Confidence Interval - Data Analysis and Visualization

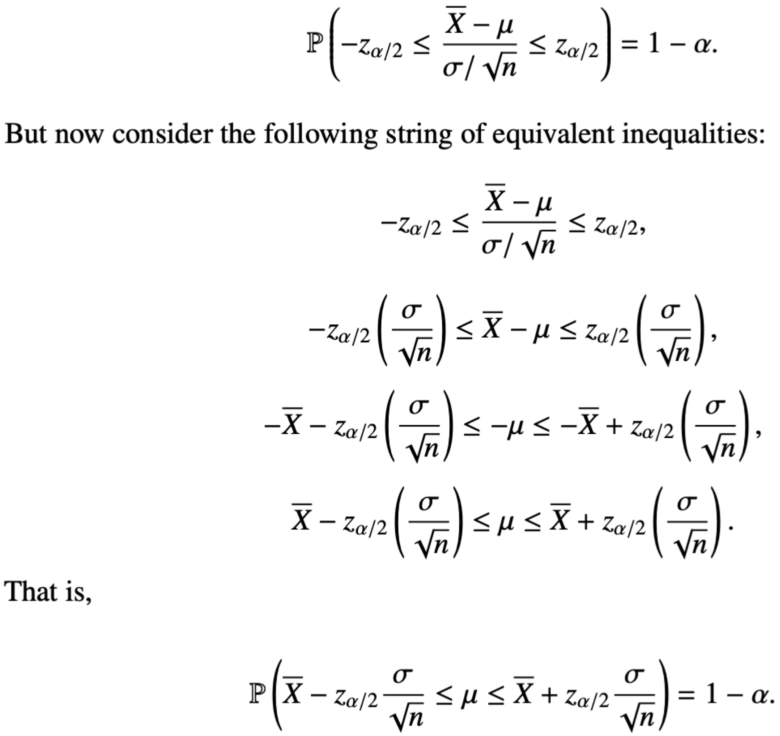
Dr. Farshid Alizadeh-Shabdiz

Spring 2021

Confidence interval for means

We are given X_1, X_2, ..., X_n that are an *SRS*(*n*) from a norm(mean = μ, sd = σ) distribution, where μ is unknown. We know that we may estimate μ with X, and we have seen that this estimator is the MLE. But how good is our estimate? We know that $(X - \mu) / (\sigma / \sqrt{n})$ ~ norm(mean = 0, sd = 1).

For a big probability 1 − α, for instance, 95%, we can calculate the quantile $z_{\alpha/2}$.



$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

But now consider the following string of equivalent inequalities:

$$-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \leq z_{\alpha/2},$$

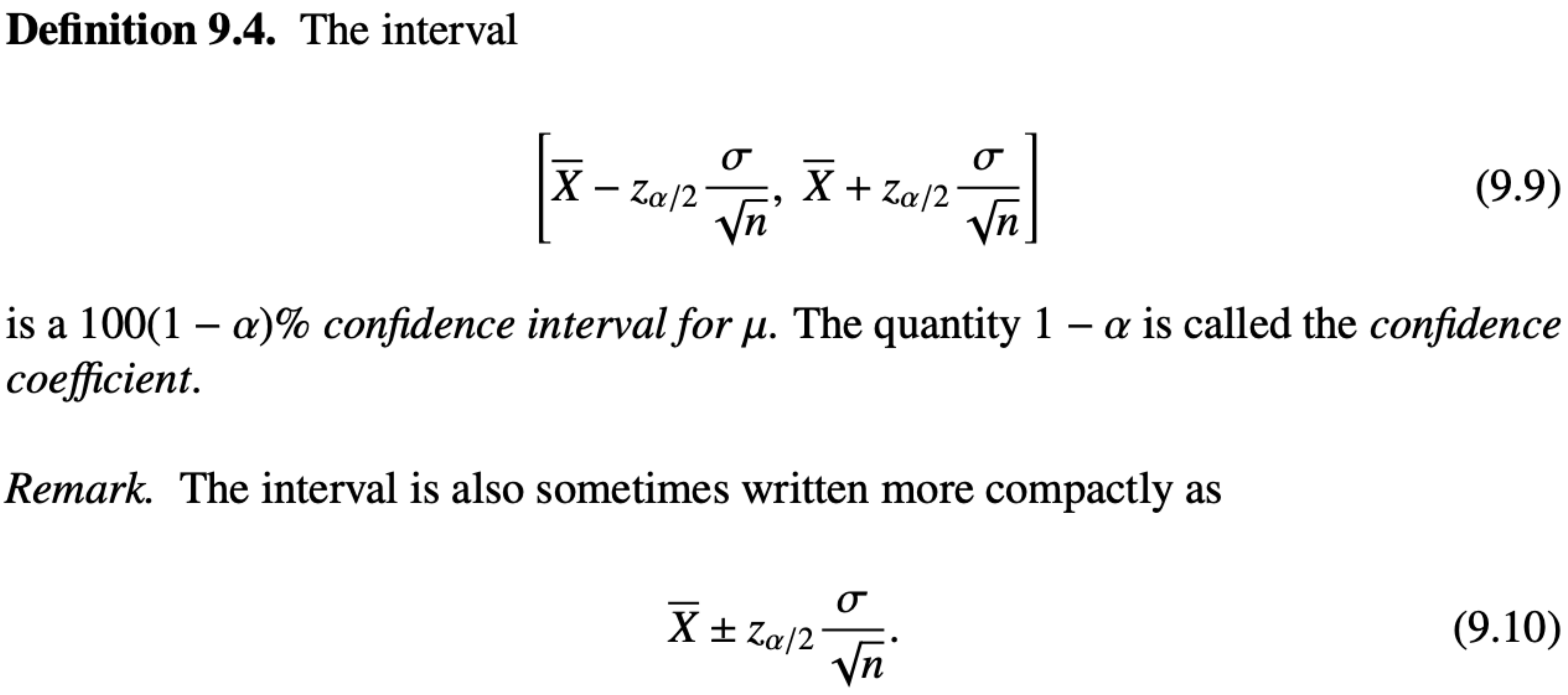
$$-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \overline{X} - \mu \leq z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),$$

$$-\overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq -\mu \leq -\overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),$$

$$\overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).$$

That is,

$$\mathbb{P}\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$



Definition 9.4. The interval

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{9.9}$$

is a $100(1-\alpha)\%$ *confidence interval for* μ. The quantity $1-\alpha$ is called the *confidence coefficient*.

