P8130: Biostatistical Methods I

Lecture 6: Methods of Inference for One-Mean

Cody Chiuzan, PhD
Department of Biostatistics
Mailman School of Public Health (MSPH)

Lecture 5: Recap

- Continuous random variables and probability distributions
- Uniform distribution: definition and statistical properties
- Normal distribution: definition and statistical properties

Lecture 6: Outline

- Sampling Distribution
- Central Limit Theorem (additional simulations in Recitation 2)
- Estimation/Confidence Interval for One-Sample Mean
- Hypothesis Testing for One-Sample Mean

Sampling Distribution

- The usual way to obtain information regarding a population parameter such as: $\mu, \sigma^2, or\ p$, is by selecting a sample from a population and compute a statistic.

  - The observed value depends on the particular sample selected, and it varies from sample to sample => *sampling variability*

  - The distribution of the values of the statistic is called *sampling distribution.*

Estimation of the Mean of a Distribution

- A common estimator of a population mean μ is the sample mean:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Even though there is only one underlying population, each sample has its own mean (and variance).

- Is the sample mean a 'good' estimator of the population mean?
  - Is the sample mean an unbiased estimator of the population mean?
  - Is the standard deviation of the estimator small?

Sampling Distribution of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ represent a simple random sample of size n from a population with mean μ and variance σ^2. The following hold true:

1. The mean of the sampling distribution of the sample mean is equal to the population mean:

$$\mu_{\bar{X}} = \mu$$

2. The variance of the sampling distribution of the sample mean is the population variance divided by the size of the sample:

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

3. Provided that n is large enough ($n \geq 30$) or the underlying distribution is normal, then the shape of the sampling distribution is approximately normal:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

3 is known as the **Central Limit Theorem (CLT).**

Recall: Statistical Inference

Questions:

1. How is a specific random sample used to estimate the population parameters of an underlying distribution for a population?
2. How to use the information we have from the sample to ‘infer’ properties of the population?

Statistical inference includes:

- Point Estimation
- Interval Estimation
- Hypothesis Testing
- Prediction

Point Estimation

A point estimate is a single number computed from the sample, that can be regarded as a plausible value of the population parameter (characteristics).

Example: the computed value of the sample mean $\bar{X}$ provides a point estimate of the population mean μ (our best single *guesstimate*).

Because of *sampling variability,* rarely is the point estimate exactly equal to the true parameter.

Solution: construct a confidence interval (CI) that contains plausible values for the population parameter.

Interval Estimation: One-Sample Mean

A $100(1-\alpha)\%$ confidence interval for the population mean with known variance σ^2:

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Where:

$\bar{X}$ is the point estimate (sample mean)

$\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean (measure of variability)

$1-\alpha$ is the desired level of confidence

Because the sample mean $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, it follows that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

Interval Estimation: One-Sample Mean

Derive the 95% confidence interval for a population mean, known variance:

$$P\left(z_{lower} < z < z_{upper}\right) = 0.95 \rightarrow P(-1.96 < z < 1.96) = 0.95$$

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

Multiplying across by -1, $\sigma/\sqrt{n}$, and then adding $\bar{X}$ leads to:

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Confidence Interval: Illustration

