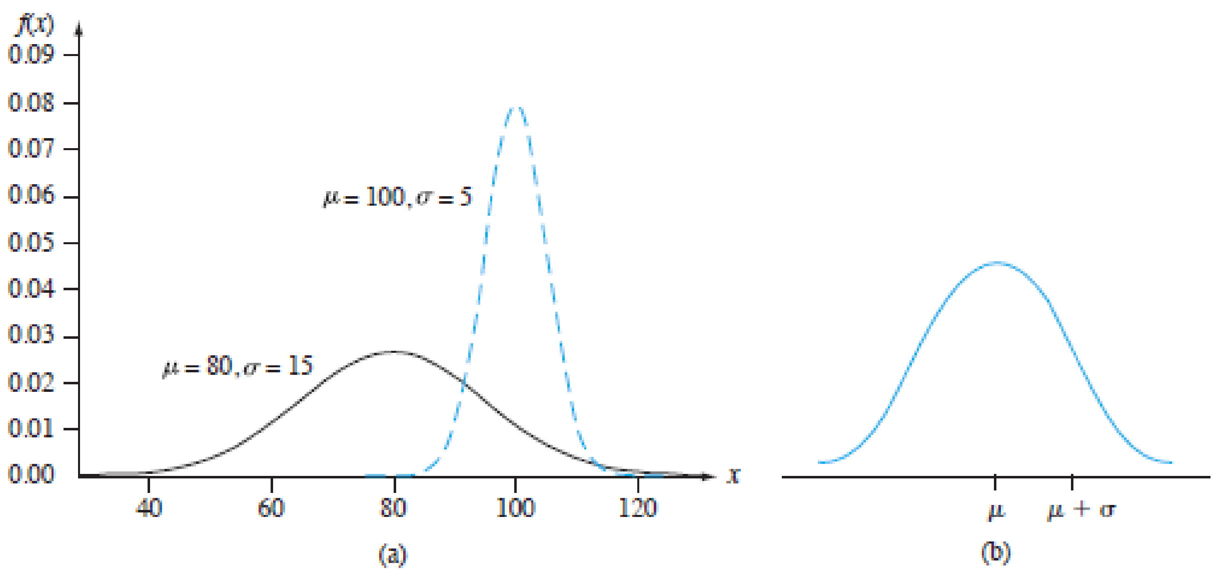
Probability and Mathematical Statistics in Data Science

Lecture 29: Section 10.3: Normal Distribution

The Normal Distribution

A continuous rv X is said to have a **normal distribution** with parameters μ and σ (or μ and σ^2), where $-\infty < \mu < \infty$ and $0 < \sigma$, if the pdf of X is

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \quad -\infty < x < \infty \tag{4.3}$$

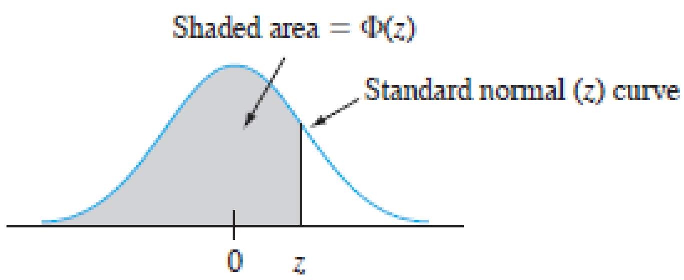


The Standard Normal Distribution

The normal distribution with parameter values $\mu = 0$ and $\sigma = 1$ is called the **standard normal distribution.** A random variable having a standard normal distribution is called a **standard normal random variable** and will be denoted by Z. The pdf of Z is

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad -\infty < z < \infty$$

The graph of $f(z; 0, 1)$ is called the *standard normal* (or z) curve. Its inflection points are at 1 and -1. The cdf of Z is $P(Z \leq z) = \int_{-\infty}^{z} f(y; 0, 1)\, dy$, which we will denote by $\Phi(z)$.



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

Differences Between Population Means

Basic Assumptions

1. $X_1, X_2, \ldots, X_m$ is a random sample from a distribution with mean μ_1 and variance σ_1^2.
2. $Y_1, Y_2, \ldots, Y_n$ is a random sample from a distribution with mean μ_2 and variance σ_2^2.
3. The X and Y samples are independent of one another.

