Lecture 18: Sampling distributions

In many applications, the population is one or several normal distributions (or approximately).

We now study properties of some important statistics based on a random sample from a normal distribution.

If $X_1,...,X_n$ is a random sample from $N(\mu,\sigma^2)$, then the joint pdf is

$$\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right), \qquad x_i\in\mathscr{R}, i=1,...,n$$

Theorem 5.3.1.

Let $X_1,...,X_n$ be a random sample from $N(\mu,\sigma^2)$ and let $\bar{X}$ and S^2 be the sample mean and sample variance. Then
a. $\bar{X}$ and S^2 are independent random variables;
b. $\bar{X}\sim N(\mu,\sigma^2/n)$;
c. $(n-1)S^2/\sigma^2$ has the chi-square distribution with $n-1$ degrees of freedom.

Proof.

We have already established property b (Chapter 4).
To prove property a, it is enough to show the independence of $\bar{Z}$ and S_Z^2, the sample mean and variance based on $Z_i = (X_i - \mu)/\sigma \sim N(0,1)$, $i = 1,...,n$, because we can apply Theorem 4.6.12 and

$$\bar{X} = \sigma \bar{Z} - \mu \quad \text{and} \quad S^2 = \frac{\sigma^2}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})^2 = \sigma^2 S_Z^2$$

Consider the transformation

$$Y_1 = \bar{Z}, \qquad Y_i = Z_i - \bar{Z}, \quad i = 2,...,n,$$

Then

$$Z_1 = Y_1 - (Y_2 + \cdots + Y_n), \quad Z_i = Y_i + Y_1, \quad i = 2,...,n,$$

and

$$\left| \frac{\partial(Z_1,...,Z_n)}{\partial(Y_1,...,Y_n)} \right| = \frac{1}{n}$$

Since the joint pdf of $Z_1,...,Z_n$ is

$$\frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} z_i^2 \right) \qquad z_i \in \mathscr{R}, i = 1,...,n,$$

the joint pdf of $(Y_1,...,Y_n)$ is

$$\frac{n}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\left(y_1-\sum_{i=2}^n y_i\right)^2\right)\exp\left(-\frac{1}{2}\sum_{i=2}^n (y_i+y_1)^2\right)$$

$$= \frac{n}{(2\pi)^{n/2}}\exp\left(-\frac{n}{2}y_1^2\right)\exp\left(-\frac{1}{2}\left[\sum_{i=2}^n y_i^2+\left(\sum_{i=2}^n y_i\right)^2\right]\right) \qquad \begin{matrix} y_i\in\mathscr{R} \\ i=1,...,n. \end{matrix}$$

Since the first exp factor involves y_1 only and the second exp factor involves $y_2,...,y_n$, we conclude (Theorem 4.6.11) that Y_1 is independent of $(Y_2,...,Y_n)$.

Since

$$Z_1-\bar{Z}=-\sum_{i=2}^n (Z_i-\bar{Z})=-\sum_{i=2}^n Y_i \quad \text{and} \quad Z_i-\bar{Z}=Y_i, \quad i=2,...,n,$$

we have

$$S_Z^2=\frac{1}{n-1}\sum_{i=1}^n (Z_i-\bar{Z})^2=\frac{1}{n-1}\left(\sum_{i=2}^n Y_i\right)^2+\frac{1}{n-1}\sum_{i=2}^n Y_i^2$$

which is a function of $(Y_2,...,Y_n)$.
Hence, $\bar{Z}$ and S_Z^2 are independent by Theorem 4.6.12.
This proves a.

Finally, we prove c (the proof in the textbook can be simplified).
Note that

$$(n-1)S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu + \mu - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu)^2 + n(\mu - \bar{X})^2$$

Then

$$n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2 + \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^n Z_i^2$$

Since $Z_i \sim N(0,1)$ and $Z_1,...,Z_n$ are independent, we have previously shown that

- each $Z_i^2 \sim$ chi-square with degree of freedom 1,
- the sum $\sum_{i=1}^n Z_i^2 \sim$ chi-square with degrees of freedom n, and its mgf is $(1-2t)^{-n/2}$, $t < 1/2$,
- $\sqrt{n}(\bar{X}-\mu)/\sigma \sim N(0,1)$ and hence $n[(\bar{X}-\mu)/\sigma]^2 \sim$ chi-square with degree of freedom 1.

