Chapter 4: Distributions of random variables

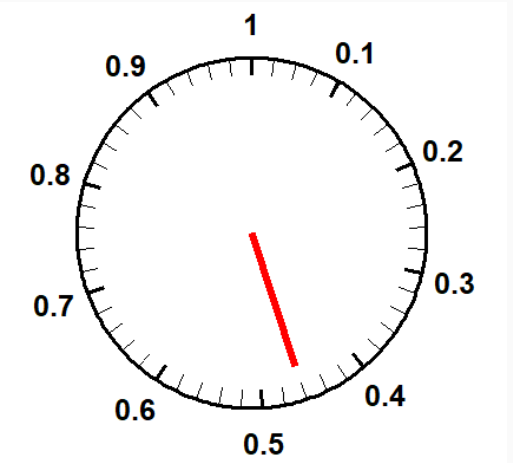
Yonghyun Kwon

Department of Mathematics, Korea Military Academy

Uniform distribution

Uniform distribution

A circular disk is marked with a scale from 0 to 1 along its edge. A needle, fixed at the center, is spun freely and lands at a random position on the scale. Consider a random variable X, a value indicated by the needle.

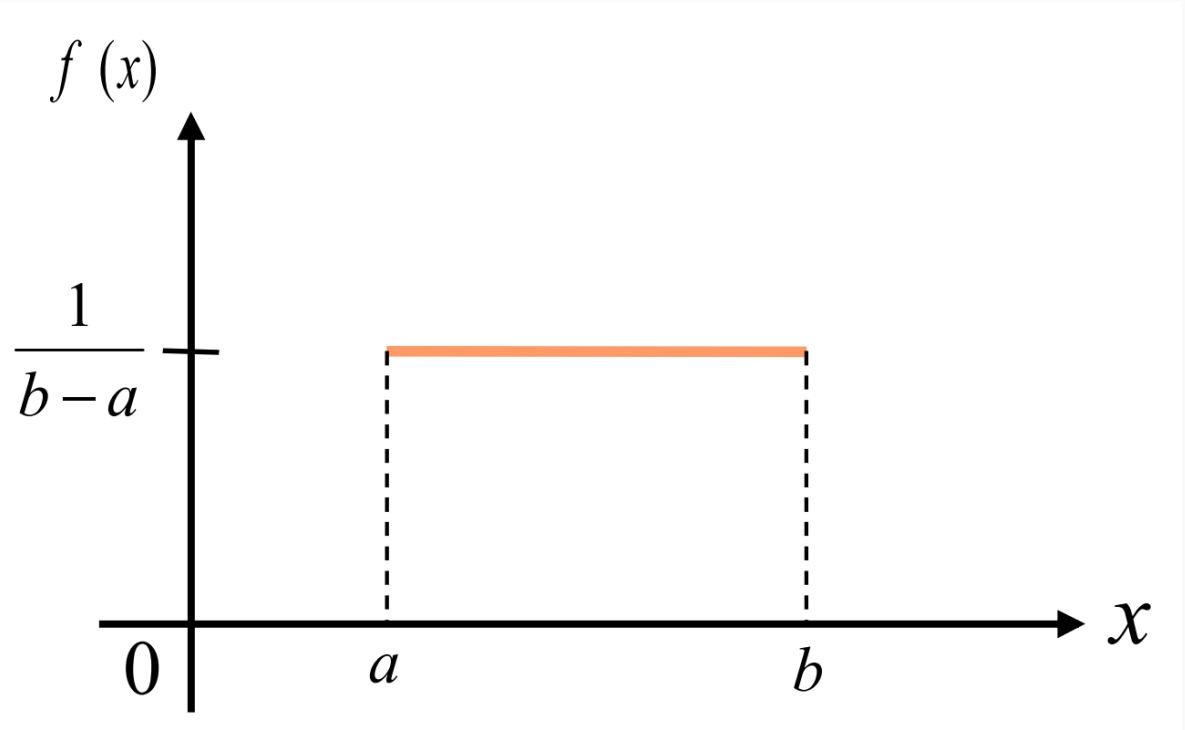


- Is the random variable X discrete or continuous?
- What is the probability distribution of X?

Uniform random variable

X is a uniform random variable on the interval (a, b) if X has a pdf

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases}$$



Expectation and variance of uniform distribution

Recall: Expectation and variance of a continuous RV X

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx, \quad E\left(X^2\right) =$$

$$Var(X) =$$

Suppose X has a uniform distribution on the interval (a, b).

