b_i, i = 1, 2, \cdots, n$*

$$\begin{aligned} & P(a_1 \leq Y_1 \leq b_1, a_2 \leq Y_2 \leq b_2, \cdots, a_n \leq Y_n \leq b_n) && (13) \\ = \quad & P(a_1 \leq Y_1 \leq b_1) P(a_2 \leq Y_2 \leq b_2) \cdots P(a_n \leq Y_n \leq b_n) && (14) \end{aligned}$$

*(c) The cdf's satisfy*

$$F(y_1, y_2, \cdots, y_n) = F_{Y_1}(y_1) F_{Y_2}(y_2) \cdots F_{Y_n}(y_n) \tag{15}$$

**Definition 15.** *If one (and hence all) of (a), (b), (c) hold, then we say $Y_1, Y_2, \cdots Y_n$ are independent.*

Usually (a) is taken to be the definition of independent and then (b) and (c) are proved to follow from idependence. Next we review some important consequences of independence.

**Theorem 7.** *If $X$ and $Y$ are independent RV's, then $E[XY] = E[X]E[Y]$.*

**Theorem 8.** *If $Y_1, Y_2, \cdots, Y_n$ are independent RV's and $g_1(y), g_2(y), \cdots g_n(y)$ are functions from $\mathbb{R}$ to $\mathbb{R}$, then $g_1(Y_1), g_2(Y_2), \cdots, g_n(Y_n)$ are independent RV's.*

**Theorem 9.** *If $Y_1, Y_2, \cdots, Y_n$ are independent then*

$$Var(Y_1 + Y_2 + \cdots Y_n) = var(Y_1) + var(Y_2) + \cdots + var(Y_n) \tag{16}$$

**END OF PROBABILITY REVIEW - for now**

We now return to the notion of a random sample.

**Definition 16.** *RV's $Y_1, Y_2, \cdots, Y_n$ are a random sample if they are independent and identically distributed (i.i.d.). A statistic is a function of the random sample. So a statistic is a random variable. The distribution of such a RV is called the sampling distribution of the statistic.*

Since the $Y_i$ are identically distibuted, they have the same pdf or pmf which we will denote by $f_Y(y)$ or $p_Y(y)$. The mean of this distribution is called the population mean and denoted by $\mu$. The variance of this distribution is called the population variance and denoted by $\sigma^2$.
**Remark:** The function that defines the statistic should not contain any unknown parameters, e.g., $\mu$ or $\sigma$. For example,

$$V = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \mu)^2 \tag{17}$$

might be used to estimate the variance but since it contains $\mu$ it is not a valid statistic, unless we know $\mu$. (It would be unusual to known $\mu$ if we do not know $\sigma^2$.)

**Definition 17.** *The sample mean of a random sample $Y_1, Y_2, \cdots, Y_n$ is the statistic defined by*

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{18}$$

*Sometimes we will denote the sample mean by $\overline{Y}_n$ to emphasize that it depends on $n$.*

The sampling distrbution of a statistic depends on $f_Y(y)$ and the function that defines the statistic. If we know these, then in principle we can compute the sampling distribution. In practice we usually cannot do this explicitly. For the sample mean the sampling distribution may not be explicitly computable, but the mean and variance always are.

**Proposition 3.** *The mean of the sample mean $\overline{Y}_n$ is $\mu$ and its variance is $\sigma^2/n$.*

There are a few cases where we can explicitly find the sampling distribution. Here is one.

**Binomial sampling distribution** Suppose that the experiment only has two outcomes which we will call success and failure. We let $p$ be the probability of success. For example, we could take the population to be all US citizens who are eligible to vote and we could look at whether or not they voted in the last election. If we define success to be that the person voted, then $p$ is the probability that a person chosen at random voted. We let $Y$ be 1 when we have success and 0 for failure. Note that the mean of $Y$ is just

$$E[Y] = 0\,P(Y=0) + 1\,P(Y=1) = p \tag{19}$$

So the population mean $\mu$ is $p$. A little computation shows the population variance $\sigma^2$ is $p(1-p)$. Our random sample $Y_1, Y_2, \cdots, Y_n$ is a string of 0's and 1's. Let

$$X = \sum_{i=1}^{n} Y_i \tag{20}$$

Then $X$ is the number of success in the random sample. It has the binomial distribution. So

$$\mathbf{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The sample mean $\overline{Y}$ is just $X/n$. So its pmf is given by

$$P(\overline{Y} = k/n) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{21}$$