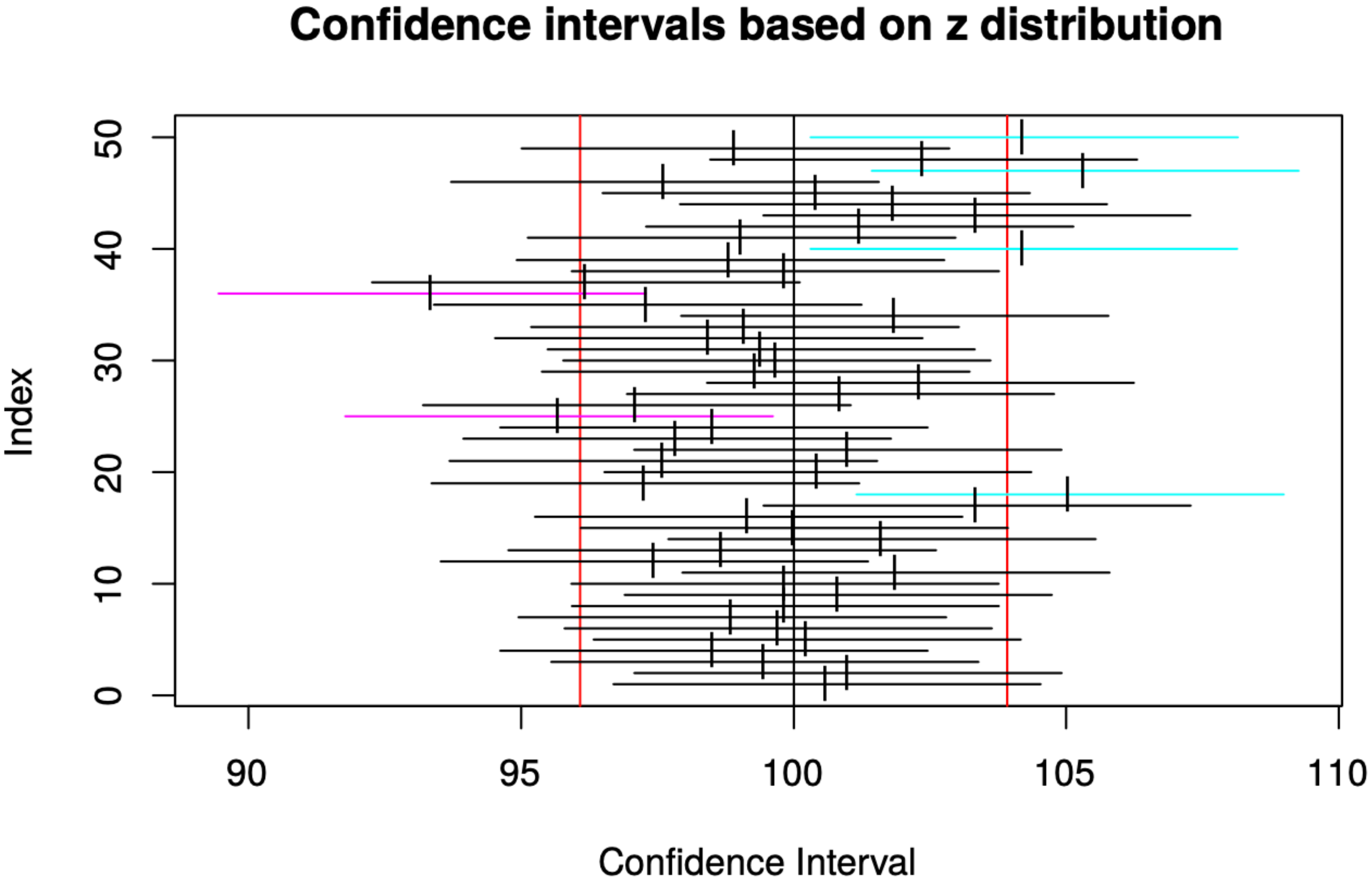
Remark. The interval is also sometimes written more compactly as

