STAT 23400 Lecture 7 Normal Distribution and joint probability distribution

Review

- A random variable X is continuous if and only if there is a function $f(x)$, such that
  - $f(x) \geq 0$ for any x
  - $P(a \leq X \leq b) = \int_a^b f(x)dx$, for any $-\infty < a < b < \infty$.
- A density curve $f(x)$ must satisfy the following conditions (a). It must be nonnegative, i.e., $f(x) \geq 0$ for all x. (b). The total area under the density curve must be 1, i.e.,

$$\int_{-\infty}^{\infty} f(x)dx = P(-\infty < X < \infty) = 1.$$

- Cumulative distribution function for a r.v. X is de
ned as

$$F(x) = F_X(x) = P(X \leq x) :$$

- Expected value:

$$\mu = E(X) = \int_{-\infty}^{\infty} x f_X(x) dx, E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

- Moments: $m_k = E(X^k) = \int_{-\infty}^{\infty} x^k f_X(x) dx :$
- Variance and Standard deviation:

$$Var(X) = E(X - \mu)^2, SD(X) = \sigma = \sqrt{Var(X)}$$

- Expected value and variance under linear transformation.
- Moment Generating Function:

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

Outline

Topics:

- Normal distribution (Section 4.1 in OIS and 4.3 in MMSA)
- Scaling Property of Normal Distributions (Section 4.1 in OIS and 4.3 in MMSA.)
- Joint probability distribution (Section 5.1 in MMSA).
- Lab04: Normal Distribution.

Normal Distributions

Normal Distributions (Section 4.3 in MMSA)

A random variable X is said to have a normal distribution (aka. Gaussian distributions) with a mean μ, and an SD σ denoted as

$$X \sim N(\mu, \sigma), \qquad \text{in OIS4}$$
$$X \sim N(\mu, \sigma^2), \qquad \text{in MMSA}$$

if this pdf is

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

The density curve is bell-shaped and symmetric about its center μ; whereas its variability is reflected by the value of σ.

* To avoid the confusion, we use $X \sim N(\mu, \sigma^2)$ throughout the lectures and assignments.*

Normal Distribution

- The pdf of $X \sim N(\mu, \sigma^2)$,

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

- One can show that
  - $\int_{-\infty}^{\infty} f(x;\mu,\sigma)dx = 1$ (not a trivial calculation, but can be found in most calculus textbooks)
  - $E(X) = \int_{-\infty}^{\infty} xf(x;\mu,\sigma)dx = \mu$
  - $Var(X-\mu) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x;\mu,\sigma)dx = \sigma^2$

Standard Normal Distribution

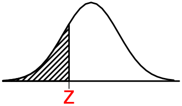
A normal distribution with $\mu = 0$, and $\sigma = 1$ is called the **standard normal distribution**, denoted as $N(0, 1)$.

The cdf of the standard normal distribution $Z \sim N(0, 1)$ is

$$\Phi(z) = P(Z < z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du.$$

- The cdf $\Phi(z)$ has no close-form formula.
- The normal probability table (on p.410-411 in OIS4 or p.792-793 in MMSA) gives the values of the cdf $\Phi(z)$ for different z.

The normal probability table (on p.410-411 in OIS4) gives

$\Phi(z) = P(Z < z) =$ area of shaded region in 

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

E.g., for $z = -0.83$, find $\Phi(z)$.

$P(Z < 1.573) = ?$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

For $Z \sim N(0, 1)$, $P(Z > -0.83) = ?$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148

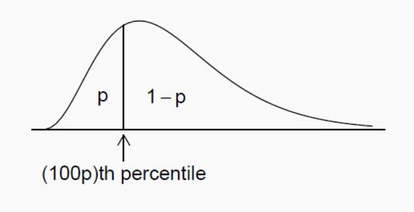
$P(-0.83 < Z < 2) =?$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857

Percentile

For $0 < p < 1$, the **(100p)th percentile** of a continuous random variable X is a value q such that

$$p = P(X \leq q)$$



Suppose we want to find the 25th percentile of the standard normal, i.e., the z such that

$$\Phi(z) = P(Z < z) = 0.25.$$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Example: Percentiles of the Standard Normal (2)

The 95th percentile of the standard normal is the z such that

$$\Phi(z) = 0.95.$$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633

Scaling Property of Normal Distributions

Scaling Property of Normal Distributions

- If $X \sim N(\mu, \sigma^2)$, its standardized z-score $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$
- Conversely, if $Z \sim N(0,1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$:
- This is because all normal distributions have the same shape; differ only in:
  - position given by the mean μ
  - scale given by the standard deviation σ.