We already know from a previous theorem that the mean of $\overline{Y}$ is $\mu = p$ and its variance is $\sigma^2/n = p(1-p)/n$. These facts can also be seen by using the mean and variance of the binomial distribution. The sample mean is an estimator for $p$, and it is sometimes denoted by $\hat{p}$ rather than $\overline{Y}$.

## 7.2 Sampling distributions for a normal population

If our random sample comes from a normal population, then we can find the sampling distribution of the sample mean and the sample variance explicitly. We first review some probability.

## PROBABILITY REVIEW

We start by reviewing moment generating functions of RV's. The definition is the same for discrete and continuous RV's.

**Definition 18.** *The moment generating function of a random variance $X$ is*

$$M_X(t) = E[e^{tX}] \tag{22}$$

As the name suggests, you can use this function to compute the moments of $X$. The $n$th moment of $X$ is $E[X^n]$.

**Proposition 4.**

$$E[X^n] = M_X^{(n)}(0) \tag{23}$$

*where $M_X^{(n)}(t)$ is the $n$th derivative of $M_x(t)$ with repect to $t$.*

The mgf of $aX + b$ is related to the mgf of $X$ in a simple way.

**Proposition 5.** *Let $a, b$ be real numbers. Let $X$ be RV and let $Y = aX + b$. Then*

$$M_Y(t) = e^{bt} M_X(at) \tag{24}$$

**Example - normal** For the normal distribution with mean $\mu$ and variance $\sigma^2$, the mgf is

$$M(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2) \tag{25}$$

Derivation: Derive this just for the standard normal in class.

Suppose $Y$ is normal with mean $\mu$ and variance $\sigma^2$. Let $Z = (Y - \mu)/\sigma$. Then the above proposition implies that $Z$ is a standard normal.

---

**End of lecture on Tues, 1/16**

---

**Proposition 6.** *If $Y_1, Y_2, \cdots, Y_n$ are independent RV's and each one has a normal distribution, then for any real numbers $c_1, c_2, \cdots, c_n$, the RV*

$$Y = \sum_{i=1}^{n} c_i Y_i \tag{26}$$

*has a normal distribution. If we let $\mu_i$ and $\sigma_i^2$ denote the mean and variance of $Y_i$, then the mean and variance of $Y$ are given by*

$$\mu_Y = \sum_{i=1}^{n} c_i \mu_i \quad \sigma_Y^2 = \sum_{i=1}^{n} c_i^2 \sigma_i^2 \tag{27}$$

**Proof:** Done in class.

**END OF PROBABILITY REVIEW - for now**

We now return to sampling distributions.

**Theorem 10.** *Suppose the population is normal with mean $\mu$ and variance $\sigma^2$ and we draw a random sample $Y_1, Y_2, \cdots, Y_n$. (This means that the $Y_i$ are i.i.d. and their common distribution is normal with mean $\mu$ and variance $\sigma^2$.) Then the sampling distribution of the sample mean*

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{28}$$

*is normal with mean $\mu$ and variance $\sigma^2/n$.*

**Proof:** This is immediate from the previous proposition. □

**A little R** The book has tables of values for various distributions. This is an anachronism. Software will give you these values. For example, R has functions that compute the following. Let $Z$ have a standard normal distribution.

$$pnorm(t) = P(Z \leq t) \tag{29}$$

*qnorm* is the inverse of this function. So

$$P(Z \leq t) = p \Rightarrow qnorm(p) = t \tag{30}$$

With just one argument these functions are for the standard normal. For a normal with mean $\mu$ and variance $\sigma^2$ the functions are $pnorm(t, \mu, \sigma)$, $qnorm(p, \mu, \sigma)$. NOTE that the third argument is the standard deviation, not the variance. There are analagous functions for all the common probability distribution.