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{9.10}$$

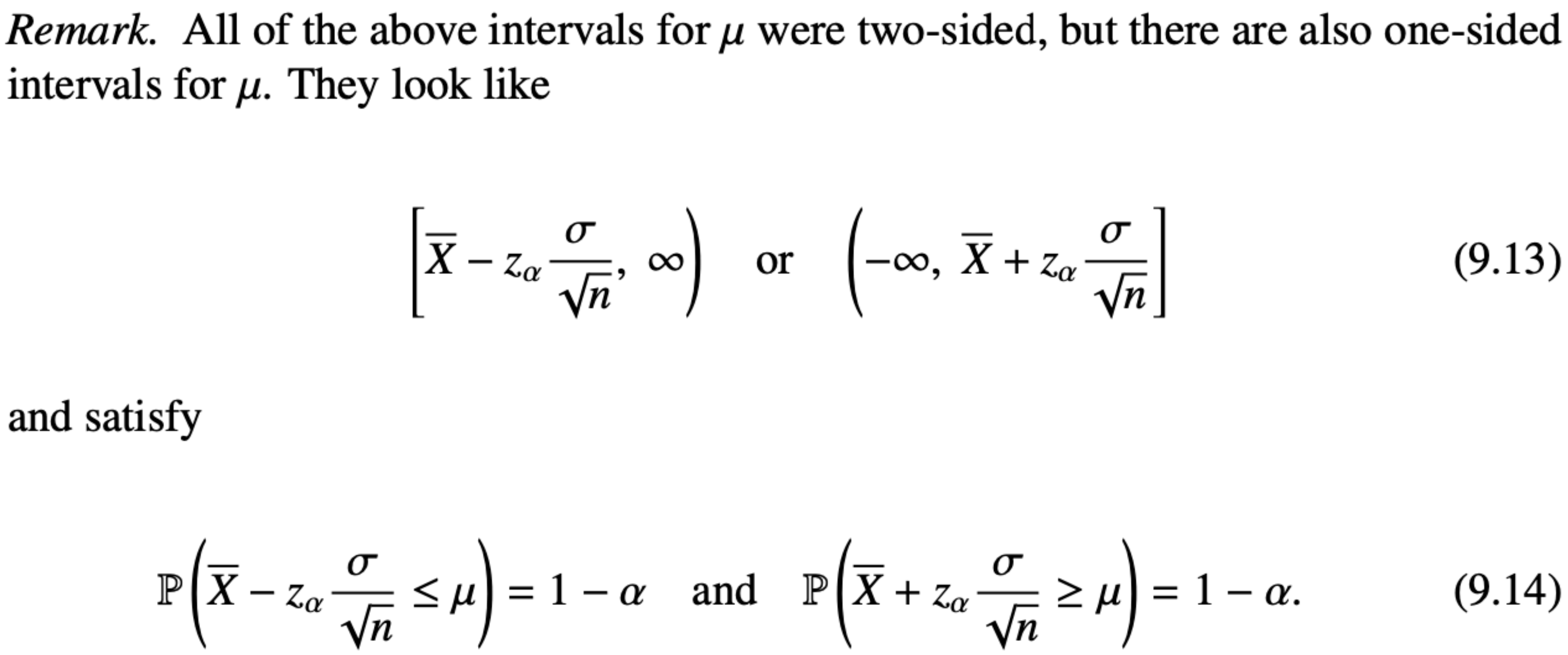
Meaning of Confidence Interval

- In the demonstration, the parameter corresponds to the chalk, the sheet of paper corresponds to the confidence interval, and the random experiment corresponds to dropping the sheet of paper. The percentage of the time that we are successful *exactly* corresponds to the *confidence coefficient*. That is, if we use a 95% confidence interval, then we can say that, in the long run, approximately 95% of our intervals will cover the true parameter (which is fixed, but unknown).

Example below shows simulation of 50 sample of size 25.



Two-sided vs one sided Confidence Interval



Remark. All of the above intervals for μ were two-sided, but there are also one-sided intervals for μ. They look like

$$\left[\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}},\ \infty\right) \quad \text{or} \quad \left(-\infty,\ \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right] \tag{9.13}$$

and satisfy

$$\mathbb{P}\left(\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \le \mu\right) = 1 - \alpha \quad \text{and} \quad \mathbb{P}\left(\overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}} \ge \mu\right) = 1 - \alpha. \tag{9.14}$$

Confidence Interval based on Sample

Remark. What if σ is unknown? We instead use the interval

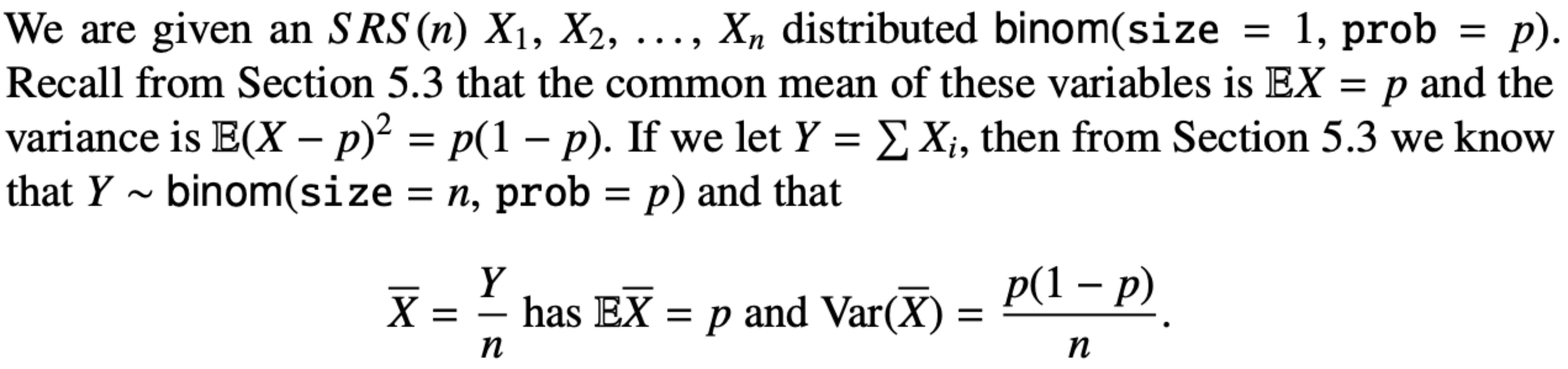
$$\overline{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}},$$

where *S* is the sample standard deviation.