Differences Between Population Means

The expected value of $\overline{X} - \overline{Y}$ is $\mu_1 - \mu_2$, so $\overline{X} - \overline{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$. The standard deviation of $\overline{X} - \overline{Y}$ is

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

- Both these results depend on the rules of expected value and variance. Since the expected value of a difference is the difference of expected values

$$E(\overline{X} - \overline{Y}) = E(\overline{X}) - E(\overline{Y}) = \mu_1 - \mu_2$$

$$V(\overline{X} - \overline{Y}) = V(\overline{X}) + V(\overline{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$$

Confidence Intervals for Difference between Population Means

Provided that m and n are both large, a CI for $\mu_1 - \mu_2$ with a confidence level of approximately $100(1 - \alpha)\%$ is

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

where $-$ gives the lower limit and $+$ the upper limit of the interval. An upper or a lower confidence bound can also be calculated by retaining the appropriate sign ($+$ or $-$) and replacing $z_{\alpha/2}$ by z_α.

Hypothesis Test for Difference between Population Means

Null hypothesis: $H_0\colon \mu_1 - \mu_2 = \Delta_0$

Test statistic value: $z = \dfrac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}}$

Alternative Hypothesis	Rejection Region for Level α Test
$H_a\colon \mu_1 - \mu_2 > \Delta_0$	$z \geq z_\alpha$ (upper-tailed)
$H_a\colon \mu_1 - \mu_2 < \Delta_0$	$z \leq -z_\alpha$ (lower-tailed)
$H_a\colon \mu_1 - \mu_2 \neq \Delta_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed)

- The p-value is calculated as we did previously

The NHANES National Youth Fitness Survey

Cardiorespiratory Endurance Dataset (Y_CEX)

First Published: January 2016

The cardiorespiratory endurance component (variable name prefix CEX) measured cardiorespiratory fitness using a treadmill exercise test. The goals of this component were to provide nationally representative data on cardiorespiratory endurance.

Participants aged 6-11 years, who did not meet any of the exclusion criteria, were eligible for this component.

https://wwwn.cdc.gov/Nchs/Nnyfs/Y_CEX.htm

The NHANES National Youth Fitness Survey

- Cardiorespiratory Endurance

- **Variable of Interest:** (Maximal) Endurance Time

- We would like to compare the mean endurance time for boys versus girls aged 6-11.

- We will complete a hypothesis test to see if there is statistical evidence in the data that the mean endurance in the population is different for boys and girls.

Hypothesis Testing: Population Mean Difference

Step 1: Null and Alternative Hypothesis

Null Hypothesis: population mean difference is equal to zero

Alternative Hypothesis: population mean difference is not equal to zero

Hypothesis Testing: Population Mean Difference

Step 2: Model

Data: Variable of Interest: (Maximal) Endurance Time

Boys: sample size = 327 sample mean = 663.3 sample standard deviation = 152.4

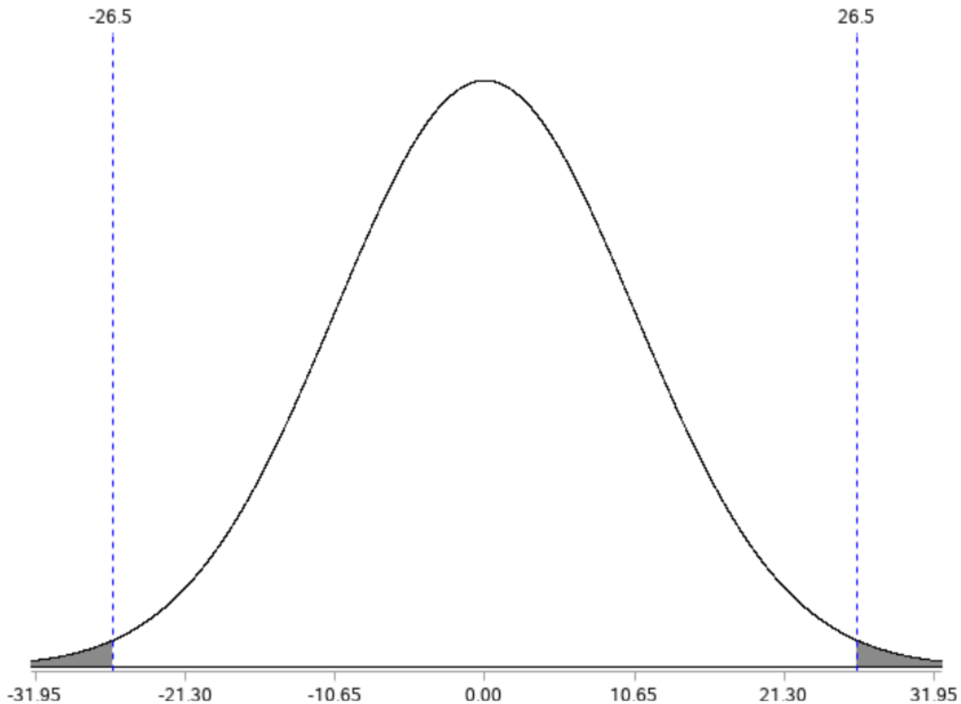
Girls: sample size = 355 sample mean = 636.8 sample standard deviation = 122.7

