What are data?

$X_1, X_2, \cdots, X_n$

measurements of sample = data

$X_1, X_2, \cdots X_n$ = $\boxed{\text{data = iid random variables}}$

What is/are statistics?

Sample statistics are, for example, mean $\bar{X}$ or SD S.

Statistics are functions of $X_1, \cdots, X_n$. Each such function is called a statistic.

A statistic is a random variable

Sampling Distributions

In statistics, a random sample is a collection of independent and identically distributed (iid) random variables, and a sampling distribution is the distribution of a function of random sample. For example, the average and the variance formula are functions of random sample. The term "statistic" is referred to an individual function of a random sample, and the understanding of sampling distributions is the major undertaking of statistics.

Moment generating function.

Let X be a random variable. The *moment generating function* (mgf) of X is given by

$$M(t) = E\left[e^{tX}\right] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx$$

The mgf $M(t)$ is a function of t defined on some open interval (c_0, c_1) around 0 with $c_0 < 0 < c_1$. In particular when X is a continuous random variable having the pdf $f(x)$, the mgf $M(t)$ can be expressed as

$$\underbrace{M(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx.}_{\text{MGF}} \longleftrightarrow \underbrace{\mathcal{L}f(s) = \int_{0}^{\infty} e^{-sx} f(x)\,dx}_{\text{Laplace transform}}$$

The most significant property of moment generating function is that the moment generating function uniquely determines the distribution.

one-to-one

$f_X(x)$

one-to-one

$f(x)$

$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

Suppose that X is a standard normal random variable. Then we can compute the mgf of X as follows.

$-\frac{1}{2}(x^2-2tx) = -\frac{1}{2}\{\underbrace{x^2-2tx+t^2}_{(x-t)^2}-t^2\}$

$$E[e^{tX}] = M_X(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}\,e^{-\frac{x^2}{2}}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2+\frac{t^2}{2}}\,dx$$

$$= e^{\frac{t^2}{2}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{(x-t)^2}{2}}\,dx = \exp\left(\frac{t^2}{2}\right)$$

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

Suppose $Y \sim N(\mu, \sigma^2)$

density for $N(t,1)$

Since $X = (Y-\mu)/\sigma$ becomes a standard normal random variable and $Y = \sigma X + \mu$, the mgf of Y can be given by

$$M_Y(t) = e^{\mu t}M_X(\sigma t) = \exp\left(\frac{\sigma^2 t^2}{2} + \mu t\right)$$

$E[e^{tY}] = e^{\mu t}E[e^{(t\sigma)X}] = e^{\mu t} \times e^{(t\sigma)^2/2}$

for $-\infty < t < \infty$.

MGF of gamma distribution.

$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \; x \geq 0$

Suppose now that X is a gamma random variable with parameter (α, λ). Then the mgf $M_X(t)$ of X can be computed as follows.

$$
\begin{aligned}
E[e^{tX}] = M_X(t) &= \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\, dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\lambda-t)x}\, dx \\
&= \frac{\lambda^\alpha}{(\lambda-t)^\alpha} \frac{1}{\Gamma(\alpha)} \int_0^\infty (\underbrace{(\lambda-t)x}_{u})^{\alpha-1} e^{-\underbrace{(\lambda-t)x}_{u}} \underbrace{(\lambda-t)dx}_{du} \\
&= \left(\frac{\lambda}{\lambda-t}\right)^\alpha \frac{1}{\Gamma(\alpha)} \underbrace{\int_0^\infty u^{\alpha-1} e^{-u}\, du}_{\Gamma(\alpha)} = \left(\frac{\lambda}{\lambda-t}\right)^\alpha
\end{aligned}
$$

where $(\lambda - t)x$ is substituted by u, and $(\lambda - t)dx$ by du. It should be noted that this mgf $M_X(t)$ is defined on the open interval $(-\infty, \lambda)$.

Joint moment generating function.

$M(t) = E[e^{tx}]$

Let X and Y be random variables having a joint density function $f(x, y)$. Then we can define the *joint moment generating function* $M(s, t)$ of X and Y by

$$M(s,t) = E\left[e^{sX+tY}\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{sx+ty} f(x,y)\,dx\,dy.$$

If X and Y are independent, then the joint mgf $M(s, t)$ becomes

$$M(s,t) = E\left[e^{sX+tY}\right] = E\left[e^{sX}\right]\cdot E\left[e^{tY}\right] = M_X(s)\cdot M_Y(t).$$

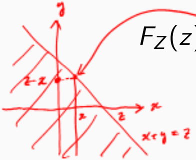
Moreover, it has been known that "if the joint mgf $M(s, t)$ of X and Y is of the form $M_1(s) \cdot M_2(t)$, then X and Y are independent and have the respective mgf $M_1(s)$ and $M_2(t)$."

Sum of independent random variables.

Let X and Y be independent random variables having the respective probability density functions $f_X(x)$ and $f_Y(y)$. Then the cumulative distribution function $F_Z(z)$ of the random variable $Z = X + Y$ can be given as follows.