- If *n* is large, then *X* will have an approximately normal distribution regardless of the underlying population (by the CLT) and *S* will be very close to the parameter σ (by the SLLN – strong low of large numbers); thus ~ 100(1 – α)% confidence of covering μ.
- If *n* is small, then
  ∘ If the underlying population is normal then we may replace *z*α/2 with
- $t_{\alpha/2}$(df = *n* – 1).
- The resulting 100(1 – α)% confidence interval is

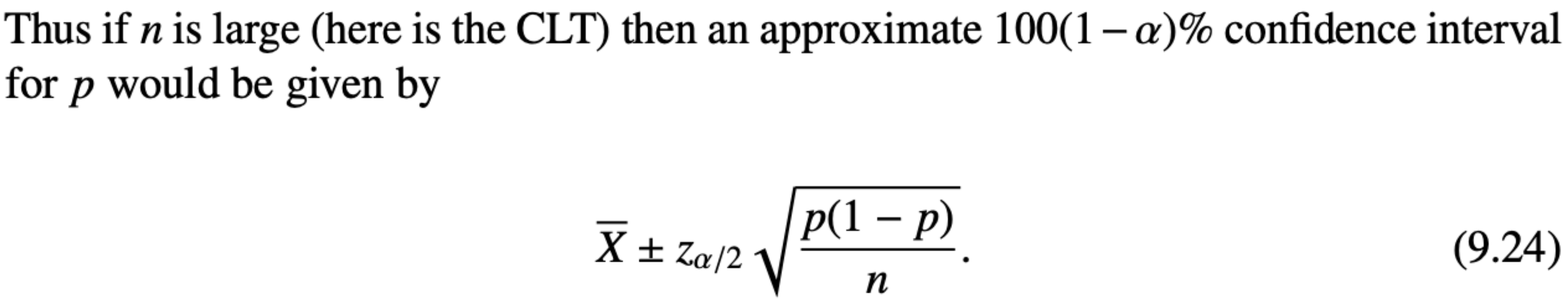
$$\overline{X} \pm t_{\alpha/2}(\mathrm{df} = n - 1)\frac{S}{\sqrt{n}}.$$

Confidence intervals for proportion



We are given an $SRS(n)$ $X_1, X_2, \ldots, X_n$ distributed binom(size $= 1$, prob $= p$). Recall from Section 5.3 that the common mean of these variables is $\mathbb{E}X = p$ and the variance is $\mathbb{E}(X-p)^2 = p(1-p)$. If we let $Y = \sum X_i$, then from Section 5.3 we know that $Y \sim$ binom(size $= n$, prob $= p$) and that

$$\overline{X} = \frac{Y}{n} \text{ has } \mathbb{E}\overline{X} = p \text{ and } \mathrm{Var}(\overline{X}) = \frac{p(1-p)}{n}.$$

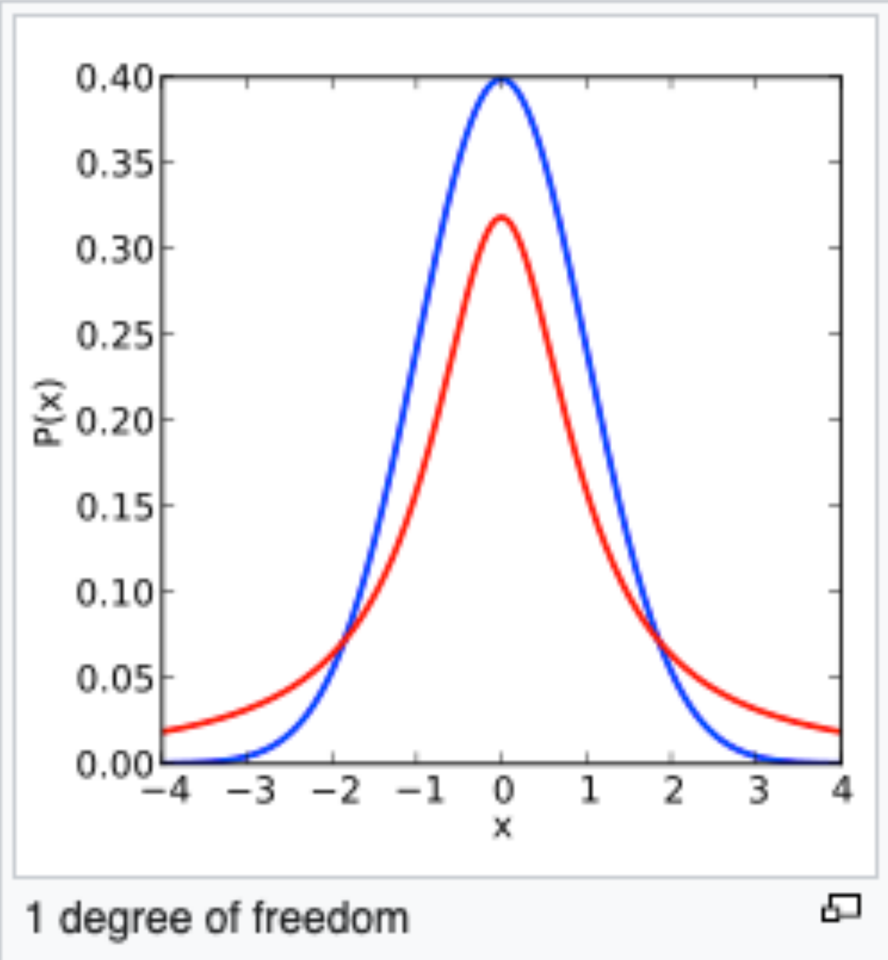


Thus if n is large (here is the CLT) then an approximate $100(1-\alpha)\%$ confidence interval for p would be given by

$$\overline{X} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}. \tag{9.24}$$