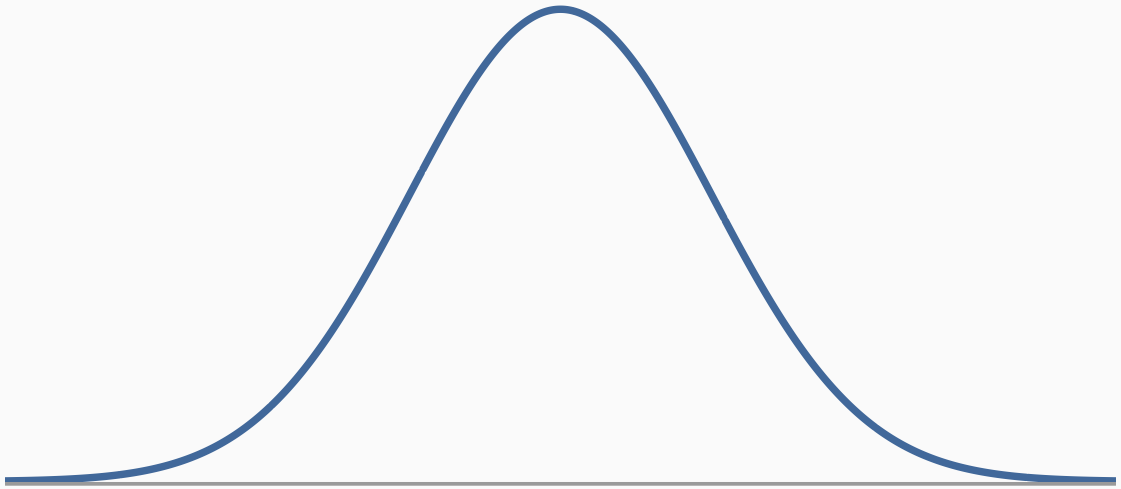
- Find $E(X)$:

- Find $Var(X)$:

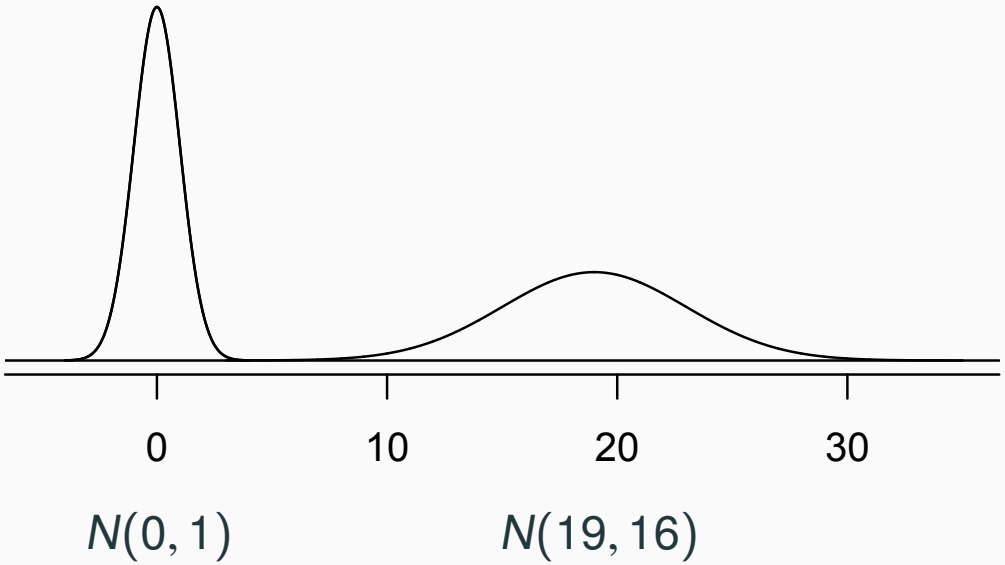
Normal distribution

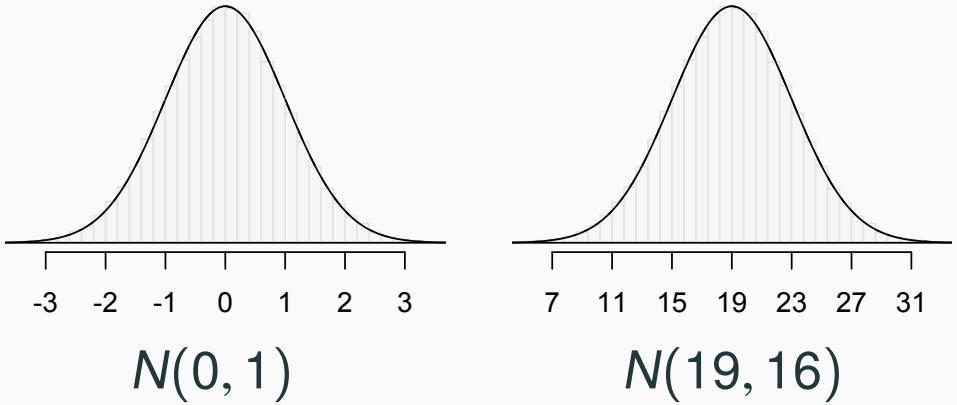
Normal distribution

- Unimodal and symmetric, bell shaped curve.
- Many variables are nearly normal, but rarely perfectly normal.
- Denoted as $N(\mu, \sigma^2)$ $\rightarrow$ Normal with mean μ and variance σ^2