$\begin{cases} Z = X+Y \\ X = x \end{cases}$

$= \iint_{\{(x,y):\, x+y\le z\}} \underset{f(x,y)}{f_X(x) f_Y(y)} \, dx\, dy$



$$F_Z(z) = P(X + Y \le z) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z-x} f_X(x) f_Y(y)\, dy \right] dx$$

$= f_X(x) \underbrace{\int_{-\infty}^{z-x} f_Y(y)\, dy}_{F_Y(z-x) = P(Y \le z-x)}$

$$= \int_{-\infty}^{\infty} f_X(x) F_Y(z - x)\, dx,$$

where F_Y is the cdf of Y. By differentiating $F_Z(z)$, we can obtain the pdf $f_Z(z)$ of Z as

$\frac{d}{dz} \int_{-\infty}^{\infty} f_X(x) F_Y(z-x)\, dz$

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} f_X(x) \left(\frac{\partial}{\partial z} F_Y(z - x) \right) dx$$

$$= \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx. = f_X * f_Y$$

Convolution.

$M_Z(t) = E[e^{tZ}] = E[e^{tX+tY}] = E[e^{tX} e^{tY}] = E[e^{tX}]\,E[e^{tY}] = M_X(t)\,M_Y(t)$

The function of z

$$(f * g)(z) = \int_{-\infty}^{\infty} f(x)g(z-x)\,dx$$

is called the *convolution* of f and g. Thus, the pdf $f_Z(z)$ for $Z = X + Y$ is given by the convolution of the pdf's $f_X(x)$ and $f_Y(y)$. Then we can easily verify

$E[e^{tZ}] = \int_{-\infty}^{\infty} e^{tz} f_Z(z)\,dz$

$$M_Z(t) = \int_{-\infty}^{\infty} e^{tz}\,\overbrace{\left[\int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\,dx\right]}^{f_Z(z)}\,dz = M_X(t)M_Y(t)$$

Therefore, as we will see in the examples below, the construction of moment generating function is much easier in order to "find the distribution of the sum of independent random variables."

$$\int_{-\infty}^{\infty} e^{tx} f_X(x) \underbrace{\left[\int_{-\infty}^{\infty} e^{t\overbrace{(z-x)}^{u}} f_Y(\underbrace{z-x}_{u}) \frac{dz}{du}\right]}_{M_Y(t)} dx = M_Y(t) \underbrace{\int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx}_{M_X(t)}$$

Example

Let X and Y be independent normal random variables with the respective parameters (μ_x, σ_x^2) and (μ_y, σ_y^2). Find the distribution for $Z = X + Y$.

Example

Let X and Y be independent normal random variables with the respective parameters (μ_x, σ_x^2) and (μ_y, σ_y^2). Find the distribution for $Z = X + Y$.

We can calculate

$$M_Z(t) = M_X(t) \cdot M_Y(t) = \exp\left(\frac{\sigma_x^2 t^2}{2} + \mu_x t\right) \cdot \exp\left(\frac{\sigma_y^2 t^2}{2} + \mu_y t\right)$$

$$= \exp\left(\frac{\overbrace{(\sigma_x^2 + \sigma_y^2)}^{\sigma^2} t^2}{2} + (\overbrace{\mu_x + \mu_y}^{\mu})t\right) = \exp\left(\frac{\sigma^2 t^2}{2} + \mu t\right) \leftarrow \text{MGF for a normal dist.}$$

which is the mgf of normal distribution with parameter $(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$. Thus, we find that Z is a normal random variable with parameter $(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

Example

Let X and Y be independent gamma random variables with the respective parameters (α_1, λ) and (α_2, λ). Find the distribution for $Z = X + Y$.

Example

Let X and Y be independent gamma random variables with the respective parameters (α_1, λ) and (α_2, λ). Find the distribution for $Z = X + Y$.

We can calculate

$$M_Z(t) = M_X(t) \cdot M_Y(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha_1} \cdot \left(\frac{\lambda}{\lambda - t}\right)^{\alpha_2} = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha_1 + \alpha_2}$$

which is the mgf of gamma distribution with parameter $(\alpha_1 + \alpha_2, \lambda)$. Thus, Z is a gamma random variable with parameter $(\alpha_1 + \alpha_2, \lambda)$.

Chi-square distribution.

Gamma with (α, λ)
$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x \geq 0$$

The gamma distribution

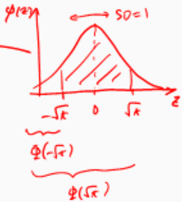
$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \quad x \geq 0$$

with parameter $(n/2, 1/2)$ is called the *chi-square distribution* with n degrees of freedom (df). Since it has only one parameter n, the mgf $M(t)$ of chi-square distribution can be simply expressed as

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

Chi-square distribution of one degree of freedom.

$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$



Let X be a standard normal random variable, and let

$$Y = X^2$$

$P(Y \le t)$ $P(|X| \le \sqrt{t})$

be the square of X. Then we can express the cdf F_Y of Y as $F_Y(t) = P(X^2 \le t) = P(-\sqrt{t} \le X \le \sqrt{t}) = \Phi(\sqrt{t}) - \Phi(-\sqrt{t})$ in terms of the standard normal cdf Φ. By differentiating the cdf, we can obtain the pdf f_Y as follows. $\frac{d}{dt}\Phi(\sqrt{t}) = \frac{1}{2} t^{-\frac{1}{2}} \phi(\sqrt{t})$; $\frac{d}{dt}\Phi(-\sqrt{t}) = -\frac{1}{2} t^{-\frac{1}{2}} \phi(-\sqrt{t})$

$$f_Y(t) = \frac{d}{dt} F_Y(t) = \frac{1}{\sqrt{2\pi}} t^{-\frac{1}{2}} e^{-\frac{t}{2}} = \frac{1}{2^{1/2}\Gamma(1/2)} t^{-1/2} e^{-t/2},$$

where we use $\Gamma(1/2) = \sqrt{\pi}$. Thus, the square X^2 of X is the chi-square random variable with 1 degree of freedom.