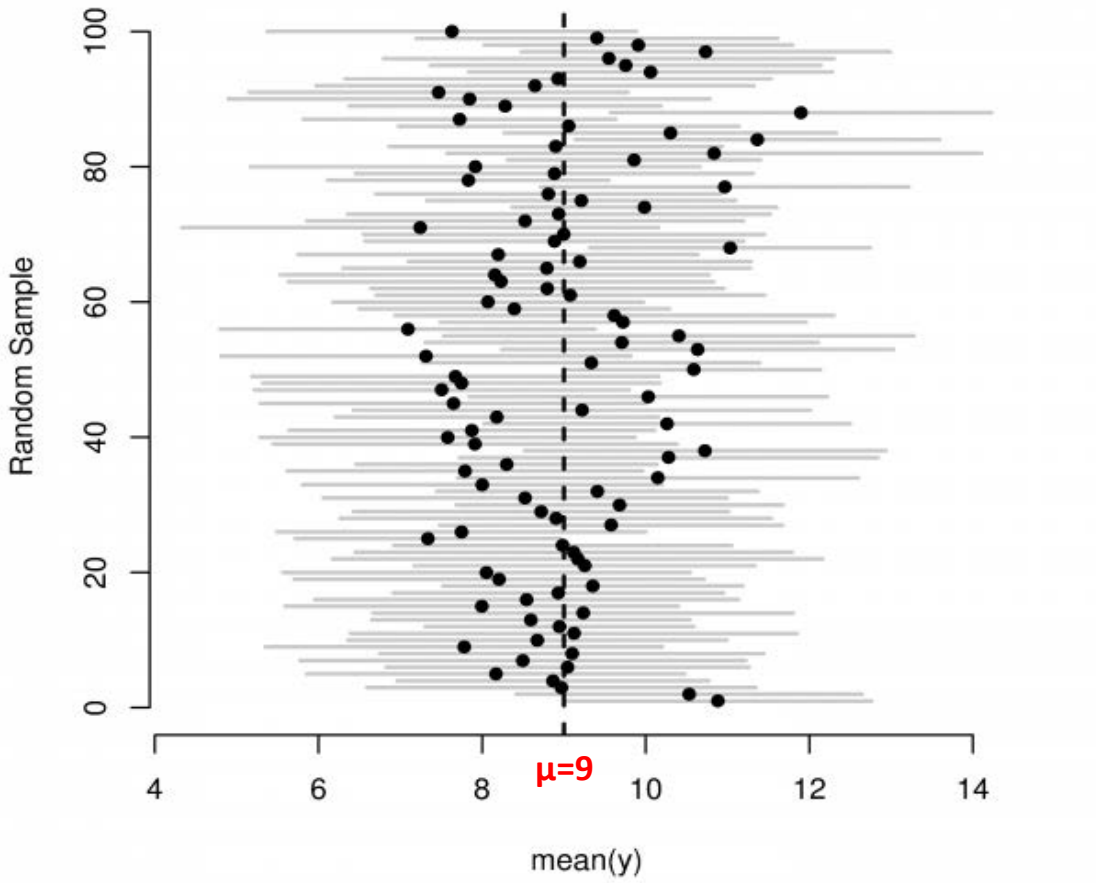
The graph on the right shows 95% CIs computed from 100 random samples, each with n=20, taken from an underlying distribution: Y$\sim N(9, 5)$.

The sample means are denoted by points and the confidence limits by gray lines.

What do you notice?

- Some intervals do not even contain the true population mean **(μ=9).**
- Because μ is fixed, the probability that any specific interval contains it is either 0 or 1.

Point Estimates and 95% CIs for 100 random samples



95% CI: Interpretations

1. Over the collection of all 95% confidence intervals that could be constructed from repeated samples of size *n,* 95% of them will contain the true population mean.

2. We are 95% confident that the population mean lies between the lower and the upper limits of the interval.

Questions:

- If the confidence level increases, the width of the CI …
- If the population variance increases, the width of the CI …
- If the sample size increases, the width of the CI …

95% CI: Example 1

Let's assume that cholesterol levels are normally distributed. Ten men are randomly selected and their average serum cholesterol was 175mg/dL. Given the population standard deviation of 15 mg/dL, construct a 95% CI for the true mean serum cholesterol level.

In class practice:

What if we only know the sample standard deviation?

Interval Estimation: One-Sample Mean

A $100(1-\alpha)\%$ confidence interval for the population mean with unknown variance σ^2:

$$\bar{X} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}$$

Where:

$\bar{X}$ is the point estimate (sample mean)

$\frac{s}{\sqrt{n}}$ is the estimated standard error, $s = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)}$