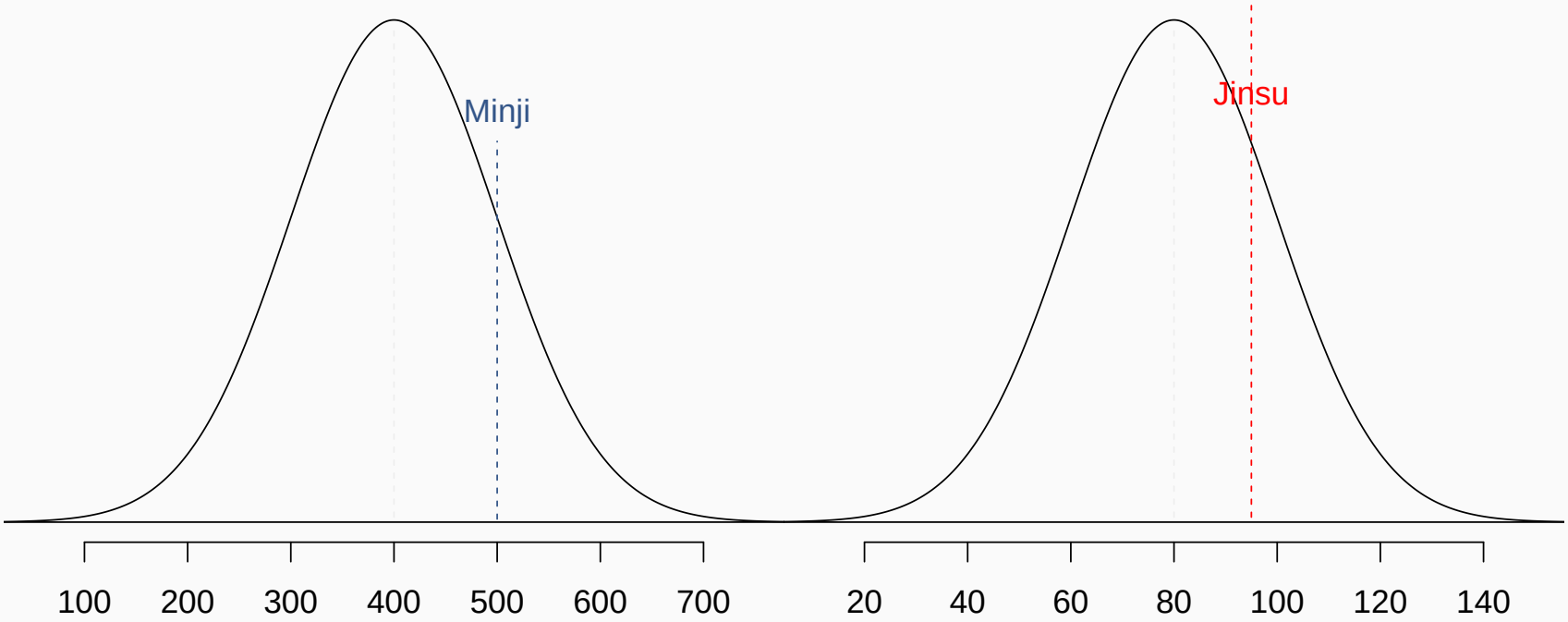
Normal distributions with different parameters





Comparing test scores using Z scores

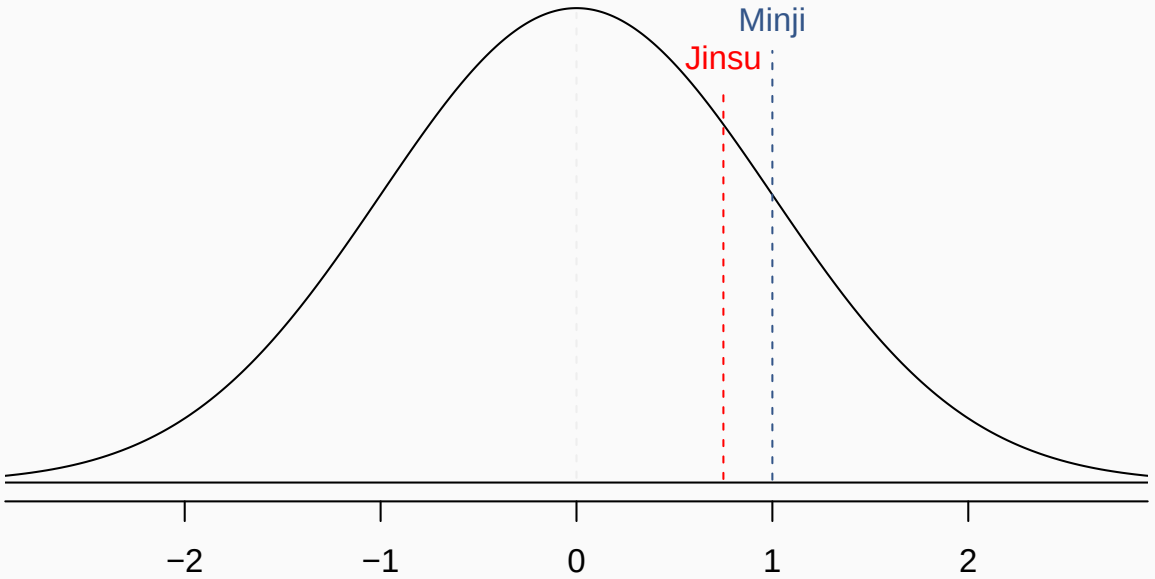
- TEPS scores: mean = 400, standard deviation = 100
- TOEFL scores: mean = 80, standard deviation = 20
- Minji scored 500 on the TEPS; Jinsu scored 95 on the TOEFL.
- Who performed better relative to their peers?



