Module 2.1
Standard Distributions

S. Lakshmivarahan

School of Computer Science
University of Oklahoma
Norman, OK, 73071
USA

Univariate normal/Gaussian distribution

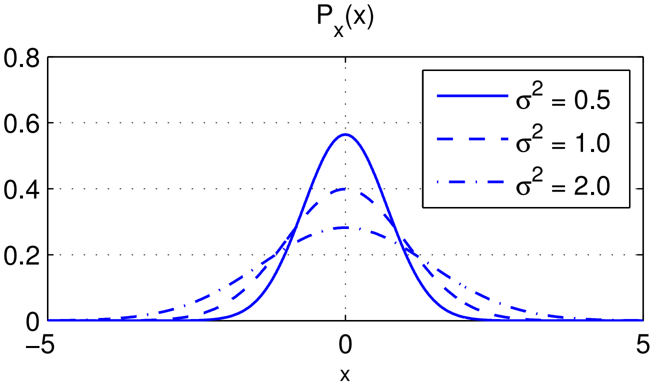
- A scalar random variable x is said to have a Gaussian, or normal distribution, if its probability density function is given by

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

- This probability density function is described by two parameters, μ the mean (location) and σ^2, the variance (the spread) of x.
- $P_x(x)$ in (1) is denoted by $\mathbf{N}(\mu, \sigma^2)$.
- When μ=0 and σ^2=1 in (1), it is called standard normal distribution, $\mathbf{N}(0, 1)$

Examples

- $P_x(x)$ is a symmetric function of x with respect to μ, that is, $P_x(x-\mu) = P_x(\mu - x)$
- Variation of $P_x(x)$ with σ^2 is illustrated below



- As σ^2 increases, the peak at $\mu = 0$ decreases, the tail gets thicker and the overall spread increases

Cumulative probability distribution

- $$\text{If } x \sim N(\mu, \sigma^2), \quad \text{then } Z = \frac{x-\mu}{\sigma} \sim N(0,1). \tag{2}$$
- By definition:

$$\begin{aligned} F(a) = \textbf{Prob}[Z \le a] &= \int_{-\infty}^{a} P_z(z)\,\mathrm{d}\,z \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} \exp\left[-\frac{z^2}{2}\right] \mathrm{d}\,z. \end{aligned} \tag{3}$$

  denotes the cumulative probability distribution of z.
- Then, $F(-\infty) = 0$, $F(0) = \frac{1}{2}, F(\infty) = 1$.

Probability mass of $N(0, 1)$ on intervals

- Since it is not easy to evaluate the integral in (1) , numerical values of $F(a)$ have been extensively tabulated.
- Using these tables: $F(a) - F(a) =$ area under the curve from a to b

$$
\begin{aligned}
F(1) - F(-1) &= \textbf{Prob}[-1 \leq z \leq 1] = 0.683 \\
F(2) - F(-2) &= \textbf{Prob}[-2 \leq z \leq 2] = 0.955 \\
F(3) - F(-3) &= \textbf{Prob}[-3 \leq z \leq 3] = 0.997
\end{aligned}
\tag{4}
$$

Probability mass of $N(\mu, \sigma^2)$ on intervals

- Using (2) and (4) it is immediate that

$$\begin{aligned} \textbf{Prob}[|x-\mu| \leq \sigma] &= 0.683 \\ \textbf{Prob}[|x-\mu| \leq 2\sigma] &= 0.955 \\ \textbf{Prob}[|x-\mu| \leq 3\sigma] &= 0.997 \end{aligned} \tag{5}$$

Sum of iid random variables

- Let $x_1, x_2, \cdots, x_n$ be a set of n independent, identically (not necessarily Gaussian) distributed random variables.
- Define

$$S_n = \sum_{i=1}^{n} x_i. \tag{6}$$

- Verify:

$$\mathbf{Mean}(S_n) = \mathbf{E}(S_n) = n\mu \tag{7}$$

$$\begin{aligned}
\mathbf{var}(S_n) &= \mathbf{E}[S_n - \mu]^2 \\
&= \mathbf{E}(\sum_{i=1}^{n} x_i - n\mu)^2 = \mathbf{E}(\sum_{i=1}^{n}(x_i - \mu)^2) \\
&= n\sigma^2
\end{aligned} \tag{8}$$

A function of S_n-centering and normalization

- Notice that S_n is such that its mean (7) and variance (8) increases linearly with n
- However, there exists a function, $g(S_n)$ of S_n whose distribution is related to the standard normal distribution
- To this end,define

$$y_n = g(S_N) = \frac{S_n - n\mu}{\sqrt{n}\sigma} \tag{9}$$

- Subtraction of the mean $n\mu$ from S_n is called *centering*, and dividing by the standard deviation $\sqrt{n}\sigma$ is called normalization.