Sample Mean Difference (Boys minus Girls)
= 663.3 – 636.8 = 26.5 seconds

Standard Error (of Sample Mean Difference) = 10.65 seconds

Hypothesis Testing: Population Mean Difference

Step 2: Model



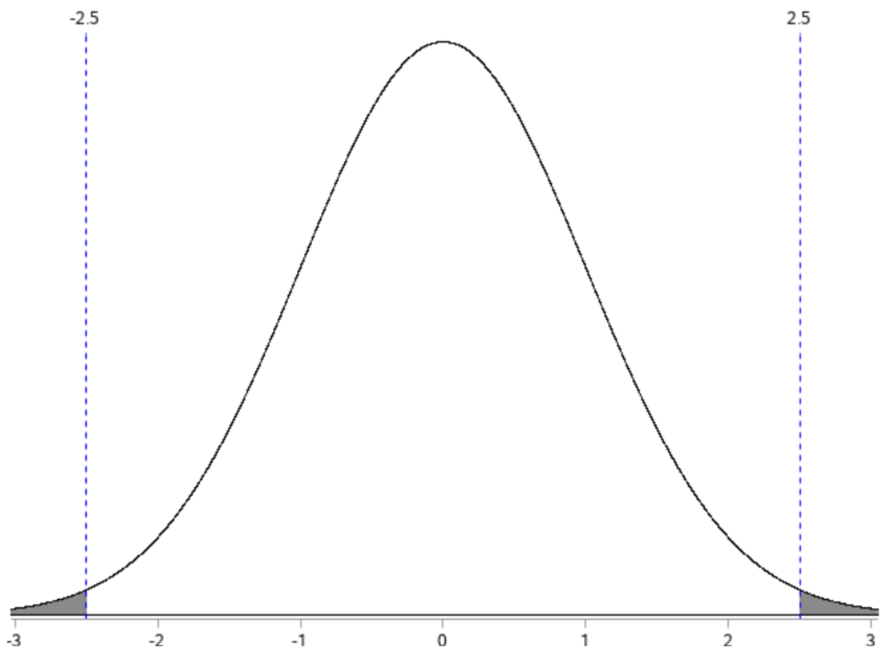
For a two-sided alternative, the p-value is the probability of obtaining a sample mean endurance time at least as far from zero (in either direction) as the one we found in our sample of data given the null hypothesis is correct.

Hypothesis Testing: Population Mean Difference

Step 3: Calculations

test statistic = (sample mean difference – null value)/standard error

= (26.5 – 0) / 10.56 = 2.5



- The p-value is the probability of obtaining a test statistic greater than 2.5 or less than -2.5 which is equal to 0.012

Hypothesis Testing: Population Mean Difference

Step 4: Conclusion

- Since the p-value equal to 0.012 is less than 0.05, we reject the null hypothesis in favor of the alternative

- We have statistical evidence that the population mean difference (boys minus girls) in endurance time is not equal to zero

- More specifically, the data indicates that the mean duration time of boys (in the population) is greater for boys than for girls

Confidence Intervals: Population Mean Difference

Data: Variable of Interest: (Maximal) Endurance Time

- The 95% confidence interval is calculated as follows:

sample mean difference ± 2 x standard error

$26.5 \pm 2 \times 10.65$

[5.2, 47.8]

Differences in Proportions

Let $\hat{p}_1 = X/m$ and $\hat{p}_2 = Y/n$, where $X \sim \text{Bin}(m, p_1)$ and $Y \sim \text{Bin}(n, p_2)$ with X and Y independent variables. Then

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

so $\hat{p}_1 - \hat{p}_2$ is an unbiased estimator of $p_1 - p_2$, and

$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{m} + \frac{p_2 q_2}{n} \quad (\text{where } q_i = 1 - p_i) \tag{9.3}$$

Differences in Proportions

- Assuming that $p_1 = p_2 = p$, instead of separate samples of size *m* and *n* from two different populations (two different binomial distributions), we really have a single sample of size m + n from one population with proportion *p*.