Standardizing with Z scores

Since we cannot just compare these two raw scores, we instead compare how many **standard deviations** beyond the mean each observation is.

- Minji's score is $\frac{500 - 400}{100} = 1$ standard deviation above the mean.
- Jinsu's score is $\frac{95 - 80}{20} = 0.75$ standard deviation above the mean.



Standardizing with Z scores (cont.)

- These are called *standardized* scores, or *Z scores*.
- Z score of an observation is the number of standard deviations it falls above or below the mean.

Standardization

$$Z = \frac{observation - mean}{SD} = \frac{X - \mu}{\sigma}$$

- Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.

Properties of normal distribution

Recall: properties of expectation and variance

$E(aX+b) =$, $Var(aX+b) =$

Properties of normal distribution(1)

- If $X \sim N(\mu, \sigma^2)$, $aX+b \sim N($, $)$

If $X \sim N(\mu, \sigma^2)$, find the distribution of $Z = \dfrac{X-\mu}{\sigma}$.

Properties of normal distribution(cont.)

Recall: a linear combination of two RVs

$$E(aX + bY) = aE(X) + bE(Y)$$

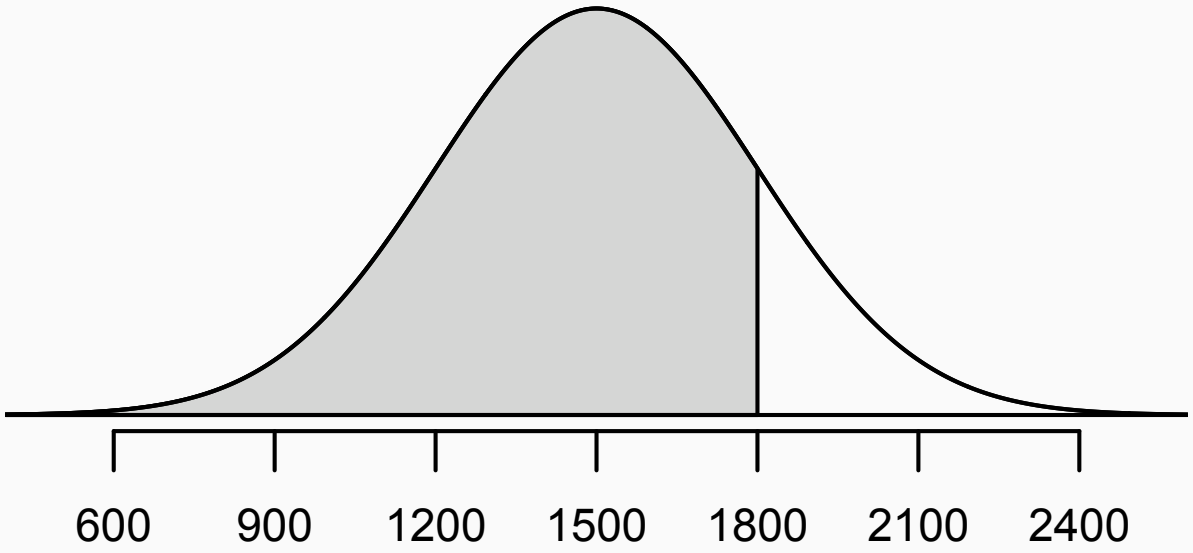
$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) \quad (X,\ Y \text{ indep.})$$

Properties of normal distribution(2)

- If $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent,
  $aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$

Area under normal distribution

- Often, we are interested in the **proportion** of observations that fall below a given data point.
- Graphically, it is the **area** below the probability distribution curve to the left of that observation.