**Example:** A grande coffee at starbucks is supposed to contain 16 oz. Suppose that the actually amount of coffee you get when you buy a grande is normally distributed with $\mu = 16$ and $\sigma = 0.5$.
(a) In the course of a week I buy 10 grandes and measure the amount of coffee in each. Let $\overline{Y}$ be the average of these 10 numbers. What is the probability that $\overline{Y}$ differs from 16 by more than 0.1?
(b) How many cups of coffee do I need to buy if I want the probability of $\overline{Y}$ being within 0.1 of $\mu$ to be at least 0.95?

For part (a) we know that $\overline{Y}$ is normal with mean 16 and variance $(0.5)^2/10 = 0.025$. So the standard deviation is 0.1581. We want

$$P(|\overline{Y} - 16| \geq 0.5) = 2P(\overline{Y} - 16 \geq 0.5) \tag{31}$$

R says that $pnorm(16.1, 16, 0.1581) = 0.7365$. So $P(\overline{Y} - 16 \geq 0.5) = 1 - 0.7365 = 0.2635$. So the answer is 0.5271.

For part $b$, we use R to see $qnorm(0.975) = 1.9600$. We find $n = (5 \times 1.96)^2 \approx 97$.

When the population is normal we can also find the distribution of the sample variance explicitly. First we introduce some important pdf's.

**The gamma distribution:** A RV $Y$ has a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$ if the pdf is

$$f(y) = \frac{y^{\alpha-1} e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)} \tag{32}$$

for $y \geq 0$. The range of a gamma RV is $[0, \infty)$, so $f(y) = 0$ for $y < 0$.

The gamma function is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\, e^{-x}\, dx \tag{33}$$

Integration by parts shows that $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. It then follows by induction that for positive integers $n$, $\Gamma(n + 1) = n!$.

**Caution:** There are different conventions for the parameters for the Gamma distribution. Sometimes $\lambda = 1/\beta$ is used in place of $\beta$. I am following the convention in the textbook.

Some calculus shows that the mean and variance of this distribution are

$$\mu = \alpha\beta, \quad \sigma^2 = \alpha\beta^2 \tag{34}$$

and the mgf is

$$M(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha} \tag{35}$$

We will derive the formula for $M(t)$. The computation of the mean and variance then just requires some differentiating which we will skip.
Derivation of $M(t)$

**Remark:** Consider the effect of the parameter $\beta$ on the gamma distribution. If $Y$ has a gamma distribution with parameters $\alpha, \beta$, then $Y/\beta$ has a gamma distribution with parameters $\alpha, 1$. Changing $\beta$ does not change the shape of the pdf, it just rescales it. By contrast, if we change $\alpha$ the shape of the pdf changes.

**Definition 19.** *(***The $\chi^2$ distribution***) A RV $Y$ has a $\chi^2$ distribution with $\nu$ degrees of freedom (read this as "chi-square distribution") if it has a gamma distribution with $\alpha = \nu/2$ and $\beta = 2$. So the pdf is*

$$f(y) = \frac{y^{\nu/2-1}e^{-y/2}}{2^{\nu/2}\Gamma(\nu/2)} \tag{36}$$

*The mean and variance are*

$$\mu = \nu, \quad \sigma^2 = 2\nu \tag{37}$$

*and the mgf is*

$$M(t) = \left(\frac{1}{1-2t}\right)^{\nu/2} \tag{38}$$

**Proposition 7.** *If $Z_1, Z_2, \cdots, Z_n$ are independent and each has a standard normal distribution, then*

$$W = \sum_{i=1}^{n} Z_i^2 \tag{39}$$

*has a $\chi^2$ distribution with $n$ degrees of freedom.*

We will prove this proposition a bit later.

**Definition 20.** *The sample variance of a random sample $Y_1, Y_2, \cdots, Y_n$ is defined to be*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{40}$$

*Sometimes we will denote the sample variance by $S_n^2$ to emphasize that it depends on $n$.*

When the population normal, the distribution of the sample variance can be found explicitly. Somewhat surprisingly the sample mean and sample variance are independent random variables.