The left hand side of the previous expression is a sum of two independent random variables and, hence, if $f(t)$ is the mgf of $(n-1)S^2/\sigma^2$, then the mgf of the sum on the left hand side is

$$(1-2t)^{-1/2}f(t)$$

Since the right hand side of the previous expression has mgf $(1-2t)^{-n/2}$, we must have

$$f(t) = (1-2t)^{-n/2}/(1-2t)^{-1/2} = (1-2t)^{-(n-1)/2} \qquad t < 1/2$$

This is the mgf of the chi-square with degrees of freedom $n-1$, and the result follows.

The independence of $\bar{X}$ and S^2 can be established in other ways.

t-distribution

Let $X_1, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$.

Using the result in Chapter 4 about a ratio of independent normal and chi-square random variables, the ratio

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{[(n-1)S^2/\sigma^2]/(n-1)}}$$

has the central t-distribution with $n-1$ degrees of freedom.

What is the distribution of $T_0 = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ for a fixed known constant $\mu_0 \in \mathscr{R}$ which is not necessarily equal to μ?

Note that T is not a statistic while T_0 is a statistic.

Since $\bar{X} - \mu_0 \sim N(\mu - \mu_0, \sigma^2/n)$, from the discussion in Chapter 4 we know that the distribution of T_0 is the noncentral t-distribution with degrees of freedom $n-1$ and noncentrality parameter $\delta = \sqrt{n}(\mu - \mu_0)/\sigma$.

F-distribution

Let $X_1, ..., X_n$ be a random sample from $N(\mu_x, \sigma_x^2)$, $Y_1, ..., Y_m$ be a random sample from $N(\mu_y, \sigma_y^2)$, X_i's and Y_i's be independent, and S_x^2 and S_y^2 be the sample variances based on X_i's and Y_i's, respectively.

From the previous discussion, $(n-1)S_x^2/\sigma_x^2$ and $(m-1)S_y^2/\sigma_y^2$ are both chi-square distributed, and the ratio $\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2}$ has the F-distribution with degrees of freedom $n-1$ and $m-1$ (denoted by $F_{n-1,m-1}$).

Theorem 5.3.8.

Let $F_{p,q}$ denote the F-distribution with degrees of freedom p and q.
a. If $X \sim F_{p,q}$, then $1/X \sim F_{q,p}$.
b. If X has the t-distribution with degrees of freedom q, then $X^2 \sim F_{1,q}$.
c. If $X \sim F_{p.q}$, then $(p/q)X/[1+(p/q)X] \sim beta(p/2, q/2)$.

Proof.

We only need to prove c, since properties a and b follow directly from the definitions of F- and t-distributions.

Note that $Z = (p/q)X$ has pdf

$$\frac{\Gamma[(p+q)/2]}{\Gamma(p/2)\Gamma(q/2)} \frac{z^{p/2-1}}{(1+z)^{(p+q)/2}}, \qquad z > 0$$

If $u = z/(1+z)$, then $z = u/(1-u)$, $dz = (1-u)^{-2}du$, and the pdf of $U = Z/(1+Z)$ is

$$\frac{\Gamma[(p+q)/2]}{\Gamma(p/2)\Gamma(q/2)} \left(\frac{u}{1-u}\right)^{p/2-1} \frac{1}{(1-u)^{-(p+q)/2}} \frac{1}{(1-u)^2}$$

$$= \frac{\Gamma[(p+q)/2]}{\Gamma(p/2)\Gamma(q/2)} u^{p/2-1}(1-u)^{q/2-1} \qquad u > 0$$

Definition 5.4.1 (Order statistics).

The order statistics of a random sample of univariate $X_1,...,X_n$ are the sample values placed in a non-decreasing order, and they are denoted by $X_{(1)},...,X_{(n)}$.