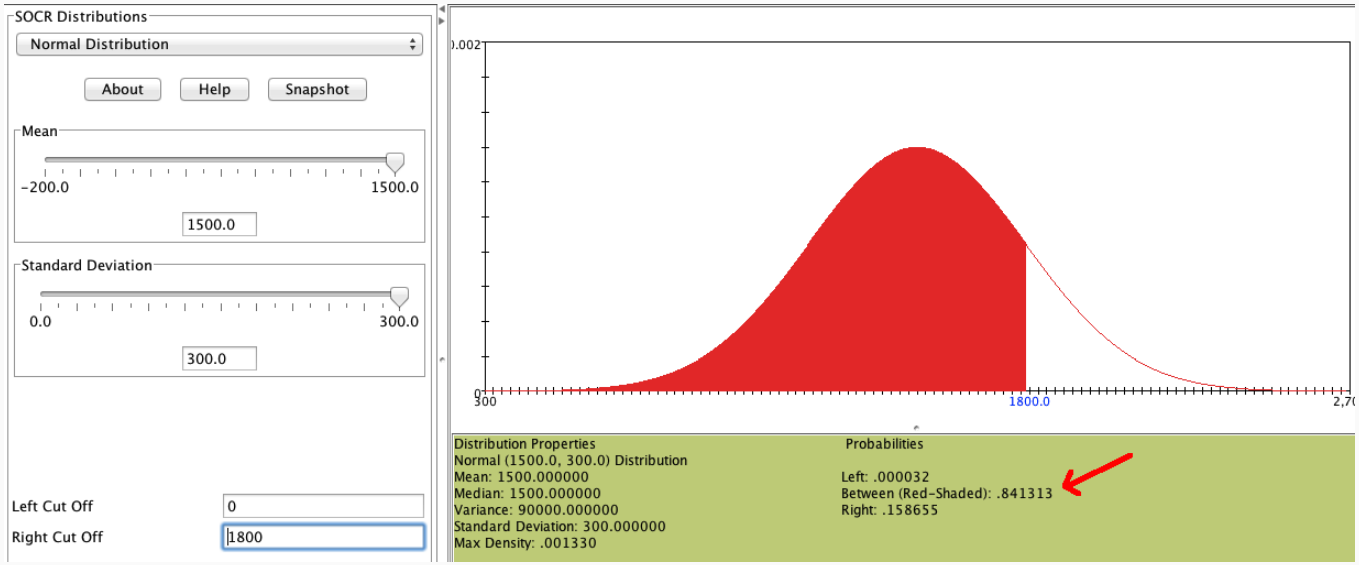
Calculating probabilities - using computation

There are many ways to compute areas under the curve:

- R:

```
> pnorm(1800, mean = 1500, sd = 300)
[1] 0.8413447
```

- Applet: *https://gallery.shinyapps.io/dist_calc/*



Calculating probabilities - using tables

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
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
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

Quality control

Bottles of Heinz ketchup are normally distributed with mean 36 oz and standard deviation 0.11 oz. A bottle fails quality control if it is below 35.8 oz or above 36.2 oz.

(1) What percent of bottles have less than 35.8 oz?

Finding the exact probability - using R

```
> pnorm(-1.82, mean = 0, sd = 1)
[1] 0.0344
```

OR

```
> pnorm(-1.82)
[1] 0.0344
```

Practice

(2) What percent of bottles pass the quality control inspection?

(a) 1.82%

(b) 3.44%

(c) 6.88%

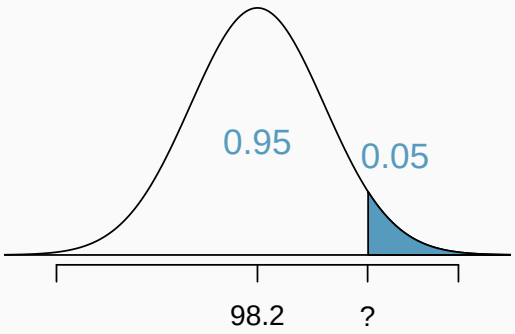
(d) 93.12%

(e) 96.56%

Finding cutoff points

Body temperatures of humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F.

(1) What is the cutoff for the highest 5% of human body temperatures?



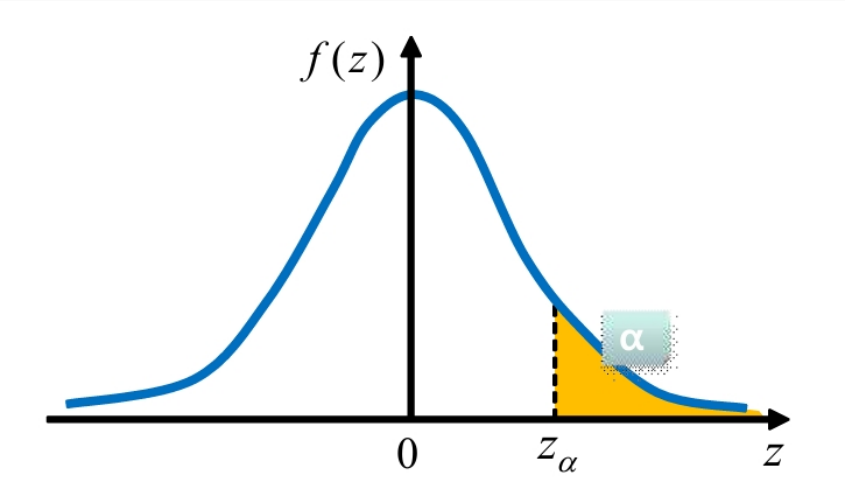
$$P(X > x) = 0.05 \leftarrow P(Z > 1.64) = 0.05$$

$$z = \frac{x - mean}{SD} =$$

$$x =$$

```
> qnorm(0.05, lower.tail = FALSE)
[1] 1.644854
```

Quantile(Percentile)



α-quantile

The (upper) α-quantile of the standard normal distribution, denoted as z_α, is the value such that a proportion α of the distribution lies *above* it. That is, $P(Z > z_\alpha) = \alpha$.

For example, $z_{0.005} = 2.58$, $z_{0.025} = 1.96$, and $z_{0.05} = 1.645$.

Practice

Body temperatures of humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F.

(2) What is the cutoff for the lowest 2.5% of human body temperatures?