Chi-square distribution with n df.

Let $X_1, \ldots, X_n$ be iid standard normal random variables. Since X_i^2's have the gamma distribution with parameter $(1/2, 1/2)$, the sum

$$Y = \sum_{i=1}^{n} X_i^2 = \overbrace{X_1^2 + X_2^2}^{\text{Gamma with } (\frac{2}{2}, \frac{1}{2}) = \chi^2 \text{ with } n=2} + X_3^2 + \cdots + X_n^2$$

$\alpha = \frac{3}{2}, \lambda = \frac{1}{2} = \chi^2$ with $n=3$

χ^2 with $n=1$ = Gamma with $(\frac{1}{2}, \frac{1}{2})$

has the gamma distribution with parameter $(n/2, 1/2)$. That is, the sum Y has the chi-square distribution with n degree of freedom.

Student's t-distribution.

Let X be a chi-square random variable with n degrees of freedom, and let Y be a standard normal random variable. Suppose that X and Y are independent. Then the distribution of the quotient

$$Z = \frac{Y}{\sqrt{X/n}} = \frac{Y}{U}$$

is called the *Student's t-distribution* with n degrees of freedom. The pdf $f_Z(z)$ is given by

$$f_Z(z) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+z^2/n)^{-(n+1)/2}, \quad -\infty < z < \infty.$$

F-distribution.

Let U and V be independent chi-square random variables with r_1 and r_2 degrees of freedom, respectively. Then the distribution of the quotient

$$Z = \frac{U/r_1}{V/r_2}$$

is called the *F-distribution* with (r_1, r_2) degree of freedom. The pdf $f_Z(z)$ is given by

$$f_Z(z) = \frac{\Gamma((r_1+r_2)/2)}{\Gamma(r_1/2)\Gamma(r_2/2)}(r_1/r_2)^{r_1/2} z^{r_1/2-1}\left(1+\frac{r_1}{r_2}z\right)^{-(r_1+r_2)/2}, \quad 0 < z < \infty.$$

Quotient of two random variables.

Let X and Y be independent random variables having the respective pdf's $f_X(x)$ and $f_Y(y)$. Then the cdf $F_Z(z)$ of the quotient

$F_Y(y) = \int_{-\infty}^{y} f_Y(u)\,du$

$$Z = Y/X$$

can be computed as follows.

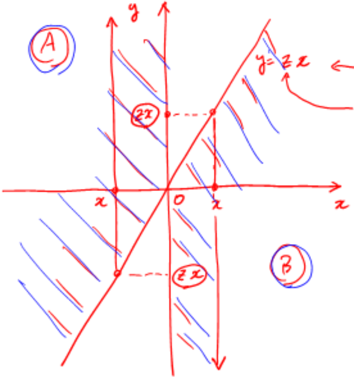
$P(X=0) = 0$

$$\begin{aligned}
F_Z(z) \overset{= P(Z \le z)}{} &= P(Y/X \le z) = P(Y \ge zX,\, X < 0) + P(Y \le zX,\, X > 0) \\
&= \int_{-\infty}^{0} \Bigg[\underbrace{\int_{xz}^{\infty} f_Y(y)\,dy}_{1 - F_Y(xz)} \Bigg] f_X(x)\,dx + \int_{0}^{\infty} \Bigg[\underbrace{\int_{-\infty}^{xz} f_Y(y)\,dy}_{F_Y(xz)} \Bigg] f_X(x)\,dx
\end{aligned}$$

By differentiating, we obtain

?

$$\begin{aligned}
f_Z(z) = \frac{d}{dz} F_Z(z) &= \int_{-\infty}^{0} [-x f_Y(xz)] f_X(x)\,dx + \int_{0}^{\infty} [x f_Y(xz)] f_X(x)\,dx \\
&= \int_{-\infty}^{\infty} |x| f_Y(xz) f_X(x)\,dx.
\end{aligned}$$



$\leftarrow$ Assume $z > 0$

slope

$$\{(x,y) : \tfrac{y}{x} \le z,\ x \ne 0\} = \{(x,y) : y \ge zx,\ x < 0\}$$

$$\cup \{(x,y) : y \le zx,\ x > 0\}$$

$$= A \cup B$$

?

$$F_Z(z) = \int_{-\infty}^{0} (1 - F_Y(xz))\, f_X(x)\,dx + \int_{0}^{\infty} F_Y(xz)\, f_X(x)\,dx$$

$$\frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} \underbrace{\left[\frac{\partial}{\partial z}(1 - F_Y(xz))\right]}_{-(x)\, f_Y(xz)} f_X(x)\,dx + \int_{0}^{\infty} \underbrace{\left[\frac{\partial}{\partial z} F_Y(xz)\right]}_{x\, f_Y(xz)} f_X(x)\,dx$$

Derivation of t-distribution.

$P(X \le nx^2) = F_X(nx^2)$

Let X be a chi-square random variable with n degrees of freedom. Then the pdf of the random variable $U = \sqrt{X/n}$ is given by

$$\frac{d}{dx}\underbrace{P(U \le x)}_{F_U(x)} = f_U(x) = \frac{d}{dx}P(\sqrt{X/n} \le x) = \frac{d}{dx}F_X(nx^2)$$

$$= 2nxf_X(nx^2) = \frac{n^{n/2}}{2^{n/2-1}\Gamma(n/2)}x^{n-1}e^{-nx^2/2}$$

for $t \ge 0$; otherwise, $f_U(t) = 0$.