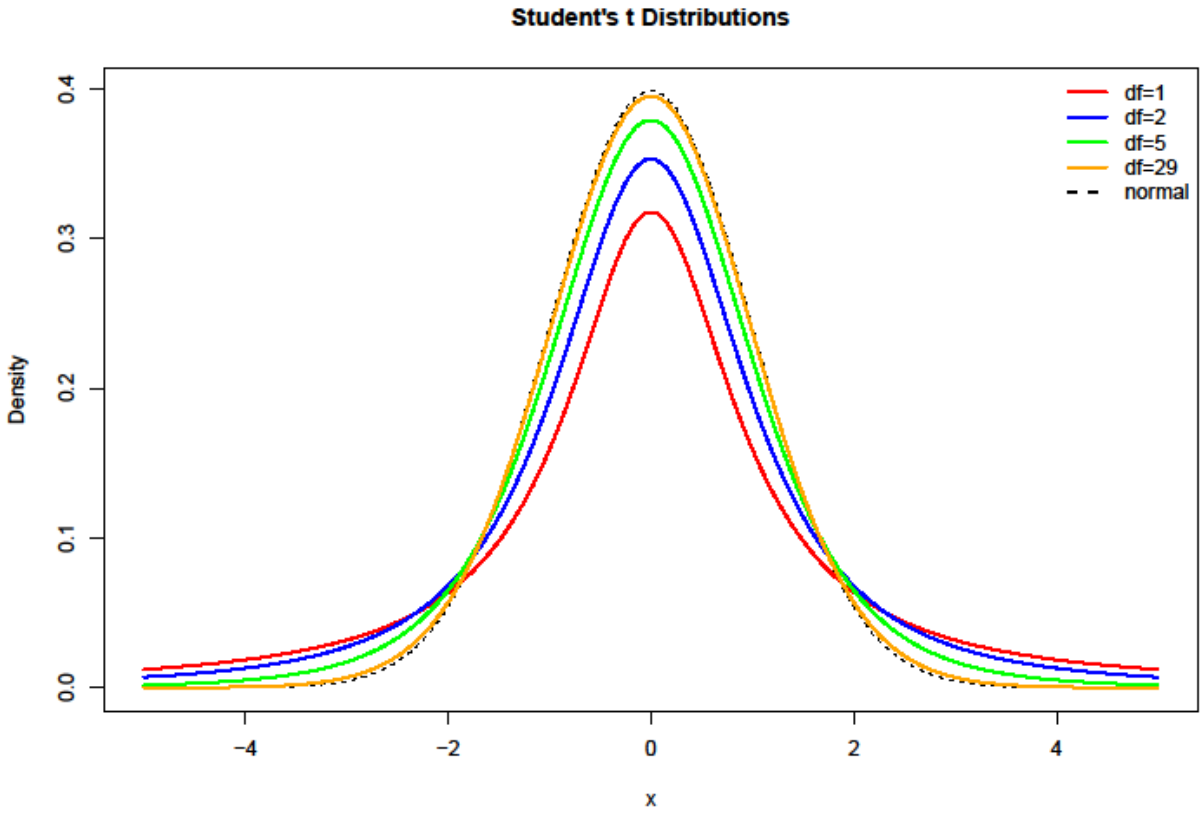
$t_{n-1,1-\alpha/2}$ is the percentile of the t-distribution with (n-1) degrees of freedom

Notice that in this case:

$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{(n-1)}$, t distribution with (n-1) degrees of freedom

Some notes on *t*-distribution

- Continuous distribution, symmetric about zero
- Heavier tails than the unit normal
- Depends on only one parameter called *degrees of freedom (df)*
- More variable because it depends on estimating both μ and σ
- As the number of *dfs* increases, the corresponding sequence of *t*-curves approaches the unit normal curve



95% CI: Example 2

Let's assume that cholesterol levels are normally distributed. Ten men are randomly selected and their average serum cholesterol was 175mg/dL. Given the sample standard deviation of 15 mg/dL, construct a 95% CI for the true mean serum cholesterol level.

In class practice:

Variance Estimation

Let $X_1, X_2, \ldots, X_n$ represent a simple random sample of size n from a population with mean μ and variance σ^2.

The sample variance s^2 is an unbiased estimator of σ^2:

$$E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)\right) = \sigma^2$$

Suppose that s^2 is the sample variance of a random sample from a normal distribution with variance σ^2. It can be shown that:

$s^2 \sim \frac{\sigma^2 \chi^2_{n-1}}{n-1} \rightarrow \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$, where χ^2_{n-1} is the chi-squared distribution with (n-1) df.