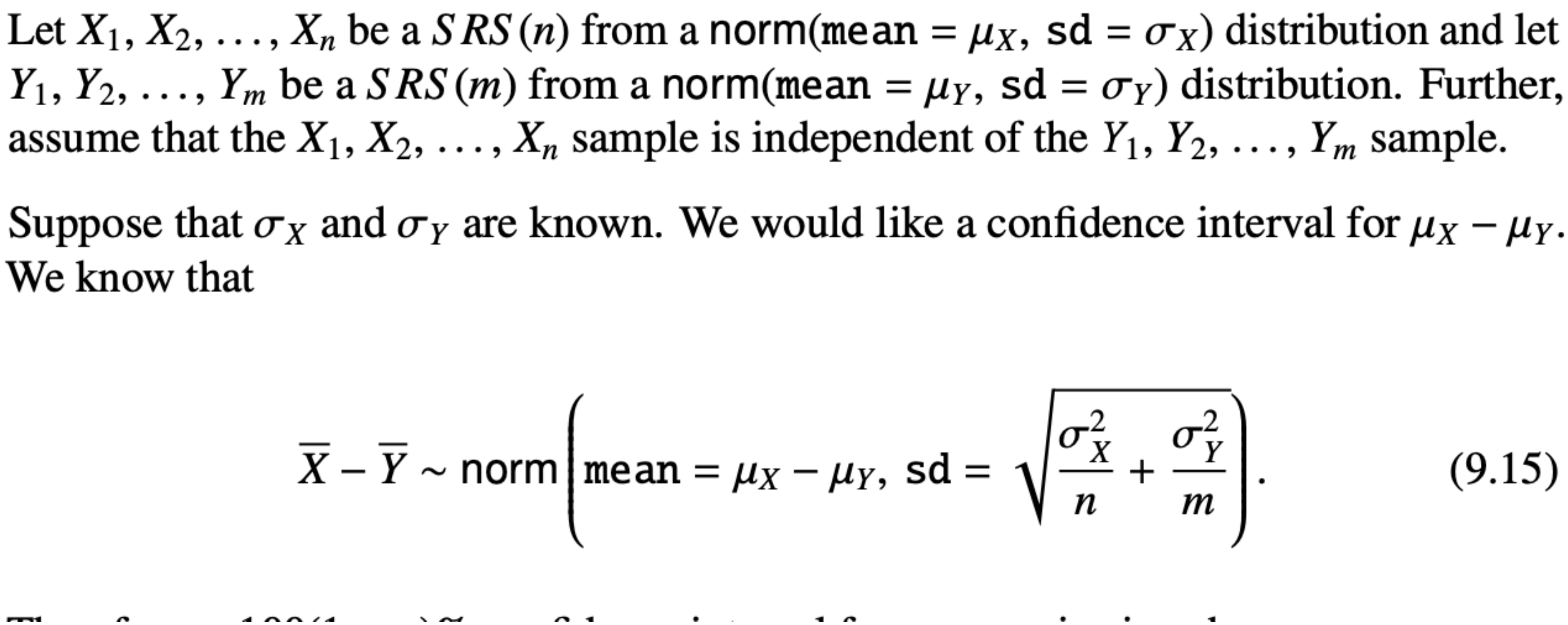
What number of samples is large enough

- Rule of thumb:
  - N>=30 => use Gaussian Distribution
  - N<30 => Use t-distribution or student’s t-distribution. As shown here, t-dist has a heavy tail