**Theorem 11.** *Suppose the population is normal with mean $\mu$ and variance $\sigma^2$ and we draw a random sample $Y_1, Y_2, \cdots, Y_n$. (This means that the $Y_i$ are i.i.d. and their common distribution is normal with mean $\mu$ and variance $\sigma^2$.) Then*

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{41}$$

*has a $\chi^2$ distribution with $n-1$ degrees of freedom. Moreover, $S^2$ and $\overline{Y}$ are independent random variables.*

We now prove proposition 7. The first step is :

**Proposition 8.** *Let $Z$ have a standard normal distribution. Then $Z^2$ has a $\chi^2$ distribution with one degree of freedom, i.e., it has a gamma distribution with $\alpha = 1/2$ and $\beta = 2$.*

**Proof:** Done in class. The idea is to first computer the CDF of $Z^2$ in terms of the CDF of $Z$. Then we can differentiate to get the pdf of $Z^2$.

**Proof of proposition 7:** Done in class. The idea is to use the previous proposition and mgf's.

---

**End of lecture on Thurs, 1/18**

---

**Example:** We continue the coffee example. Suppose that the actual amount of coffee you get when you buy a grande coffee is normally distributed with $\mu = 16$ and $\sigma = 0.5$. I buy 10 cups of coffee and compute the sample variance $S^2$.
(a) Find the probability that the sample variance is greater than $(0.5)^2$. Answer: The RV $(n-1)S^2/\sigma^2$ has a $\chi^2$ distrbution with $n-1 = 9$ degrees of freedom. So

$$P(S^2 \geq (0.5)^2) = P(9S^2/(0.5)^2 \geq 9) = P(\chi^2 \geq 9) = 1 - pchisq(9,9) = 0.4372$$

where $\chi^2$ has a $\chi^2$ distribution with 9 d.f., and we have used R to get the final number.
(b) Find the probabiilty that the sample variance is greater than $(0.75)^2$.

$$\begin{aligned} P(S^2 \geq (0.75)^2) &= P(9S^2/(0.5)^2 \geq 9(1.5)^2) = P(\chi^2 \geq 9(1.5)^2) \\ &= 1 - pchisq(9 * (1.5)^2, 9) = 0.0164 \end{aligned}$$

(c) Find numbers $b_1$ and $b_2$ such that

$$P(b_1 \leq S^2 \leq b_2) = 0.9$$

Answer: the question has many answers. The probabity that $S^2$ is outside $[b_1, b_2]$ should be 0.1, but the question does not say how much of this should be less that $b_1$ and how much greater than $b_2$. We look for a solution with

$$P(S^2 < b_1) = 0.05, \quad P(S^2 > b_2) = 0.05$$

So

$$P(9S^2/(0.5)^2 < 9b_1/(0.5)^2) = 0.05, \quad P(9S^2/(0.5)^2 > 9b_2/(0.5)^2) = 0.05$$

Using R we find $qchisq(0.95, 9) = 16.92, qchisq(0.05, 9) = 3.325$. So

$$9b_1/(0.5)^2 = 3.325, \quad 9b_2/(0.5)^2 = 16.92$$

When we do confidence intervals and hypothesis testing we will look at

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \tag{42}$$

If the population is normal, then $Z$ will have a standard normal distribution. (Note that we have the very nice feature that the distribution of $Z$ does not depend on the sample size.) If we want to test whether our data agree with a particular hypothesized value of $\mu$, then we can look at the value we get for $Z$ and see if it is reasonable for the standard normal. But to do this we need to know $\sigma$.

If we don't know $\sigma$, then we can try to approximate it by $\sqrt{S^2}$. We will denote $\sqrt{S^2}$ by $S$. So we could look at

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \tag{43}$$

If the population is normal, $T$ will not have a normal distribution but it will have a distribution that only depends on the sample size, not on $\mu$ or $\sigma$. We first define that distribution.