Once $X_{(1)},...,X_{(n)}$ are given, the information left in the sample is the particular positions from which $X_{(i)}$ is observed, $i = 1,...,n$.

Functions of order statistics

Many useful statistics are functions of order statistics.

- Both sample mean and variance are functions of order statistics, because
$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_{(i)} \quad \text{and} \quad \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_{(i)}^2$$
- The **sample range** $R = X_{(n)} - X_{(1)}$, the distance between the smallest and largest observations, is a measure of the dispersion in the sample and should reflect the dispersion in the population.

- For any fixed $p \in (0,1)$, the $(100p)$th **sample percentile** is the observation such that about np of the observations are less than this observation and $n(1-p)$ of the observations are greater:

$$\begin{array}{ll} X_{(1)} & \text{if } p \le (2n)^{-1} \\ X_{(\{np\})} & \text{if } (2n)^{-1} < p < 0.5 \\ X_{((n+1)/2)} & \text{if } p = 0.5 \text{ and } n \text{ is odd} \\ (X_{(n/2)} + X_{(n/2+1)})/2 & \text{if } p = 0.5 \text{ and } n \text{ is even} \\ X_{(n+1-\{n(1-p)\})} & \text{if } 0.5 < p < 1-(2n)^{-1} \\ X_{(n)} & \text{if } p \ge 1-(2n)^{-1} \end{array}$$

  where $\{b\}$ is the number b rounded to the nearest integer, i.e., if k is an integer and $k - 0.5 \le b < k + 0.5$, then $\{b\} = k$.

  Other textbooks may define sample percentiles differently.
- The **sample median** is the 50th sample percentile.
  It is a measure of location, alternative to the sample mean.
- The **sample lower quartile** is the 25th sample percentile and the **upper quartile** is the 75th sample percentile.
- The **sample mid-range** is defined as $V = (X_{(1)} + X_{(n)})/2$.

If $X_1, ..., X_n$ is a random sample of discrete random variables, then the calculation of probabilities for the order statistics is mainly a counting task.

Theorem 5.4.3.

Let $X_1, ..., X_n$ be a random sample from a discrete distribution with pmf $f(x_i) = p_i$, where $x_1 < x_2 < \cdots$ are the possible values of X_1. Define

$$P_0 = 0,\ P_1 = p_1,\ ...,\ P_i = p_1 + \cdots + p_i,\ ...$$

Then, for the jth order statistic $X_{(j)}$,

$$P(X_{(j)} \le x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}$$

$$P(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{k} [P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}]$$

Proof.

For any fixed i, let Y be the number of $X_1, ..., X_n$ that are less than or equal to x_i.

If the event $\{X_j \leq x_i\}$ is a "success", then Y is the number of successes in n trials and is distributed as $binomial(n, P_i)$.
Then, the result follows from $\{X_{(j)} \leq x_i\} = \{Y \geq j\}$,

$$P(X_{(j)} \leq x_i) = P(Y \geq j) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1-P_i)^{n-k}$$

and $P(X_{(j)} = x_i) = P(X_{(j)} \leq x_i) - P(X_{(j)} \leq x_{i-1})$.

If $X_1, ..., X_n$ is a random sample from a continuous population with pdf $f(x)$, then

$$P(X_{(1)} < X_{(2)} < \cdots < X_{(n)}) = 1$$

i.e., we do not need to worry about ties, and the joint pdf of $(X_{(1)}, ..., X_{(n)})$ is

$$h(x_1, ..., x_n) = \begin{cases} n! f(x_1) \cdots f(x_n) & x_1 < x_2 < \cdots < x_n \\ 0 & \text{otherwise} \end{cases}$$

The n! naturally comes into this formula because, for any set of values $x_1, ..., x_n$, there are n! equally likely assignments for these values to $X_1, ..., X_n$ that all yield the same values for the order statistics.

Theorem 5.4.4.

Let $X_{(1)},...,X_{(n)}$ be the order statistics of a random sample $X_1,...,X_n$ from a continuous population with cdf F and pdf f.
Then the pdf of $X_{(j)}$ is

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!}[F(x)]^{j-1}[1-F(x)]^{n-j}f(x) \qquad x \in \mathscr{R}$$

Proof.