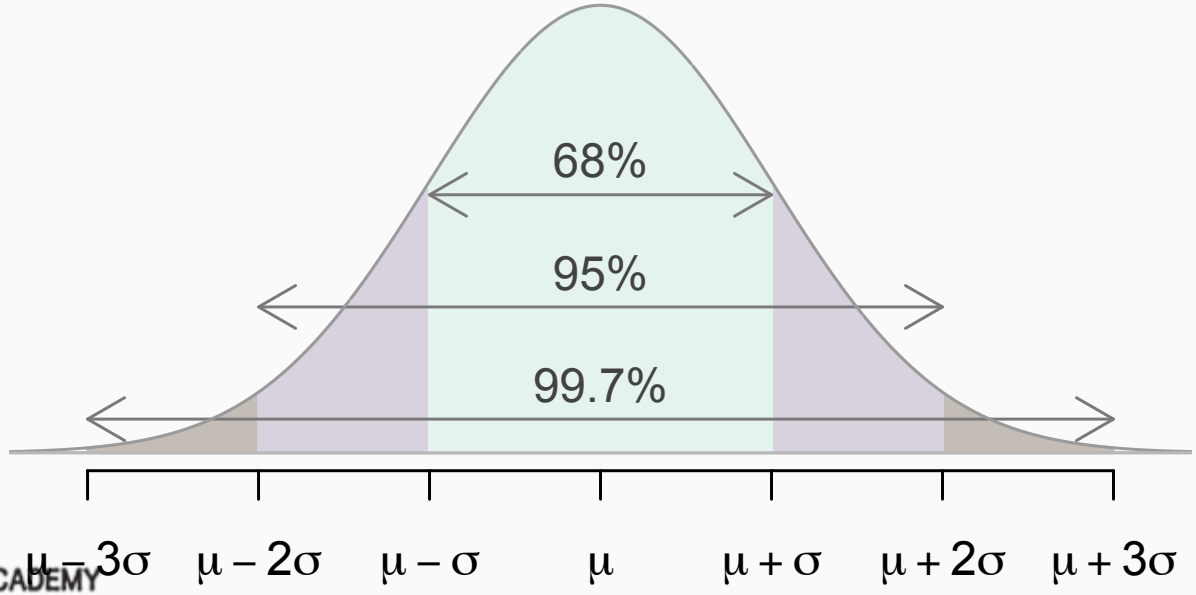
(You may use $z_{0.005} = 2.58$, $z_{0.025} = 1.96$, or $z_{0.05} = 1.645$)

$$P(X < x) = P\left(\frac{X - 98.2}{0.73} < \frac{x - 98.2}{0.73}\right) = P\left(Z < \frac{x - 98.2}{0.73}\right) =$$

$$\frac{x - 98.2}{0.73} =$$

68-95-99.7 Rule

- For nearly normally distributed data,
  - about 68% falls within 1 SD of the mean,
  - about 95% falls within 2 SD of the mean,
  - about 99.7% falls within 3 SD of the mean.
- It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



Practice

Which of the following is false?

(a) Majority of Z scores in a right skewed distribution are negative.

(b) In skewed distributions the Z score of the mean might be different than 0.

(c) For a normal distribution, IQR is less than $2 \times SD$.

(d) Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

Binomial distribution

Health insurance plans

- In South Korea, everyone is covered under the National Health Insurance Service (NHIS).
- Each year, some insured individuals use medical services, while others do not.
- We can model such “use vs. no use” outcomes as a random variable.

Bernoulli random variable in health insurance

- An insurance agency reports: 70% of people used hospital services last year.
- Think of each insured person as a single *trial*.
- Define a random variable X with two possible outcomes:
  - $X = 1$ (*Success*): Used hospital services,
  - $X = 0$ (*Failure*): No hospital use.
- The *probability of success*, p, is the probability that $X = 1$:

$$P(X = 1) = p = 0.7, \qquad P(X = 0) = 1 - p = 0.3 (= q)$$

- Then X follows a *Bernoulli distribution* with parameter $p = 0.7$.

Bernoulli random variable

A random variable X has a Bernoulli distribution with probability of success $p(0 \le p \le 1)$ if the pmf of X is $f(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$

Recall: Expectation and variance of a discrete RV X

$$\mu = E(X) = \qquad , \quad Var(X) = E\left[(X-\mu)^2\right] = \sum_{\forall x} (x-\mu)^2 f(x)$$

Suppose X follows a Bernoulli distribution with success probability p.