**Definition 21.** *Let $Z$ and $W$ be independent RV's and suppose $Z$ has a standard normal distribution and $W$ has a $\chi^2$ distribution with $\nu$ degrees of freedom. Define a new RV by*

$$T = \frac{Z}{\sqrt{W/\nu}} \tag{44}$$

*Then the distribution of $T$ is called the student's t-distribution (or just the t-distribution) with $\nu$ degrees of freedom.*

**Proposition 9.** *The pdf of the t-distribution is given by*

$$f(t) = c(\nu) \left[1 + \frac{t^2}{\nu}\right]^{-(\nu+1)/2} \tag{45}$$

*$c(\nu)$ is the constant that makes the integral of this function equal to 1. There is an explicit formula for $c(\nu)$ in terms of the gamma function. See problem 7.98 in the text for more details.*

A derivation of this formula can be found in the text problem above. Note that this pdf is an even function, i.e., the distribution is symmetric in the sense that $T$ and $-T$ are identically distributed. In particular the mean is zero. The variance can be computed and is given by $\nu/(\nu-2)$ for $\nu > 2$. For $\nu \leq 2$ the variance is infinite.

Recall that

$$\lim_{n\to\infty} (1 + x/n)^n = e^x \tag{46}$$

It follows from this that the limit as $\nu \to \infty$ of the pdf for the t-distribution converges to the standard normal pdf. You can also see this from the CLT.

**Plot** t-distribution and standard normal.

Having defined the t-distribution we now show how it comes up in samples from a normal population. We have already seen that the distribution of the sample mean is normal and the distribution of the sample variance is related to the $\chi^2$ distribution. And we saw that the sample mean and sample variance are idependent.

**Theorem 12.** *Suppose the population is normal with mean $\mu$ and variance $\sigma^2$ and we draw a random sample $Y_1, Y_2, \cdots, Y_n$. Then*

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \tag{47}$$

*has a t-distribution with $n-1$ degrees of freedom. Here $S$ denotes $\sqrt{S^2}$ where $S^2$ is the sample variance.*

**Proof:** Define

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \tag{48}$$

$$W = \frac{(n-1)S^2}{\sigma^2} \tag{49}$$

Then since $\overline{Y}$ and $S^2$ are independent, $Z$ and $W$ are independent. We already know that $Z$ has a standard normal distribution and $W$ has a $\chi^2$ distribution with $n-1$ degrees of freedom. So by def. of the t-distribution,

$$\frac{Z}{\sqrt{W/(n-1)}}$$

has a t-distribution with $n-1$ degrees of freedom. But this just simplifes to $T$.

**Example:** Suppose that a population has a normal distribution with mean $\mu$ and variance $\sigma^2$. We define

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}, \tag{50}$$

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \tag{51}$$

For sample sizes of 4, 10, 100, find $c$ so that $P(-c \leq Z \leq c) = 0.95$ and so that $P(-c \leq T \leq c) = 0.95$ Answer: For $Z$, $c$ is 1.960 for all sample sizes. For $T$, $c = 3.182, 2.262, 1.984$

---

**End of lecture on Tues, 1/16**

---

We introduce one last distrbution in this subsection. Suppose we have two normal populations with variances $\sigma_1^2$ and $\sigma_2^2$. We want to test if these are equal. We could draw random samples from each population and compute their sample variances $S_1^2$ and $S_2^2$. Then we look at $S_1^2/S_2^2$ and see how close it is to 1. This statistic is the ratio of two independent random variables with $\chi^2$ distributions. So we may the following definitions

**Definition 22. F-distribution** *Let $W_1$ and $W_2$ be independent RV's with $\chi^2$ distributions with $\nu_1$ and $\nu_2$ degrees of freedom. Define*

$$F = \frac{W_1/\nu_1}{W_2/\nu_2} \tag{52}$$

*Then the distribution of $F$ is called the F-distribution with $\nu_1$ numerator degrees of freedom and $\nu_2$ denominator degrees of freedom.*

**Remark:** If we have two normal populations with variances $\sigma_1^2$ and $\sigma_2^2$, and we take random samples from each one with sizes $n_1$ and $n_2$ and sample variances $S_1^2$ and $S_2^2$, then we know that $(n_i - 1)S_i^2/\sigma_i^2$ have $\chi^2$ distributions. So the following has an F-distribution:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \tag{53}$$

The F stands for Sir Ronald Fisher, one of the founders of statistics.

## 7.3 Central Limit Theorem

This subsection is a review of one of the most important topics from 464 - the central limit theorem.