Remark. Not all distributions have scaling properties, ex., $(X-a)/b$ is not binomial even if X is binomial r.v.

CDF of N(μ, σ^2)

If $X \sim N(\mu, \sigma^2)$, then by definition, the CDF of X is

$$F(x) = P(X \le x) = P\Big(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\Big) = \Phi\Big(\frac{x-\mu}{\sigma}\Big).$$

Since $\frac{X-\mu}{\sigma} \sim N(0,1)$.

Example: SAT (1)

The distribution of SAT scores was about normal with mean 1500 and SD 300. About what percent of students score below 1800 on the SAT?

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830

Example: SAT (2)

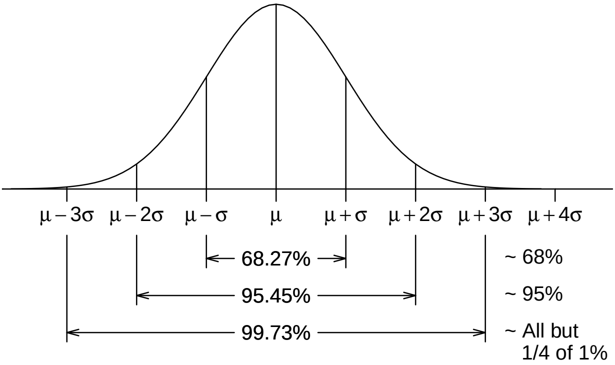
Q: What proportion of SAT test takers scored between 1650 and 1800?

Example: SAT (3)

Z	0.05	0.06	0.07	0.08	0.09
1.1	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9115	0.9131	0.9147	0.9162	0.9177

What is the SAT score that is at the 90% percentile?

68-95-99.7% Rule for the Normal Distributions



In terms of the standard normal CDF $\Phi(z)$.

$$\Phi(1) - \Phi(-1) \approx 0.6827$$
$$\Phi(2) - \Phi(-2) \approx 0.9545$$
$$\Phi(3) - \Phi(-3) \approx 0.9973$$

Normal Distribution in R

Type '`help(rnorm)`' in R console to get help documentation.

Density, distribution function, quantile function and random generation for the normal distribution with mean equal to mean and standard deviation equal to sd.

Usage

- Evaluate pdf at x: `dnorm(x, mean = 0, sd = 1, log = FALSE)`.
- Evaluate cdf at x: `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)`.
- Get the 100pth percentile: `qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)`.
- Generate random samples: `rnorm(n, mean = 0, sd = 1)`.

Joint Probability Distribution: Discrete Case

Multivariate Random Variables

- Our focus so far has been on the distribution of a single random variable.
- In many situations, there are two or more variables of interest, and we want to know how they are related. For example, I am interested to know

  X_1: the number of hours spent on studying per week

  X_2: final grade of stat234.

- Since the relationship is important, we cannot study them separately.
- As a matter of fact, we need to study their joint behavior, thus we need to define multivariate random variables.

Joint Distribution of Two Discrete Random Variables

The **joint probability mass function (joint pmf)**, or, simply **joint distribution**, of two discrete r.v. X and Y is defined as

$$p(x, y) = P(X = x, Y = y) = P(\{X = x\} \cap \{Y = y\})$$

Properties of the joint probability distribution:

1. $p(x, y) \geq 0$.
2. Define the probability for an event A as,

$$P(A) = P((X, Y) \in A) = \sum_{(x,y) \in A} p(x, y)$$

3. If we set A = S in (2), then

$$P(S) = \sum_x \sum_y p(x, y) = 1.$$

Example: Joint Distribution of Two Random Variables

A white die and a black die are rolled. What is the joint pmf for X = the smaller number, and Y = the bigger number?

$p(x, y)$	1	2	3	4	5	6
1	1/36	2/36	2/36	2/36	2/36	2/36
2	0	1/36	2/36	2/36	2/36	2/36
3	0	0	1/36	2/36	2/36	2/36
4	0	0	0	1/36	2/36	2/36
5	0	0	0	0	1/36	2/36
6	0	0	0	0	0	1/36

- What is $P(X = 1, Y = 1)$?
- If $x = y, P(X = x, Y = y) =?$
- $P(X = 2, Y = 1) =?$
- If $x > y, P(X = x, Y = y) =?$
- What is $P(X = 1, Y = 2)$?