Let Y be the number of $X_1,...,X_n$ less than or equal to x.
Then, similar to the proof of Theorem 5.4.3, $Y \sim binomial(n, F(x))$, $\{X_{(j)} \le x\} = \{Y \ge j\}$ and

$$F_{X_{(j)}}(x) = P(X_{(j)} \le x) = P(Y \ge j) = \sum_{k=j}^{n} \binom{n}{k} [F(x)]^k [1-F(x)]^{n-k}$$

We now obtain the pdf of $X_{(j)}$ by differentiating the cdf $F_{X_{(j)}}$:

$$f_{X_{(j)}}(x) = \frac{d}{dx} F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} \frac{d}{dx} [F(x)]^k [1-F(x)]^{n-k}$$

$$= \sum_{k=j}^{n} \binom{n}{k} \{ k[F(x)]^{k-1}[1-F(x)]^{n-k} - (n-k)[F(x)]^{k}[1-F(x)]^{n-k-1} \} f(x)$$

$$= \binom{n}{j} j[F(x)]^{j-1}[1-F(x)]^{n-j} f(x) + \sum_{l=j+1}^{n} \binom{n}{l} l[F(x)]^{l-1}[1-F(x)]^{n-l} f(x)$$

$$- \sum_{k=j}^{n-1} \binom{n}{k} (n-k)[F(x)]^{k}[1-F(x)]^{n-k-1} f(x)$$

$$= \frac{n!}{(j-1)!(n-j)!} [F(x)]^{j-1}[1-F(x)]^{n-j} f(x)$$

$$+ \sum_{k=j}^{n-1} \binom{n}{k+1} (k+1)[F(x)]^{k}[1-F(x)]^{n-k-1} f(x)$$

$$- \sum_{k=j}^{n-1} \binom{n}{k} (n-k)[F(x)]^{k}[1-F(x)]^{n-k-1} f(x)$$

The result follows from the fact that the last two terms cancel, because

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

Example 5.4.5.

Let $X_1,...,X_n$ be a random sample from *uniform*$(0,1)$ so that $f(x)=1$ and $F(x)=x$ for $x\in[0,1]$.
By Theorem 5.4.4, the pdf of $X_{(j)}$ is

$$\frac{n!}{(j-1)!(n-j)!}x^{j-1}(1-x)^{n-j}=\frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)}x^{j-1}(1-x)^{n-j+1-1}\qquad 0<x<1$$

which is the pdf of *beta*$(j,n-j+1)$.

Theorem 5.4.6.

Let $X_{(1)},...,X_{(n)}$ be the order statistics of a random sample $X_1,...,X_n$ from a continuous population with cdf F and pdf f.
Then the joint pdf of $X_{(i)}$ and $X_{(j)}$, $1\le i<j\le n$, is

$$\begin{aligned} f_{X_{(i)},X_{(j)}}(x,y) &= \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(x)]^{i-1}[F(y)-F(x)]^{j-i-1} \\ &\quad \times[1-F(y)]^{n-j}f(x)f(y) \qquad x<y,\ (x,y)\in\mathscr{R}^2 \end{aligned}$$

The proof is left to Exercise 5.26.

Example 5.4.7.

Let $X_1, ..., X_n$ be a random sample from *uniform*$(0, a)$, $R = X_{(n)} - X_{(1)}$ be the range, and $V = (X_{(1)} + X_{(n)})/2$ be the midrange.

We want to obtain the joint pdf of R and V as well as the marginal distributions of R and V.