- The total number of individuals in this combined sample having the characteristic of interest is X + Y

- The natural estimator of *p* is then

$$\hat{p} = \frac{X+Y}{m+n} = \frac{m}{m+n} \cdot \hat{p}_1 + \frac{n}{m+n} \cdot \hat{p}_2$$

Hypothesis Test for Comparing Proportions

Null hypothesis: $H_0: p_1 - p_2 = 0$

Test statistic value (large samples): $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{m} + \dfrac{1}{n}\right)}}$

Alternative Hypothesis	Rejection Region for Approximate Level α Test
$H_a: p_1 - p_2 > 0$	$z \geq z_\alpha$
$H_a: p_1 - p_2 < 0$	$z \leq -z_\alpha$
$H_a: p_1 - p_2 \neq 0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

A P-value is calculated in the same way as for previous z tests.
The test can safely be used as long as $m\hat{p}_1$, $m\hat{q}_1$, $n\hat{p}_2$, and $n\hat{q}_2$ are all at least 10.

Example

The article "Aspirin Use and Survival After Diagnosis of Colorectal Cancer" (*J. of the Amer. Med. Assoc.*, 2009: 649–658) reported that of 549 study participants who regularly used aspirin after being diagnosed with colorectal cancer, there were 81 colorectal cancer-specific deaths, whereas among 730 similarly diagnosed individuals who did not subsequently use aspirin, there were 141 colorectal cancer-specific deaths.

Does this data suggest that the regular use of aspirin after diagnosis will decrease the incidence rate of colorectal cancer-specific deaths? Let's test the appropriate hypotheses using a significance level of .05.

Example

The parameter of interest is the difference $p_1 - p_2$, where p_1 is the true proportion of deaths for those who regularly used aspirin and p_2 is the true proportion of deaths for those who did not use aspirin.

The use of aspirin is beneficial if $p_1 < p_2$, which corresponds to a negative difference between the two proportions. The relevant hypotheses are therefore:

$$H_0: p_1 - p_2 = 0 \qquad \text{versus} \qquad H_a: p_1 - p_2 < 0$$

Example

Parameter estimates are $\hat{p}_1 = 81/549 = .1475$, $\hat{p}_2 = 141/730 = .1932$, and $\hat{p} = (81 + 141)/(549 + 730) = .1736$. A z test is appropriate here because all of $m\hat{p}_1, m\hat{q}_1, n\hat{p}_2$, and $n\hat{q}_2$ are at least 10. The resulting test statistic value is

$$z = \frac{.1475 - .1932}{\sqrt{(.1736)(.8264)\left(\frac{1}{549} + \frac{1}{730}\right)}} = \frac{-.0457}{.021397} = -2.14$$

- The corresponding P-value for a lower-tailed z test is $\Phi(-2.14) = 0.0162 < 0.05$.

Confidence Interval for Differences in Proportions

A CI for $p_1 - p_2$ with confidence level approximately $100(1 - \alpha)\%$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m} + \frac{\hat{p}_2\hat{q}_2}{n}}$$

This interval can safely be used as long as $m\hat{p}_1, m\hat{q}_1, n\hat{p}_2$, and $n\hat{q}_2$ are all at least 10.

Example

The authors of the article "Adjuvant Radiotherapy and Chemotherapy in Node-Positive Premenopausal Women with Breast Cancer" (*New Engl. J. of Med.*, 1997: 956–962) reported on the results of an experiment designed to compare treating cancer patients with chemotherapy only to treatment with a combination of chemotherapy and radiation.

Of the 154 individuals who received the chemotherapy-only treatment, 76 survived at least 15 years, whereas 98 of the 164 patients who received the hybrid treatment survived at least that long. With $p1$ denoting the proportion of all such women who, when treated with just chemotherapy, survive at least 15 years and $p2$ denoting the analogous proportion for the hybrid treatment

Example

$$\hat{p}_1 = 76/154 = .494 \text{ and } 98/164 = .598$$

A confidence interval for the difference between proportions with a confidence level of 99% is

$$\begin{aligned} .494 - .598 \pm (2.58)\sqrt{\frac{(.494)(.506)}{154} + \frac{(.598)(.402)}{164}} &= -.104 \pm .143 \\ &= (-.247, .039) \end{aligned}$$