- Find $E(X) =$
- Find $Var(X) =$

Sample proportion

- Suppose we observe ten trials:

$$1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0$$

- *Sample proportion*, $\hat{p}$, is the sample mean of the observations:

$$\begin{aligned}\hat{p} &= \frac{\text{\# of successes}}{\text{\# of trials}} \\ &= \frac{1+1+1+0+1+0+0+1+1+0}{10} = 0.6\end{aligned}$$

Probability of Using Hospital Services

Suppose we select 4 insured individuals. What is the probability that exactly 1 of them used hospital services last year?

$$\text{Scenario 1:} \quad \underset{\text{(A) } \textit{used}}{0.3} \times \underset{\text{(B) not}}{0.7} \times \underset{\text{(C) not}}{0.7} \times \underset{\text{(D) not}}{0.7} = 0.103$$

$$\text{Scenario 2:} \quad \underset{\text{(A) not}}{0.7} \times \underset{\text{(B) } \textit{used}}{0.3} \times \underset{\text{(C) not}}{0.7} \times \underset{\text{(D) not}}{0.7} = 0.103$$

$$\text{Scenario 3:} \quad \underset{\text{(A) not}}{0.7} \times \underset{\text{(B) not}}{0.7} \times \underset{\text{(C) } \textit{used}}{0.3} \times \underset{\text{(D) not}}{0.7} = 0.103$$

$$\text{Scenario 4:} \quad \underset{\text{(A) not}}{0.7} \times \underset{\text{(B) not}}{0.7} \times \underset{\text{(C) not}}{0.7} \times \underset{\text{(D) } \textit{used}}{0.3} = 0.103$$

The probability that exactly one of the four used hospital services is the sum of all scenarios:

$$0.103 + 0.103 + 0.103 + 0.103 = 4 \times 0.103 = 0.412$$

Binomial distribution

The previous question asked for the probability of getting

$$k = 1 \quad \text{success in} \quad n = 4 \quad \text{trials.}$$

We found this by:

$$\underbrace{\#\text{ of scenarios}}_{\binom{n}{k}} \times \underbrace{P(\text{one specific scenario})}_{p^k(1-p)^{n-k}}$$

The *Binomial distribution* describes the probability of k successes in n *independent* Bernoulli trials with probability of success p.

Counting the number of scenarios

For $n = 4$ and $k = 1$, we could list all cases by hand.

But what if n were larger? For example, $n = 9$ and $k = 2$.

$$UUNNNNNNN$$
$$NUUNNNNNN$$
$$NNUUNNNNN$$
$$\cdots$$
$$NNUNNUNNN$$
$$\cdots$$
$$NNNNNNNUU$$

Listing them all would be tedious and error-prone.

Calculating the # of scenarios

Choose function

The *choose function* is useful for calculating the number of ways to choose k successes in n trials.

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

- $k = 1, n = 4$: $\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4 \times 3 \times 2 \times 1}{1 \times (3 \times 2 \times 1)} = 4$
- $k = 2, n = 9$: $\binom{9}{2} =$

Binomial distribution (cont.)

- p:
- n:
- X:

Binomial Distribution

The probability mass function of X is given by:

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

where $k \in \{0, 1, \ldots, n\}$. We say that X follows a binomial distribution with parameters n and p, and denoted it as:

$$X \sim B(n, p).$$

Note: If X has the Bernoulli distribution with probability of success p, $X \sim B(1, p)$.

Which of the following is not a condition that needs to be met for the binomial distribution to be applicable?

(a) the trials must be independent

(b) the number of trials, n, must be fixed

(c) each trial outcome must be classified as a *success* or a *failure*

(d) the number of desired successes, k, must be greater than the number of trials

(e) the probability of success, p, must be the same for each trial

A 1998 Gallup Korea survey suggests that 26.2% of Koreans are obese.

(1) Among a random sample of 10 Koreans, what is the probability that exactly 8 are obese?

(a) $0.262^8 \times 0.738^2$

(b) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(c) $\binom{10}{8} \times 0.262^8 \times 0.738^2$

(d) $\binom{10}{8} \times 0.262^2 \times 0.738^8$

Expected value

A 1998 Gallup Korea survey suggests that 26.2% of Koreans are obese.

(2) Among a random sample of 100 Koreans, how many would you expect to be obese?

Expected value and its variability

Mean and standard deviation of binomial distribution

If $X \sim B(n, p)$, $E(X) = np$, $Var(X) = np(1-p)$.

- Going back to the obesity rate:

  $$Var(X) = np(1-p) =$$

- We would expect 26.2 out of 100 randomly sampled Koreans to be obese, with a standard deviation of 4.4.