- CLAIM: The distribution of the random variable y_n in (9) tends towards the standard normal as $n \to \infty$.That is,

$$\lim_{n\to\infty} \textbf{Prob}[a < y_n \le b] = \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(-\frac{y^2}{2}\right) \mathrm{d}\, y \qquad (10)$$

- Gaussian distribution called a "stable" distribution.

- Let $x_1, x_2, \cdots, x_n$ be the iid samples from a distribution with unknown mean μ and known variance σ^2
- A standard estimate for μ is the sample mean.

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{11}$$

Properties of $\overline{X}_n$

- Verify that:

$$E(\overline{X}_n) = \mu$$
$$var(\overline{X}_n) = \frac{\sigma^2}{n} \tag{12}$$

- By CLT, the sampling distribution of $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ is standard normal as n $\to \infty$, that is,

$$\lim_{n\to\infty} \textbf{Prob}\left[\frac{\sigma}{\sqrt{n}}a \leq (\overline{X}_n - \mu) \leq \frac{\sigma}{\sqrt{n}}b\right] = \frac{1}{\sqrt{2\pi}}\int_a^b \exp\left(-\frac{y^2}{2}\right) \mathrm{d}\, y \tag{13}$$

Confidence interval for $\overline{X}_n$

- Let

$$Z_n = \frac{(\overline{X}_n - \mu)}{(\frac{\sigma}{\sqrt{n}})} \tag{14}$$

- By CLT in (13), for $0 < \alpha < 1$, if

$$\textbf{Prob}[-Z_{\frac{\alpha}{2}} \leq Z_n \leq Z_{\frac{\alpha}{2}}] = 1 - \alpha \tag{15}$$

then Z_n in (14) lies in the interval $[-Z_{\frac{\alpha}{2}}, Z_{\frac{\alpha}{2}}]$ with probability $(1 - \alpha)$, where $Z_{\frac{\alpha}{2}} > 0$

- Verify from the Tables of standard normal:

$Z_{\frac{\alpha}{2}}$	$(1-\alpha)$	α
1	0.683	0.317
2	0.955	0.045
3	0.997	003

Table : Confidence intervals

Confidence interval for $\overline{X}_n$

- Substituting Z_n from (14) in (15):

$$\textbf{Prob}\left[\mu - \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}} \leq \overline{X}_n \leq \mu + \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}\right] = 1 - \alpha. \qquad (16)$$

- That is, $\overline{X}_n$ lies in the interval $[\mu - \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}, \mu + \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}]$ with probability $(1-\alpha)$. This interval is a function of n and α.
- α is called the level of confidence

Relation between n and α

- Let $d > 0$ be such that

$$|\overline{X}_n - \mu| \leq d \tag{17}$$

- Then

$$-\frac{d}{(\frac{\sigma}{\sqrt{n}})} \leq \frac{\overline{X}_n - \mu}{(\frac{\sigma}{\sqrt{n}})} \leq \frac{d}{(\frac{\sigma}{\sqrt{n}})} \tag{18}$$

- By CLT,

$$\textbf{Prob}[-\frac{\sqrt{n}d}{\sigma} \leq Z_n \leq \frac{\sqrt{n}d}{\sigma}] = 1 - \alpha \tag{19}$$

$$\text{where } Z_{\frac{\alpha}{2}} = \frac{\sqrt{n}d}{\sigma} \ \ or \ \ n = \frac{\sigma^2}{d^2} Z^2{}_{\frac{\alpha}{2}} \tag{20}$$

- Thus, α decides $Z_{\frac{\alpha}{2}}$ which in turn decides n through (20).

Chi-square (χ^2) distribution

- A scalar random variable x is said to be $\chi^2(n)$-distributed with n degrees of freedom if

$$f_x(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\left(\frac{n}{2}-1\right)} e^{-\frac{x}{2}}, \text{ for } \ x > 0 \tag{21}$$

denoted as $x \sim \chi^2(n)$

- $\Gamma(r)$ is the standard Gamma function
- (21) is a special case Gamma distribution

$$f_x(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} \tag{22}$$

with $\lambda = \frac{1}{2}$ *and* $r = \frac{n}{2}$

Properties of $\Gamma(r)$

- $\Gamma(r) = \int_0^\infty x^{r-1} e^{-\lambda} \,\mathrm{d}\, x \;\; r > 0$
- $\Gamma(1) = 1, \Gamma(\frac{1}{2}) = \sqrt{\pi}$
- $\Gamma(r+1) = r\Gamma(r)$ when r is real and positive
- $\Gamma(r+1) = r!$ when r is an integer
- $\binom{n+r-1}{n} = \frac{\Gamma(n+r)}{\Gamma(n+1)\Gamma(r)}$
- $\frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)} = \int_0^1 u^{r-1}(1-u)^{s-1} \,\mathrm{d}\, u$