CI for σ^2 using χ^2 distribution

A $100(1-\alpha)\%$ confidence interval for the population variance σ^2 is given by:

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \quad \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right).$$

Note: this interval is valid under the assumption that the underlying distribution is normal.

Hypothesis Testing

Hypothesis testing provides a framework for making decisions on an objective basis rather than on a subjective basis by simply looking at the data.

Objective basis – evaluation of the relative probabilities of different hypotheses.

The null hypothesis (H_0): hypothesis to be tested

vs.

The alternative hypothesis ($H_1\ or\ H_a$): hypothesis contradicting the null

Note that our decisions will always be with respect to the null hypothesis.

Hypothesis Testing: Motivation

Court room example - assume innocence until proven guilty.

H_0: the accused is innocent

vs.

H_1: the accused is <u>not</u> innocent

<u>Decisions:</u>

- There is not enough evidence to show that the hypothesis of innocence is false (fail to reject H_0).

OR

- There is enough evidence to show that the accused is not innocent (reject H_0).

Hypothesis Testing: Decision Table

	Decision	
Truth	Fail to reject H_0	Reject H_0
H_0 is true	**Correct Decision**	**Type I Error (α)**
H_0 is false	**Type II Error (β)**	**Correct Decision**

Type I error is also called the significance of a test: P(reject $H_0|H_0$ is true) (e.g., 0.05)
Type II error : P(fail to reject $H_0|H_0$ is false)
Power of a test: 1– type II error = P(reject $H_0|H_0$ is false) (e.g., 0.80, 0.90)

Type I error and type II are inverse proportional. Type I increases -> type II decreases

How to Conduct Hypothesis Testing

1. State the question of interest!
2. Set up the null/alternative hypotheses and the significance level
3. Clearly state the statistical methodology to be used and assumptions - why?
4. Collect the data
5. State the test statistic and determine the critical region/p-value
6. Interpret the findings in the context of the question/problem
7. Draw conclusions and compare the results to other findings (if available)

One-Sample, 1-sided Tests

Tests for the Mean of a Normal Distribution with Known Variance

H_0: $\mu = \mu_0$ vs H_1: $\mu < \mu_0$

With significance level α pre-specified, compute the test statistic:

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Reject H_0: if $z < -z_\alpha$

Fail to reject H_0: if $z \geq -z_\alpha$

H_0: $\mu = \mu_0$ vs H_1: $\mu > \mu_0$

With significance level α pre-specified, compute the test statistic:

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Reject H_0: if $z > z_\alpha$

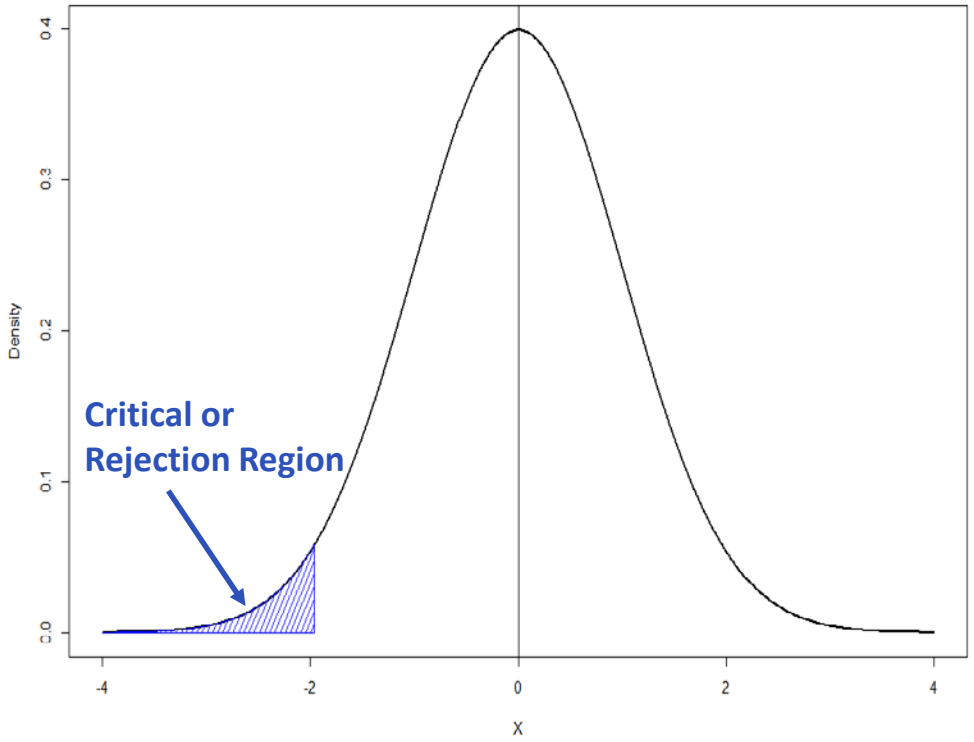
Fail to reject H_0: if $z \leq z_\alpha$