1 degree of freedom

Confidence interval for differences in mean

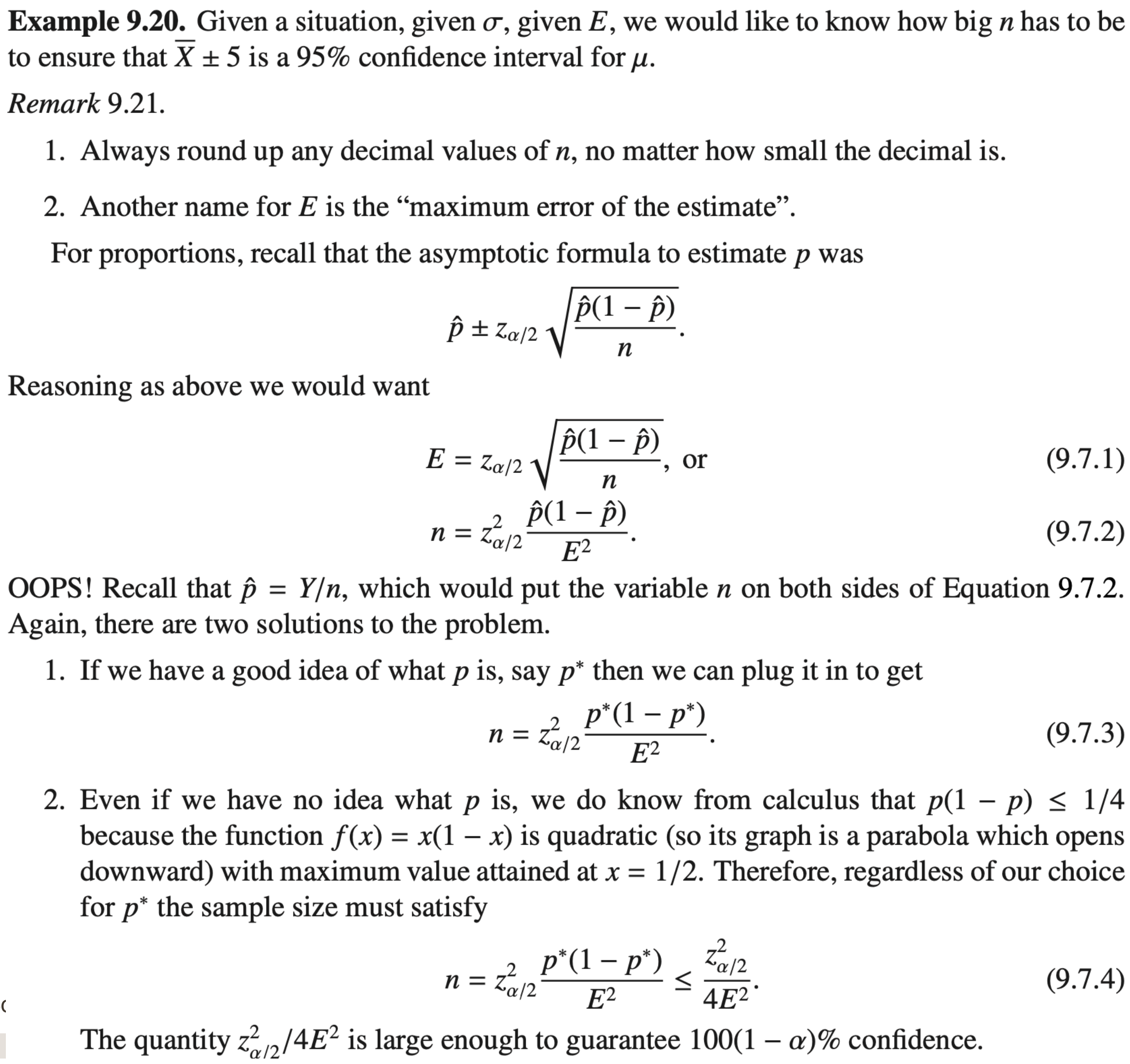


Let $X_1, X_2, \ldots, X_n$ be a $SRS(n)$ from a $\mathsf{norm}(\mathtt{mean} = \mu_X,\ \mathsf{sd} = \sigma_X)$ distribution and let $Y_1, Y_2, \ldots, Y_m$ be a $SRS(m)$ from a $\mathsf{norm}(\mathtt{mean} = \mu_Y,\ \mathsf{sd} = \sigma_Y)$ distribution. Further, assume that the $X_1, X_2, \ldots, X_n$ sample is independent of the $Y_1, Y_2, \ldots, Y_m$ sample.

Suppose that σ_X and σ_Y are known. We would like a confidence interval for $\mu_X - \mu_Y$. We know that

$$\overline{X} - \overline{Y} \sim \mathsf{norm}\left(\mathtt{mean} = \mu_X - \mu_Y,\ \mathsf{sd} = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right). \tag{9.15}$$

Sample Size and Margin of Error



Example 9.20. Given a situation, given σ, given E, we would like to know how big n has to be to ensure that $\overline{X} \pm 5$ is a 95% confidence interval for μ.

Remark 9.21.

1. Always round up any decimal values of n, no matter how small the decimal is.

2. Another name for E is the "maximum error of the estimate".

For proportions, recall that the asymptotic formula to estimate p was

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Reasoning as above we would want

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \text{ or} \tag{9.7.1}$$

$$n = z^2_{\alpha/2}\frac{\hat{p}(1-\hat{p})}{E^2}. \tag{9.7.2}$$

OOPS! Recall that $\hat{p} = Y/n$, which would put the variable n on both sides of Equation 9.7.2. Again, there are two solutions to the problem.

1. If we have a good idea of what p is, say p^* then we can plug it in to get

$$n = z^2_{\alpha/2}\frac{p^*(1-p^*)}{E^2}. \tag{9.7.3}$$

2. Even if we have no idea what p is, we do know from calculus that $p(1-p) \leq 1/4$ because the function $f(x) = x(1-x)$ is quadratic (so its graph is a parabola which opens downward) with maximum value attained at $x = 1/2$. Therefore, regardless of our choice for p^* the sample size must satisfy

$$n = z^2_{\alpha/2}\frac{p^*(1-p^*)}{E^2} \leq \frac{z^2_{\alpha/2}}{4E^2}. \tag{9.7.4}$$

The quantity $z^2_{\alpha/2}/4E^2$ is large enough to guarantee $100(1-\alpha)\%$ confidence.

Hypothesis testing

Example

1. Example of selecting a card
2. The mean of the defects is at most 2%
3. Most CEOs are male
2. The average starting salary of BU students: $53k
3. Drug causes immunization at least 76% of the time

Terminology

- Null hypothesis – H_0
- Alternative hypothesis – H_1
- Processes/procedure
  - Collect random samples
  - Construct confidence interval for P
  - If the confidence interval covers H0, we “fail to reject H0”. Otherwise, “reject H0”.
- The rejection region is called “critical region”.
  - One sided : H1>= or H1<=
  - Two sided
- Critical Value

Error types

- Error Type I: if we reject H0 by mistake – H1 gets approved
- Error Type II: if we fail to reject H0 by mistake – don’t know about H1

- Type I errors are usually considered worse.

- Null hypothesis: H0 always contains 0
- Alternative hypothesis: H1 always is (> or < or !=)
- Prove H1 by rejecting H0.

Examples of hypothesis Testing

- Examples with (mean and proportion)

1. The mean of the defects is at most 2%

   - H0: defect <= 2%

   - H1: defect > 2%

2. Most CEOs are male

   - H0: p = 0.5

   - H1: p > 0.5

Examples of hypothesis Testing

- Examples with (mean and proportion)

2. The average income of BU students is $53k
   - H0: avg = $53k
   - H1: avg != $53k
3. Drug causes immunization at least 76% of the time
   - H0: immunization <=76%
   - H1: Immunization < 76%

Test statistics

- Proportion P:
- $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$
- Mean μ:
- $Z = \frac{\hat{x}-\mu}{\sigma/\sqrt{n}}$ (if population sd in known) or
- $Z = \frac{\hat{x}-\mu}{S/\sqrt{n}}$ (if population sd is unknown)

- One sided
    - Left
- Two sided

Example 10.1 of Intro to Stats CS544

- **Example 10.1.** We have a machine that makes widgets.
  • Under normal operation, about 0.10 of the widgets produced are defective.
- • Go out and purchase a torque converter.
  • Install the torque converter, and observe $n = 100$ widgets from the machine.
- • Let Y = number of defective widgets observed.
- If
  • $Y = 0$, then the torque converter is great!
  • $Y = 4$, then the torque converter seems to be helping.
  • $Y = 9$, then there is not much evidence that the torque converter helps.
  - $Y = 17$, then throw away the torque converter.