Derivation of t-distribution, continued.

Suppose that Y is a standard normal random variable and independent of X. Then the quotient

$$Z = \frac{Y}{\sqrt{X/n}} = \frac{Y}{U}$$

has a t-distribution with n degrees of freedom. The pdf $f_Z(z)$ can be computed as follows.

$$\int_{-\infty}^{\infty} |x| f_Y(xz) f_U(x)\, dx = \frac{n^{n/2}}{2^{(n-1)/2}\sqrt{\pi}\Gamma(n/2)} \int_0^{\infty} x^n e^{-\underbrace{(z^2+n)x^2/2}_{t}}\, dx$$

$$= \frac{1}{\sqrt{n\pi}\Gamma(n/2)}(1+z^2/n)^{-(n+1)/2} \underbrace{\int_0^{\infty} t^{(n-1)/2} e^{-t}\, dt}_{\shortparallel} \leftarrow \begin{cases} dt = (z^2+n)\, x\, dx \\ x = 2^{\frac{1}{2}}(z^2+n)^{-\frac{1}{2}} t^{\frac{1}{2}} \end{cases}$$

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t}\, dt \quad \text{with} \quad \alpha = \frac{n+1}{2}$$

$$= \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+z^2/n)^{-(n+1)/2}$$

for $-\infty < z < \infty$.

Derivation of F-distribution.

Let U and V be independent chi-square random variables with r_1 and r_2 degrees of freedom, respectively. Then the quotient

$$Z = \frac{U/r_1}{V/r_2}$$

has F-distribution with (r_1, r_2) degree of freedom. By putting $Y = U/r_1$ and $X = V/r_2$, we can compute the pdf $f_Z(z)$ as follows.

$$\int_{-\infty}^{\infty} |x| f_Y(xz) f_X(x)\, dx \qquad f_Y(y) = r_1 f_U(r_1 y),\ f_X(x) = r_2 f_V(r_2 x)$$

$$= \frac{r_1^{r_1/2} r_2^{r_2/2}}{2^{(r_1+r_2)/2}\Gamma(r_1/2)\Gamma(r_2/2)}\, z^{r_1/2-1} \int_0^\infty x^{(r_1+r_2)/2-1} e^{-\underbrace{(r_1 z + r_2)x/2}_{t}}\, dx$$

$$= \frac{r_1^{r_1/2} r_2^{r_2/2}}{\Gamma(r_1/2)\Gamma(r_2/2)}\, (r_1 z + r_2)^{-(r_1+r_2)/2}\, z^{r_1/2-1} \int_0^\infty t^{(r_1+r_2)/2-1} e^{-t}\, dt \qquad \Gamma\left(\frac{r_1+r_2}{2}\right)$$

$$= \frac{\Gamma((r_1+r_2)/2)}{\Gamma(r_1/2)\Gamma(r_2/2)} (r_1/r_2)^{r_1/2} z^{r_1/2-1} \left(1 + \frac{r_1}{r_2} z\right)^{-(r_1+r_2)/2}, \quad 0 < z < \infty.$$

Population and random sample.

iid random variables

$$\underbrace{X_1, \dots, X_n}_{\text{Data}}$$

$$\left.\begin{aligned} \bar{X} &= \frac{X_1 + \dots + X_n}{n} \\ S^2 &= \frac{(X_1-\bar{X})^2 + \dots + (X_n-\bar{X})^2}{n-1} \end{aligned}\right\} \text{statistics}$$

The data values recorded $x_1, \dots, x_n$ are typically considered as the observed values of iid random variables $X_1, \dots, X_n$ having a common probability distribution $f(x)$. Specifically we may assume that $f(x)$ is a normal distribution with parameter (μ, σ^2). In a typical statistical problem it is useful to envisage a *population* from which the sample should be drawn. A *random sample* is chosen at random from the population, ensuring that the sample is representative of the population.

$$E[X_i] = \mu$$

$$Var(X_i) = \sigma^2$$

Sample mean.

Let $X_1, \ldots, X_n$ be iid normal random variables with parameter (μ, σ^2). The function of random sample

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

is called the *sample mean*. Then we obtain

$$E[\bar{X}] = \mu \text{ and } \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

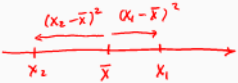
Furthermore, $\bar{X}$ has the normal distribution with parameter $(\mu, \sigma^2/n)$.

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

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

Sample variance.

The function of random sample



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

Why not by n ?

is called the *sample variance*. The sample variance S^2 can be rewritten as

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

Why ?