If $X$ is a random variable with mean $\mu$ and variance $\sigma^2$, then

$$Z = \frac{X - \mu}{\sigma}$$

will have mean 0 and variance 1. We refer to this process of substracting off the mean and then dividing by the standard deviation as "standardizing" the random variable. The central limit theorem says that if we add a large number of independent, identically distributed random variables and then standardize the sum, then the distribution of the resulting random variable is approximately the standard normal.

**Theorem 13.** *Let $Y_1, Y_2, \cdots, Y_n$ be independent, identically distributed random variables with mean $\mu$ and variance $\sigma^2 < \infty$. Define*

$$Z_n = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}, \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{54}$$

*Then the distribution of $Z_n$ converges to a standard normal in the sense that*

$$\lim_{n\to\infty} P(a \le Z_n \le b) = P(a \le Z \le b) \tag{55}$$

*for all real $a < b$, where $Z$ has a standard normal distribution.*

**Remark:** This theorem does not say anything about how fast $P(a \le Z_n \le b)$ converges to $P(a \le Z \le b)$. Our main use of this theorem will be to approximate $P(a \le Z_n \le b)$ by $P(a \le Z \le b)$, so it is important to know $n$ is large enough for this to be valid. The rule of thumb in statistics is that $n = 30$ is usually big enough. The Berry-Esseen theorem gives a quantitative bond for how close $P(a \le Z_n \le b)$ is to $P(a \le Z \le b)$.

**Example:** Recall that if $Z_i$ are independent standard normals, then $\sum_{i=1}^n Z_i^2$ has a $\chi^2$ distribution with $n$ d.f. This is a sum of independent RV's so CLT says it should be approximately normal when $n$ is large.

**Important application:** We already know that if the population is normally distributed, then the sample mean $\overline{Y}$ will have a normal distribution even if the number of samples is not large. The central limit theorem implies that even when the population is not normally distributed, for large sample sizes the distribution of the sample mean will be approximately normal.

**The continuity approximation or the one-half game** Suppose that $Y_1, \cdots, Y_n$ is a random sample from a population in which the random variable only takes on integer values. So $S_n = \sum_{i=1}^n Y_i$ only takes on integer values. Suppose we want to use the CLT to approximate $P(S_n = k)$ for a particular $k$. Note that the mean of $S_n$ is $n\mu$ and its variance is $n\sigma^2$. So we standardize it by letting

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\mu} \tag{56}$$

We could try

$$\begin{aligned} P(S_n = k) &= P(k \leq S_n \leq k) = P(\frac{k - n\mu}{\sqrt{n}\sigma} \leq Z_n \leq \frac{k - n\mu}{\sqrt{n}\sigma}) && (57) \\ &\approx P(\frac{k - n\mu}{\sqrt{n}\sigma} \leq Z \leq \frac{k - n\mu}{\sqrt{n}\sigma}) && (58) \end{aligned}$$

But $Z$ is a continuous RV, so this last probability is zero.

The problem is that for the discrete RV $S_n$, the probabiity is concentrated on the integers, while for a continuous random variable it is spread over all real numbers. We get a better approximation is we think of the event $S_n = k$ as the event $k - 1/2 \leq S_n \leq k + 1/2$. Then we get

$$\begin{aligned} P(S_n = k) &= P(k - 1/2 \leq S_n \leq k + 1/2) && (59) \\ &= P(\frac{k - 1/2 - n\mu}{\sqrt{n}\sigma} \leq Z_n \leq \frac{k + 1/2 - n\mu}{\sqrt{n}\sigma}) && (60) \\ &\approx P(\frac{k - 1/2 - n\mu}{\sqrt{n}\sigma} \leq Z \leq \frac{k + 1/2 - n\mu}{\sqrt{n}\sigma}) && (61) \end{aligned}$$

which is not zero. More generally, if we want to compute $P(k \leq S_n \leq l)$ where $k$ and $l$ are integers, then we should think of it as $P(k - 1/2 \leq S_n \leq l + 1/2)$ where $k$ and

**Example:** Suppose $X$ has a binomial distribution with $p = 0.3$ and $n = 20$. Note that $E[X] = 6$. Compute $P(5 \leq X \leq 8)$ exactly and approximately with the CLT. Answers: 0.4851 from normal, 0.4703 from exact computation.