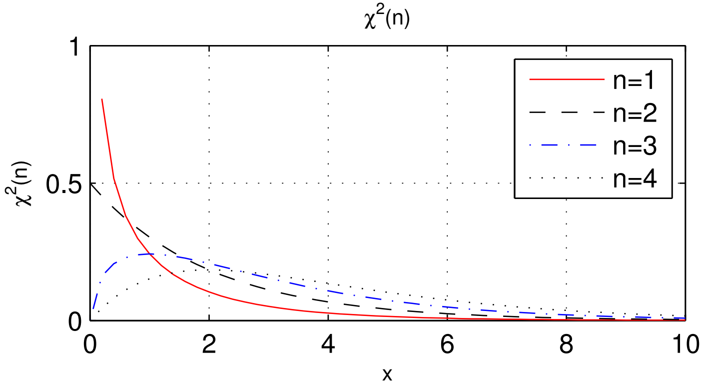
Mean and variance

Factors of x	Gamma	$\chi^2(n)$
Mean	$\frac{r}{\lambda}$	n
Variance	$\frac{r}{\lambda^2}$	$2n$

Table : Mean and variance of x

(23)

Sample plots of $\chi^2(n)$



Examples of $\chi^2(n)$ random variables

- Let $z_1, z_2, \cdots, z_n$ be iid samples from $N(0, 1)$
- Then
$$\sum_{i=1}^{n} (z_i)^2 \sim \chi^2(n) \tag{24}$$

Examples of $\chi^2(n)$ random variables

- Let $x_1, x_2, \cdots, x_n$ be iid samples from $N(0, \sigma^2)$
- If μ is known, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ is an estimator of σ^2
- If μ and σ are <u>not</u> known, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X}_n)^2$ are the estimators of μ and σ^2.
- Then,

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n) \text{ and } \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \tag{25}$$

F-distribution — after Sir Ronald Fisher

- Let $U \sim \chi^2(m), V \sim \chi^2(n)$ and be independent.
- Then

$$X = \frac{U/m}{V/n} \sim F_{m,n}, \quad \text{called } F - \text{distribution} \tag{26}$$

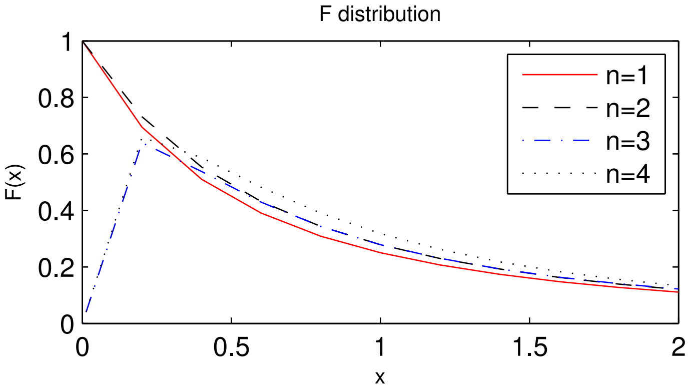
- It is given by

$$f_x(x) = \frac{\Gamma(\frac{(m+n)}{2}) m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\frac{m}{2}-1}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})(n+mx)^{\frac{m+n}{2}}} \tag{27}$$

Application of F-distribution

- F-distribution is used to test the properties of a statistic which is the ratio of two χ^2-distributed variables.
- See the module on linear least squares for an illustration.

Sample plots of F-distribution



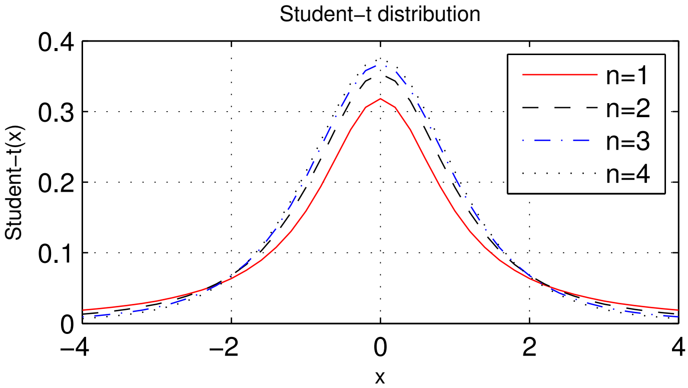
Student-t distribution

- Let $Z \sim N(0,1)$, $V \sim \chi^2(n)$and be independent.
- The ratio $\mathrm{x} = \frac{Z}{\sqrt{\frac{V}{n}}}$ is said to inherit the student-t distribution

$$f_x(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})(1+\frac{x^2}{n})^{\frac{n+1}{2}}} \tag{28}$$

for $-\infty < x < \infty$

- This distribution is symmetric with respect to the y-axis.


Student–t distribution
Student–t(x)
x
0.4
0.3
0.2
0.1
0
–4
–2
0
2
4
n=1
n=2
n=3
n=4

References

- The following book contain a wealth of information on various distributions:

Krishnan, V. (2016) *"Probability and random processes."* Wiley