In particular, we can obtain

$$\begin{aligned} E[S^2] &= \frac{1}{n-1}\left(\sum_{i=1}^{n} \underbrace{E[(X_i-\mu)^2]}_{E[X_i]=\mu} - n\underbrace{E[(\bar{X}-\mu)^2]}_{E[\bar{X}]=\mu}\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} \underbrace{\mathrm{Var}(X_i)}_{\sigma^2} - n\underbrace{\mathrm{Var}(\bar{X})}_{\sigma^2/n}\right) = \sigma^2. \end{aligned}$$

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \frac{n-1}{n}E[S^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

Claim: $\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2 = (n-1)S^2$

Proof: $\sum_{i=1}^{n}\{(x_i-\mu)-(\bar{x}-\mu)\}^2 = \sum_{i=1}^{n}\{(x_i-\mu)^2 - 2(x_i-\mu)(\bar{x}-\mu) + (\bar{x}-\mu)^2\}$

$$= \sum_{i=1}^{n}(x_i-\mu)^2 - \underbrace{2(\bar{x}-\mu)\underbrace{\sum_{i=1}^{n}(x_i-\mu)}_{=\underbrace{\underbrace{\sum_{i=1}^{n}x_i}_{n\bar{x}} - n\mu}_{n(\bar{x}-\mu)}}}_{2n(\bar{x}-\mu)^2} + n(\bar{x}-\mu)^2$$

$$= \sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2$$

We don't prove it

Theorem 3

$\bar{X}$ has the normal distribution with parameter $(\mu, \sigma^2/n)$, and $\bar{X}$ and S^2 are independent. Furthermore,

$$\boxed{W_{n-1} = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

has the chi-square distribution with $(n-1)$ degrees of freedom.

$$W_{n-1} = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right\}$$

$$= \underbrace{\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}_{V} - \underbrace{\frac{n}{\sigma^2}(\bar{x}-\mu)^2}_{W_1}$$

$$\Rightarrow V = W_{n-1} + W_1$$

Rough explanation of Theorem 3.

Assuming that $\bar{X}$ and S^2 are independent (which is not so easy to prove), we can show that the random variable W_{n-1} has the chi-square distribution with $(n-1)$ degrees of freedom. By introducing $\boxed{V = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2}$ and $\boxed{W_1 = \frac{n}{\sigma^2}(\bar{X}-\mu)^2}$ we can write $V = W_{n-1} + W_1$. Since W_{n-1} and W_1 are independent, the mgf $M_V(t)$ of V can be expressed as

S^2 Indep. $\bar{X}$

$$E[e^{tW_{n-1}}e^{tW_1}] = E[e^{tV}] = M_V(t) = M_{W_{n-1}}(t) \times M_{W_1}(t), = E[e^{tW_{n-1}}] \times E[e^{tW_1}]$$

where $M_{W_{n-1}}(t)$ and $M_{W_1}(t)$ are the mgf's of W_{n-1} and W_1, respectively. Observe that W_1 has the chi-square distribution with 1 degrees of freedom. Thus, we can obtain

χ^2 with n d.f.

$$M_{W_{n-1}}(t) = \frac{M_V(t)}{M_{W_1}} = \frac{\boxed{(1-2t)^{-n/2}}}{\boxed{(1-2t)^{-1/2}}} = (1-2t)^{-(n-1)/2}$$

χ^2 with 1 d.f.

which is the mgf of the chi-square distribution with $(n-1)$ degrees of freedom.

$$V = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}\Bigg(\underbrace{\frac{X_i-\mu}{\sigma}}_{N(0,1)}\Bigg)^2 \sim \chi^2\text{-dist. with } n \text{ d.f.}$$

$$W_1 = \frac{n}{\sigma^2}(\bar{x}-\mu)^2 = \Bigg(\underbrace{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}_{N(0,1)}\Bigg)^2 \sim \chi^2 \text{ dist. with } 1 \text{ d.f.}$$

t-statistic.

Assuming iid normal random variables $X_1, \dots, X_n$ with parameter (μ, σ^2), we can construct a standard normal random variable $\boxed{Z_1 = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}$ and a chi-square random variable W_{n-1} of Theorem 3. Furthermore, the function of random sample

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\frac{\sqrt{S^2/n}}{\sigma/\sqrt{n}}}$$

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} = \frac{Z_1}{\sqrt{\frac{W_{n-1}}{n-1}}}$$

$\bar{X}$

Indep.

S^2

has the t-distribution with $(n-1)$ degrees of freedom, and it is called t-statistic. The statistic $S = \sqrt{S^2}$ is called the *sample standard deviation*.

Sampling distributions under normal assumption.

1. The sample mean $\bar{X}$ is normally distributed with parameter $(\mu, \sigma^2/n)$, and

$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

$$Z_1 = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$$

2. The statistic

$$W_{n-1} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

has a chi-square distribution with $(n-1)$ df, and independent of $\bar{X}$ (Theorem 3).

3. The t-statistic has t-distribution with $(n-1)$ df

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} = \frac{Z_1}{\sqrt{\frac{W_{n-1}}{n-1}}} \sim t_{n-1}$$

Critical points.

Given a random variable X, the *critical point* for level α is defined as the value c satisfying $P(X > c) = \alpha$.

1. Suppose that a random variable X has the chi-square distribution with n degrees of freedom. Then the critical point of chi-square distribution, denoted by $\chi_{\alpha,n}$, is defined by

$$P(X > \chi_{\alpha,n}) = \alpha.$$

2. Suppose that a random variable Z has the t-distribution with m degrees of freedom. Then the critical point is defined as the value $t_{\alpha,m}$ satisfying

$$P(Z > t_{\alpha,m}) = \alpha.$$

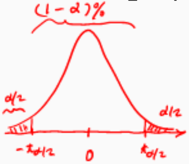
Critical points, continued.

Suppose that a random variable Z has the t-distribution with m degrees of freedom. Since the t-distribution is symmetric, the critical point $t_{\alpha,m}$ can be equivalently given by

$$P(Z < -t_{\alpha,m}) = \alpha.$$

Together we can obtain the expression

$$\begin{aligned} &P(-t_{\alpha/2,m} \le Z \le t_{\alpha/2,m}) \\ &= 1 - P(Z < -t_{\alpha/2,m}) - P(Z > t_{\alpha/2,m}) = 1 - \alpha. \end{aligned}$$



Extrema.