By Theorem 5.4.6, the joint pdf of $Z = X_{(1)}$ and $Y = X_{(n)}$ is

$$f_{Z,Y}(z,y) = \frac{n(n-1)}{a^2}\left(\frac{y}{a} - \frac{z}{a}\right)^{n-2} = \frac{n(n-1)(y-z)^{n-2}}{a^n}, \quad 0 < z < y < a$$

Since $R = Y - Z$ and $V = (Y + Z)/2$, we obtain $Z = V - R/2$ and $Y = V + R/2$,

$$\left|\frac{\partial(Z,Y)}{\partial(R,V)}\right| = \begin{vmatrix} -\frac{1}{2} & 1 \\ \frac{1}{2} & 1 \end{vmatrix} = -1$$

The transformation from (Z, Y) to (R, V) maps the sets

$$\{(z,y) : 0 < z < y < a\} \to \{(r,v) : 0 < r < a, r/2 < v < a - r/2\}$$

Obviously $0 < r < a$, and for a fixed r, the smallest value of v is $r/2$ (when $z = 0$ and $y = r$) and the largest value of v is $a - r/2$ (when $z = a - r$ and $y = a$).

Thus, the joint pdf of R and V is

$$f_{R,V}(r,v) = \frac{n(n-1)r^{n-2}}{a^n}, \qquad 0 < r < a,\ r/2 < v < a - r/2$$

The marginal pdf of R is

$$f_R(r) = \int_{r/2}^{a-r/2} \frac{n(n-1)r^{n-2}}{a^n} dv = \frac{n(n-1)r^{n-2}(a-r)}{a^n}, \qquad 0 < r < a$$

The marginal pdf of V is

$$\begin{aligned} f_V(v) &= \int_0^{2v} \frac{n(n-1)r^{n-2}}{a^n} dr = \frac{n(2v)^{n-1}}{a^n} \qquad 0 < v < a/2 \\ &= \int_0^{2(a-v)} \frac{n(n-1)r^{n-2}}{a^n} dr = \frac{n(2(a-v))^{n-1}}{a^n} \qquad a/2 < v < a \end{aligned}$$

because the set where $f_{R,V}(r,v) > 0$ is

$$\begin{aligned} & \{(r,v) : 0 < r < a, r/2 < v < a - r/2\} \\ = & \{(r,v) : 0 < v \leq a/2, 0 < r < 2v\} \\ & \bigcup \{(r,v) : a/2 < v \leq a, 0 < r < 2(a-v)\} \end{aligned}$$

Example.

Let $X_1, ..., X_n$ be a random sample from *uniform*$(0,1)$.
We want to find the distribution of $X_1/X_{(1)}$.
For $s > 1$,

$$
\begin{aligned}
P\left(\frac{X_1}{X_{(1)}} > s\right) &= \sum_{i=1}^{n} P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_i\right) \\
&= \sum_{i=2}^{n} P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_i\right) \\
&= (n-1) P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_n\right) \\
&= (n-1) P(X_1 > sX_n, X_2 > X_n, ..., X_{n-1} > X_n) \\
&= (n-1) P(sX_n < 1, X_1 > sX_n, X_2 > X_n, ..., X_{n-1} > X_n) \\
&= (n-1) \int_0^{1/s} \left[\int_{sx_n}^{1} \left(\prod_{i=2}^{n-1} \int_{x_n}^{1} dx_i \right) dx_1 \right] dx_n \\
&= (n-1) \int_0^{1/s} (1-x_n)^{n-2} (1-sx_n) dx_n
\end{aligned}
$$

Thus, for $s > 1$,

$$\frac{d}{ds}P\left(\frac{X_1}{X_{(1)}} \leq s\right) = \frac{d}{ds}\left[1-(n-1)\int_0^{1/s}(1-t)^{n-2}(1-st)dt\right]$$
$$= (n-1)\int_0^{1/s}(1-t)^{n-2}tdt$$
$$= (n-1)\int_0^{1/s}(1-t)^{n-2}tdt - (n-1)\int_0^{1/s}(1-t)^{n-1}dt$$
$$= (n-1)\int_0^{1/s}(1-t)^{n-2}tdt - (n-1)\int_0^{1/s}(1-t)^{n-1}dt$$
$$= 1-\left(1-\frac{1}{s}\right)^{n-1} - \frac{n-1}{n}\left[1-\left(1-\frac{1}{s}\right)^{n-1}\right]$$

For $s \leq 1$, obviously

$$P\left(\frac{X_1}{X_{(1)}} \leq s\right) = 0 \qquad \frac{d}{ds}P\left(\frac{X_1}{X_{(1)}} \leq s\right) = 0$$