Example 10.1

- Let p denote the proportion of. Before the installation of the torque converter p was 0.10. Then we installed the torque converter. Did p change? Our method is to observe data and construct a 95% confidence interval for p,
- If the confidence interval is
  - [0.01, 0.05], then we are 95% confident that $0.01 \leq p \leq 0.05$, so there is evidence that the torque converter is helping.
  - [0.15, 0.19], then we are 95% confident that $0.15 \leq p \leq 0.19$, so there is evidence that the torque converter is hurting.
  - [0.07, 0.11], then there is not enough evidence to conclude that the torque converter is doing anything at all, positive or negative.

Example – US Berkeley Admission

- Suppose p = the proportion of students who are admitted to the graduate school of the University of California at Berkeley, and suppose that a public relations officer boasts that UCB has historically had a 40% acceptance rate for its graduate school. Consider the data stored in the table UCBAdmissions from 1973. Assuming these observations consti- tuted a simple random sample, are they consistent with the officer's claim, or do they provide evidence that the acceptance rate was significantly less than 40%? Use an $\alpha = 0.01$ significance level.
- Our null hypothesis in this problem is $H0 : p = 0.4$ and the alternative hypothesis is $H1 : p < 0.4$.

Example

- P-value example 1
  - Alpha = 0.05 or 95%
  - H1: proportion > 0.25 – you get a test statistics of Z=1.18
  - Solution:
  - Traditional: Z = 1.685 since H1 is “right hand tail”. 1.18<1.685 => failed to reject H0
  - P-value: Area corresponding to Z = 1.18 [pnorm(1.18,mean=0,sd=1)] is

  1-0.8810 = 0.1190

  0.1190 > 0.05 => failed to reject H0
- P-value example: a = 0.05 , H1!=0.25 and Z=2.3
  - Area = 0.0096 on either side
  - 0.0096 < 0.025 or (0.0096*2) < 0.05
  - Reject Ho
- P-value example: a = 0.05, H1!= 0.8, Z=2.3 (area of z=2.3 is 0.0107)

Example - Calculating P-value for an one-sided test

A gym is interested in whether a 6-week weight loss training program they launched has been successful in helping their clients lose weight. To assess this, they took a sample of 30 participants.

They are interested in testing the following hypotheses:

- **$H_0 : \mu = 0$ (there is no effect on weight change of program participants)**
- **$H_1 : \mu < 0$ (program participants lose weight on average)**

Suppose we know that for the general population, the standard deviation of changes in weights over a six-week interval is 6 pounds.

The sample mean of the change in weight for the 30 participants in the sample was -2.98 pounds.

Calculate the value of the test statistic and the associated p-value.

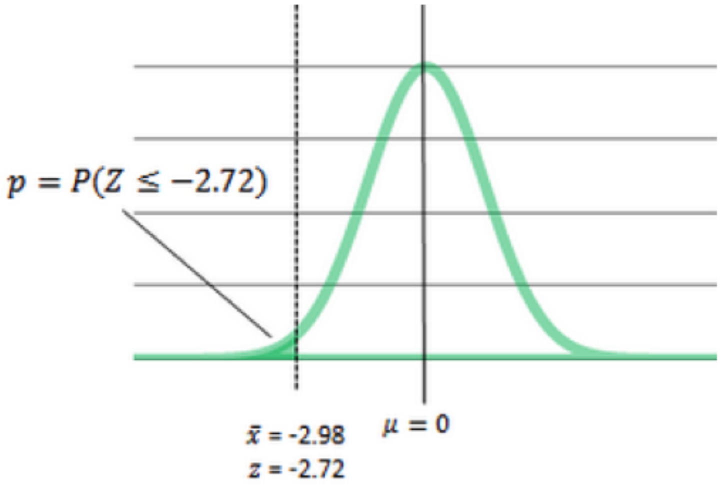
Calculating the value of the test statistic and p-value

$\sigma = 6$, $n = 30$, $\bar{x} = -2.98$, $\mu_{H_0} = 0$

The p-value is the probability that the test statistic is -2.72 or more extreme, which is the probability that $Z \leq -2.72$. Using the standard normal table, we can calculate: $P = P(Z \leq -2.72) = 0.0033$

$$z = \frac{\bar{x} - \mu_{H_0}}{\frac{\sigma}{\sqrt{n}}} = \frac{-2.98 - 0}{\frac{6}{\sqrt{30}}} \approx \frac{-2.98}{1.0954} \approx -2.72$$

This is a small p-value. It appears that the sample mean (= -2.98) is highly unlikely to have occurred if the true population mean $\mu = 0$. **Thus we have strong evidence against the null hypothesis.**



Example: Chemical in Water - Two-Sided

Normal levels of this chemical are 15 parts per million (ppm).

Samples from 50 water sources throughout the county are taken and the levels of this chemical are measured.

They are interested in testing the following hypotheses:

$H_0 : \mu = 15$ (the mean level of the chemical is normal)

$H_1 : \mu \neq 15$ (the mean level of the chemical is abnormal)

Suppose we know that the population standard deviation is 6.2.

The sample mean from the 50 samples was 16.4 ppm.

Calculate the value of the test statistic and the associated p-value.