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables (iid random variables) with the common cdf $F(x)$ and the pdf $f(x)$. Then we can define the random variable $\boxed{V = \min(X_1, \ldots, X_n)}$ of the minimum of $X_1, \ldots, X_n$. To find the distribution of V, consider the survival function $P(V > x)$ of V and calculate as follows.

$1 - P(X_n \le x) = 1 - F(x)$

$$
\begin{aligned}
P(V > x) &= P(X_1 > x, \ldots, X_n > x) \\
&= P(X_1 > x) \times \cdots \times P(X_n > x) = [1 - F(x)]^n.
\end{aligned}
$$

$\{\omega \in \Omega : X_n(\omega) > x\}$ is a subset of Ω where $X_n : \Omega \ni \omega \to X_n(\omega) \in \mathbb{R}$ is a function on Ω.

$P(\underbrace{\{X_1 > x\}}_{A_1} \cap \underbrace{\{X_2 > x\}}_{A_2} \cap \cdots \cap \underbrace{\{X_n > x\}}_{A_n})$

independent $\longrightarrow$ $P(A_1 \cap \cdots \cap A_n) = P(A_1) \times \cdots \times P(A_n)$

$1 - P(A) = P(A^c)$

$P(V \le x)$

Since $F_V(x) = 1 - P(V > x)$, we obtain the cdf $F_V(x)$ and the pdf $f_V(x)$ of V

$$F_V(x) = 1 - [1 - F(x)]^n$$

$\frac{d}{dx} F(x) = \frac{d}{dx} \int_{-\infty}^{x} f(u)du = f(x)$

and

$$f_V(x) = \frac{d}{dx} F_V(x) = nf(x)[1 - F(x)]^{n-1}.$$

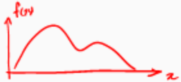
In particular, if $X_1, \ldots, X_n$ be independent exponential random variables with the common parameter λ, then the minimum $\min(X_1, \ldots, X_n)$ has the density $(n\lambda)e^{-(n\lambda)x}$ which is again the exponential density function with parameter $(n\lambda)$.

$f(x) = \lambda e^{-\lambda x}$

$F(x) = 1 - e^{-\lambda x}$

$1 - F(x) = e^{-\lambda x}$

Order statistics.



Let $X_1, \ldots, X_n$ be iid random variables with the common cdf $F(x)$ and the pdf $f(x)$. When we sort $X_1, \ldots, X_n$ as

$$\min\{X_1, \ldots, X_n\} = X_{(1)} < X_{(2)} < \cdots < X_{(n)},$$

the random variable $X_{(k)}$ is called the *k-th order statistic*. (k-th smallest value).

Theorem

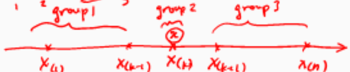
The pdf of the k-th order statistic $X_{(k)}$ is given by

$$f_k(x) = \frac{n!}{(k-1)!(n-k)!} f(x) F^{k-1}(x)[1 - F(x)]^{n-k}. \qquad (6.1)$$

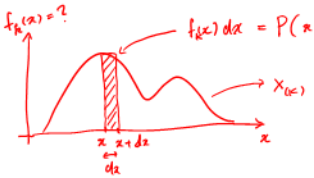
$$= \binom{n}{\underbrace{k-1}_{1},\ \underbrace{1}_{2},\ \underbrace{n-k}_{3}} [F(x)]^{k-1} f(x) [1 - F(x)]^{n-k}$$

trinomial coefficient

$$\binom{n}{i,\ j,\ k} = \frac{n!}{i!\,j!\,k!}$$



Infinitesimal probability



$$f_k(x)\,dx = P(x < X_{(k)} \le x + dx)$$

How to construct k-th order statistic $X_{(k)}$?

$X_1, \ldots, X_n$

choose $(k-1)$ out of n ①

Choose 1 index ②

Rest goes to here

$\underbrace{X_{(1)}, \ldots, X_{(k-1)}}_{(k-1) \text{ values are smaller than } X_{(k)}},\ X_{(k)},\ \underbrace{X_{(k+1)}, \ldots, X_{(n)}}_{(n-k) \text{ values are larger than } X_{(k)}}$

$\boxed{X_{(1)}, \ldots, X_{(k-1)} \le x}$ $\rightarrow$ $[F(x)]^{k-1}$

$\boxed{\text{Consider } x < X_{(k)} \le x + dx}$ $\rightarrow$ $f(x)\,dx$

$\boxed{X_{(k+1)}, \ldots, X_{(n)} > x}$ $\rightarrow$ $[1 - F(x)]^{n-k}$

① How many different ways to choose? $\binom{n}{k-1}$

② " $n-k+1$

①+② How many?

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

$$= \binom{n}{k-1,\,1,\,n-k} \leftarrow \text{Trinomial coefficient}$$

$\Downarrow$

$$\Rightarrow P(x < X_{(k)} \le x + dx) = \binom{n}{k-1,\,1,\,n-k}[F(x)]^{k-1} f(x)\,dx\,[1 - F(x)]^{n-k}$$

Proof of theorem: Infinitesimal probability.