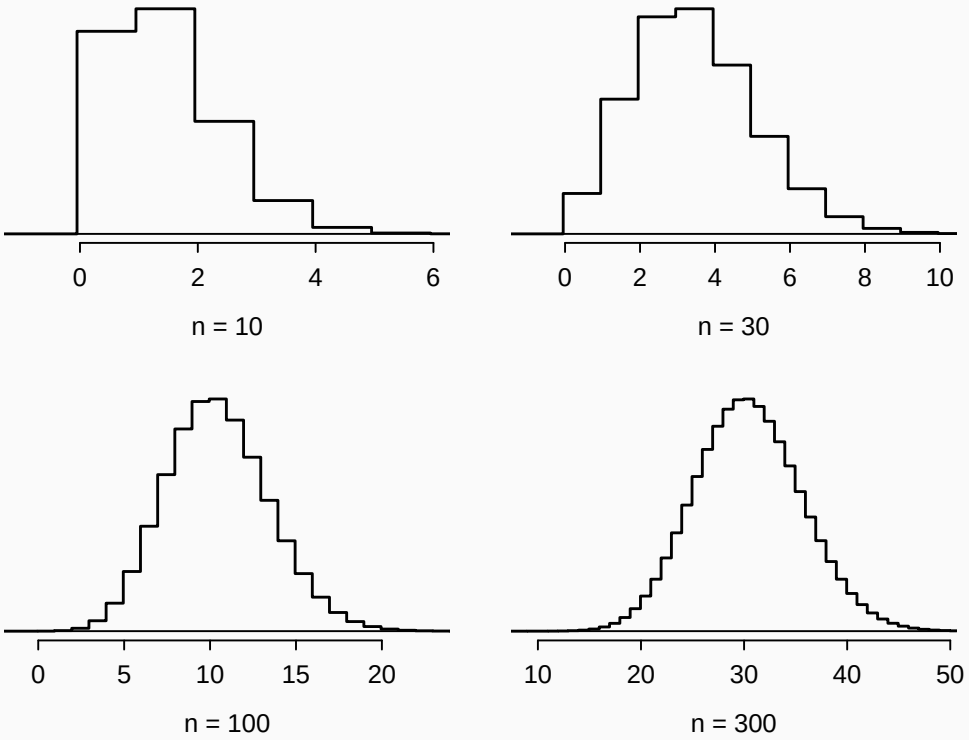
Note: If $X_i \sim B(1, p)$ are independent Bernoulli randomvariables with probability of success p, $X = \sum_{i=1}^{n} X_i \sim B(n, p)$. Thus,

$$E(X) = E(X_1 + \cdots + X_n) = E(X_1) + \cdots E(X_n) = np$$

$$Var(X) = Var(X_1 + \cdots + X_n) = Var(X_1) + \cdots Var(X_n) = np(1-p)$$

Distributions of number of successes

Hollow histograms of samples from the binomial distribution where $p = 0.10$ and $n = 10$, 30, 100, and 300. What happens as n increases?



How large is large enough?

The sample size is considered large enough if the expected number of successes and failures are both at least 10.

$$np \geq 10 \qquad \text{and} \qquad n(1-p) \geq 10$$

Below are four pairs of Binomial distribution parameters. Which distribution can be approximated by the normal distribution?

(a) $n = 100, p = 0.95$

(b) $n = 25, p = 0.45$

(c) $n = 150, p = 0.05$

(d) $n = 500, p = 0.015$

On a multiple-choice exam, each question has 4 options, and a student randomly guesses on every question. What is the probability that the student gets at least 70 correct out of 245 questions?

We are given that $n = 245$, $p = 0.25$, and we are asked for the probability $P(X \geq 70)$.

$$\begin{aligned} P(X \geq 70) &= P(X = 70 \text{ or } X = 71 \text{ or } \cdots \text{ or } X = 245) \\ &= P(X = 70) + P(X = 71) + \cdots + P(X = 245) \end{aligned}$$

Clearly, calculating this directly would be very tedious...

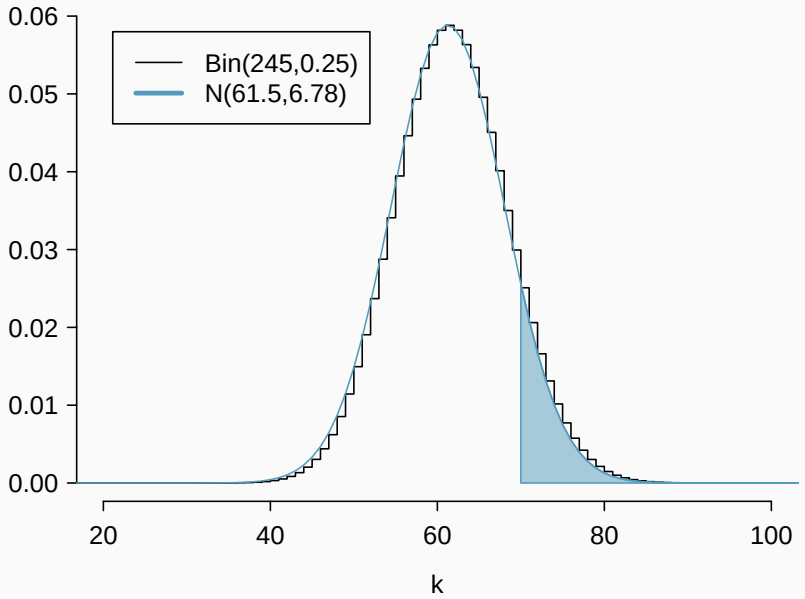
Normal approximation to the binomial

When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal distribution with parameters $\mu = np$ and $\sigma^2 = np(1-p)$.

- For the exam problem, $n = 245$ and $p = 0.25$.

  $\mu =$ $\sigma^2 =$

- $B(n = 245, p = 0.25) \approx N(\mu = \quad , \sigma^2 = \quad)$.



Using the normal approximation, what is the probability that the student gets at least 70 correct answers out of 245 questions?

Exercises

Exercises in OpenIntro Statistics 4th ed.

- Normal distribution: Exercise 4.7, 4.9
- Binomial distribution: Exercise 4.18 (b), (d), 4.20 (a), (c)