Example: Chemical in Water - Two-Sided

Givens are:

$\bar{x} = 16.4$, $\mu_{H_0} = 15$, $\sigma = 6.2$, $n = 50$

Now, we can just plug these values in to calculate the value of z.

$$z = \frac{\bar{x} - \mu_{H_0}}{\frac{\sigma}{\sqrt{n}}} = \frac{16.4 - 15}{\frac{6.2}{\sqrt{50}}} \approx \frac{1.4}{0.8768} \approx 1.60$$

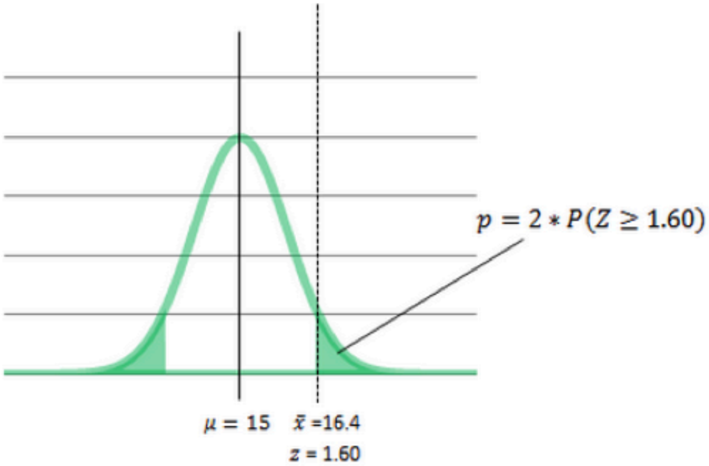
The p-value is the probability that the test statistic is 1.60 or more extreme.

$\boldsymbol{H_0 : \mu = 15}$

$\boldsymbol{H_1 : \mu \neq 15}$

Example: Chemical in Water - Two-Sided

That is, the p-value is the probability that $Z \geq 1.60$ or $Z \leq -1.60$.



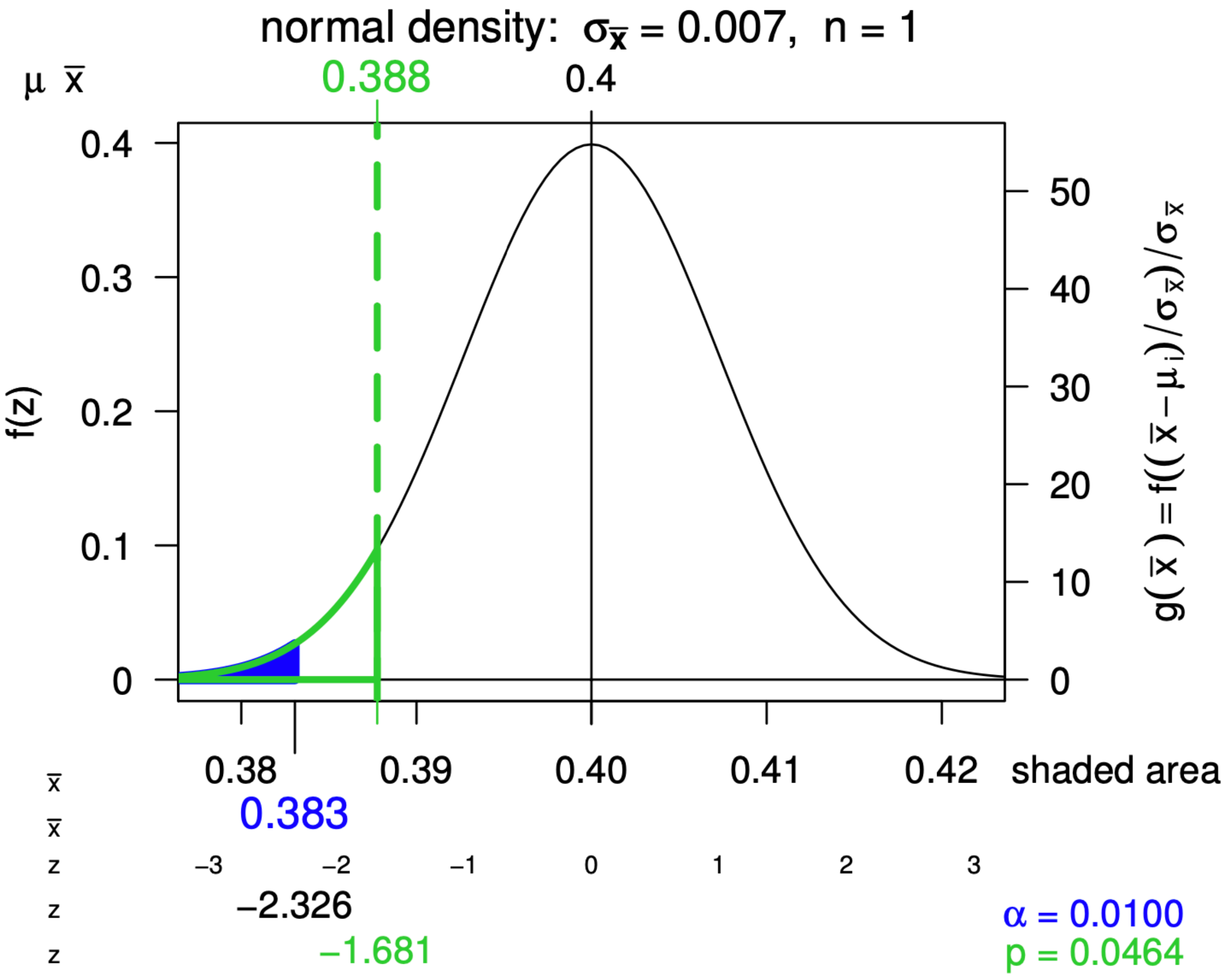
$$P = P(Z \leq -1.60 \text{ or } Z \geq 1.60) = P(Z \geq 1.60) + P(Z \leq -1.60)$$

$$= 2 \times P(Z \geq 1.60) = 2 \times 0.0548 = 0.1096$$

It appears that the sample mean that we observed ($\bar{x} = 16.4$) is moderately likely to have occurred if the true population mean was 15 ppm (if $\mu = 15$).

We don't have strong evidence against the null hypothesis.

P-value or Observed Significance Level or Significance Level


normal density: $\sigma_{\overline{x}} = 0.007$, n = 1
μ x̄
f(z)
$g(\overline{x}) = f((\overline{x} - \mu_i)/\sigma_{\overline{x}})/\sigma_{\overline{x}}$
x̄
shaded area
x̄
z
z
z
α = 0.0100
p = 0.0464