If f is continuous at x and $\delta > 0$ is small, then the probability that a random variable X falls into the small interval $[x, x+\delta]$ is approximated by $\delta f(x)$. In differential notation, we can write

$$P(x \le X \le x + dx) = f(x)dx.$$

Now consider the k-th order statistic $X_{(k)}$ and the event that $x \le X_{(k)} \le x + dx$. This event occurs when $(k-1)$ of X_i's are less than x and $(n-k)$ of X_i's are greater than $x \approx x + dx$. Since there are $\binom{n}{k-1,1,n-k}$ arrangements for X_i's, we can obtain the probability of this event as

$$P(x < X_{(k)} \le x+dx) = f_k(x)dx = \underbrace{\frac{n!}{(k-1)!(n-k)!}(F(x))^{k-1}(f(x)dx)[1-F(x)]^{n-k}}_{\Downarrow\ f_k(x)},$$

which implies (6.1).

Beta distribution.

$f(x) = 1 \quad \text{if } 0 \le x \le 1$

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \le x \le 1 \\ 1 & \text{if } x > 1 \end{cases}$$

When $X_1, \ldots, X_n$ be iid uniform random variables on $[0, 1]$, the pdf of the k-th order statistic $X_{(k)}$ becomes

$$f_k(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}, \quad 0 \le x \le 1,$$

which is known as the *beta distribution* with parameters $\alpha = k$ and $\beta = n - k + 1$. in

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1$$

Exercises

Problem

Consider a sample $X_1, \ldots, X_n$ *of normally distributed random variables with variance* $\sigma^2 = 5$*. Suppose that* $n = 31$*.*

1. *What is the value of* c *for which* $P(S^2 \leq c) = 0.90$*?*
2. *What is the value of* c *for which* $P(S^2 \leq c) = 0.95$*?*

Problem

Consider a sample $X_1, \ldots, X_n$ of normally distributed random variables with mean μ. Suppose that $n = 16$.

1. *What is the value of c for which $P(|4(\bar{X} - \mu)/S| \leq c) = 0.95$?*
2. *What is the value of c for which $P(|4(\bar{X} - \mu)/S| \leq c) = 0.99$?*

Problem

Suppose that $X_1, \ldots, X_{10}$ are iid normal random variables with parameters $\mu = 0$ and $\sigma^2 > 0$.

1. *Let $Z = X_1 + X_2 + X_3 + X_4$. Then what is the distribution for Z?*
2. *Let $W = X_5^2 + X_6^2 + X_7^2 + X_8^2 + X_9^2 + X_{10}^2$. Then what is the distribution for $\frac{W}{\sigma^2}$?*
3. *Describe the distribution for the random variable $\frac{Z/2}{\sqrt{W/6}}$.*
4. *Find the value c for which $P\left(\frac{Z}{\sqrt{W}} \leq c\right) = 0.95$.*

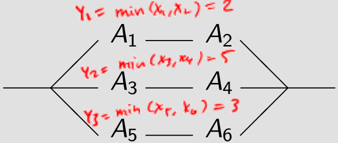
Problem

Suppose that system's components are connected in series and have lifetimes that are independent exponential random variables $X_1, \dots, X_n$ *with respective parameters* $\lambda_1, \dots, \lambda_n$.

1. *Let* V *be the lifetime of the system. Show that* $V = \min(X_1, \dots, X_n)$.
2. *Show that the random variable* V *is exponentially distributed with parameter* $\sum_{i=1}^{n} \lambda_i$.

Problem 5

Each component A_i of the following system has lifetime that is iid exponential random variable X_i with parameter λ.



1. *Let V be the lifetime of the system. Show that* $V = \max(\min(X_1, X_2), \min(X_3, X_4), \min(X_5, X_6))$. $= Y_{(3)}$
2. *Find the cdf and the pdf of the random variable V.*

Problem

A card contains n chips and has an error-correcting mechanism such that the card still functions if a single chip fails but does not function if two or more chips fail. If each chip has an independent and exponentially distributed lifetime with parameter λ, find the pdf of the card's lifetime.

$\boxed{1}$	$\cancel{\boxed{2}}$	$\cancel{\boxed{3}}$	$\cdots$	$\boxed{n}$
X_1	X_2	X_3		X_n
$\shortparallel$	$\shortparallel$	$\shortparallel$		$\shortparallel$
10	2	4		8
$X_{(1)}$	$\boxed{X_{(2)}}$	$\cdots$		$X_{(n)}$

Problem

Suppose that a queue has n servers and that the length of time to complete a job is an exponential random variable. If a job is at the top of the queue and will be handled by the next available server, what is the distribution of the waiting time until service? What is the distribution of the waiting time until service of the next job in the queue?

Optional problems

Joint distribution of order statistics.

Let U and V be iid uniform random variable on $[0,1]$, and let $X = \min(U,V)$ and $Y = \max(U,V)$. We have discussed that the probability $P(x \le X \le x + dx)$ is approximated by $f(x)\,dx$. Similarly, we can find the following formula in the case of joint distribution.

$$P(x \le X \le x + dx, y \le Y \le y + dy) = f(x,y)\,dx\,dy.$$

In particular, we can obtain

$$P(x \le U \le x + dx, y \le V \le y + dy) = dx\,dy.$$

Problem

By showing that

$$\begin{aligned} P(x \le X \le x+dx, y \le Y \le y+dy) \\ &= P(x \le U \le x+dx, y \le V \le y+dy) \\ &+ P(x \le V \le x+dx, y \le U \le y+dy), \end{aligned}$$

justify that the joint density function $f(x,y)$ of X and Y is given by

$$f(x,y) = \begin{cases} 2 & \text{if } 0 \le x \le y \le 1; \\ 0 & \text{otherwise.} \end{cases}$$

Problem

Continue Problem 8, and answer the following questions.

1. *Find the conditional density* $f_{X|Y}(x|y)$ *and* $f_{Y|X}(y|x)$.
2. *Find the conditional expectations* $E(X|Y=y)$ *and* $E(Y|X=x)$.

Problem

Let

$\frac{n!}{x!(n-x)!} y^x (1-y)^{n-x}$

$$f(x,y) = \begin{cases} \binom{n}{x} y^x (1-y)^{n-x} & \text{if } x = 0, 1, \ldots, n \text{ and } 0 \le y \le 1; \\ 0 & \text{otherwise.} \end{cases}$$

1. *Find the marginal density* $f_Y(y) = \sum_{x=0}^{n} f(x,y)$ *for* $0 \le y \le 1$, *and the conditional frequency function* $f_{X|Y}(x|\theta)$ *given* $Y = \theta$. *Then identify the distribution for* $f_{X|Y}(x|\theta)$.
2. *Find the marginal freqency function* $f_X(x) = \int_0^1 f(x,y)\, dy$.
3. *Given* $x = k$, *find the conditional density and* $f_{Y|X}(y|k)$, *and identify the distribution for* $f_{Y|X}(y|k)$.

In Bayesian statistics (later in this semester) $f_Y(y)$ *is called a prior distribution of parameter* θ *of binomial distribution* (n, θ), *and* $f_{Y|X}(y|k)$ *is a posterior distribution of parameter* θ *provided that we have observed* $X = k$ *from the binomial distribution.*

Answers to exercises

Problem 1.

Observe that $30S^2/5 = 6S^2$ has a χ^2-distribution with 30 degrees of freedom.

(a) $P(S^2 \leq c) = P(6S^2 \leq 6c) = 0.90$ implies $6c = \chi^2_{0.1,30} \approx 40.26$. Thus, $c \approx 6.7$.

(b) $P(S^2 \leq c) = P(6S^2 \leq 6c) = 0.95$ implies $6c = \chi^2_{0.05,30} \approx 43.77$. Thus, $c \approx 7.3$.

Problem 2.

Observe that $\frac{4(\bar{X}-\mu)}{S} = \frac{\bar{X}-\mu}{S/\sqrt{16}}$ has a t-distribution with 15 degrees of freedom.

(a) $c = t_{0.025,15} \approx 2.13$.

(b) $c = t_{0.005,15} \approx 2.95$.

Problem 3.

(a) $N(0, 4\sigma^2)$.

(b) Since $W/\sigma^2 = (X_5/\sigma)^2 + \cdots + (X_{10}/\sigma)^2$ and each (X_i/σ) has the standard normal distribution, W/σ^2 has the chi-square distribution with 6 degrees of freedom.

(c) We can write $\dfrac{Z/2}{\sqrt{W/6}} = \dfrac{Z/(2\sigma)}{\sqrt{(W/\sigma^2)/6}}$. Since $Z/(2\sigma)$ is independent of W/σ^2, and has the standard normal distribution, $\dfrac{Z/2}{\sqrt{W/6}}$ has the t-distribution with 6 degree of freedom.

(d) Since $P\left(\dfrac{Z}{\sqrt{W}} \leq c\right) = P\left(\dfrac{Z/2}{\sqrt{W/6}} \leq \dfrac{\sqrt{6}}{2}c\right)$, the result of (c) implies that $\frac{\sqrt{6}}{2}c = t_{0.05,6} \approx 1.94$. Thus, $c \approx 1.58$.

Problem 4.

(b) $P(V > t) = P(X_1 > t) \times \cdots \times P(X_1 > t) = e^{-\lambda_1 t} \times \cdots \times e^{-\lambda_n t} = \exp\left(-\sum_{i=1}^{n} \lambda_i\right)$ is the survival function of V. Therefore, V is an exponential random variable with parameter $\sum_{i=1}^{n} \lambda_i$.

Problem 5.

(b) Let $Y_1 = \min(X_1, X_2)$, $Y_2 = \min(X_3, X_4)$, and $Y_3 = \min(X_5, X_6)$. Then Y_1, Y_2 and Y_3 are iid exponential random variables with parameter 2λ. Since $V = Y_{(3)}$, we obtain

$$f_V(t) = 6\lambda e^{-2\lambda t}(1 - e^{-2\lambda t})^2, \quad t > 0.$$

By integrating it, we can find the cdf $F_V(t) = (1 - e^{-2\lambda t})^3$.

Problem 6.

Let X_i denote the lifetime of the i-th chip, and let V denote the lifetime of the card. Then X_i's are iid exponential random variables with λ, and $V = X_{(2)}$. Thus, we obtain

$$f_V(t) = n(n-1)\lambda[e^{-\lambda(n-1)t} - e^{-\lambda nt}], \quad t > 0.$$

Problem 7.

Each of n servers completes a job independently in an exponentially distributed time with λ. Then it is known that the sojourn time between the jobs is exponential with $n\lambda$, and that the sojourn times are independent each other. Therefore, (i) the distribution of the waiting time at the top of the queue is exponential with $n\lambda$, and (ii) the distribution of the waiting time for the second in the queue is given by the sum of two iid exponential random variables, that is, the gamma distribution with $(2, n\lambda)$.