z_α is called the critical value and it's 'fixed.'
Values can be found in tables or using software.

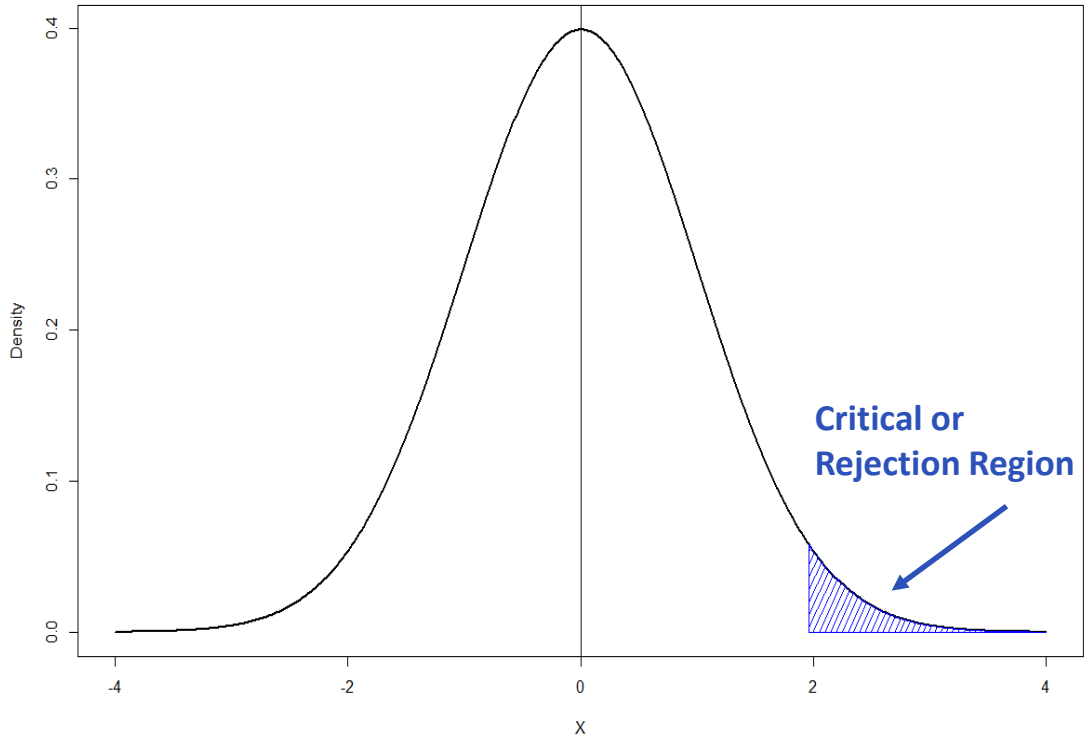
One-Sample, 1-sided Tests

Tests for the Mean of a Normal Distribution with Known Variance

$H_0: \mu = \mu_0$ vs $H_1: \mu < \mu_0$



$H_0: \mu = \mu_0$ vs $H_1: \mu > \mu_0$



One-Sample, 2-sided Tests

Tests for the Mean of a Normal Distribution with Known Variance

H_0: $\mu = \mu_0$ vs H_1: $\mu \neq \mu_0$

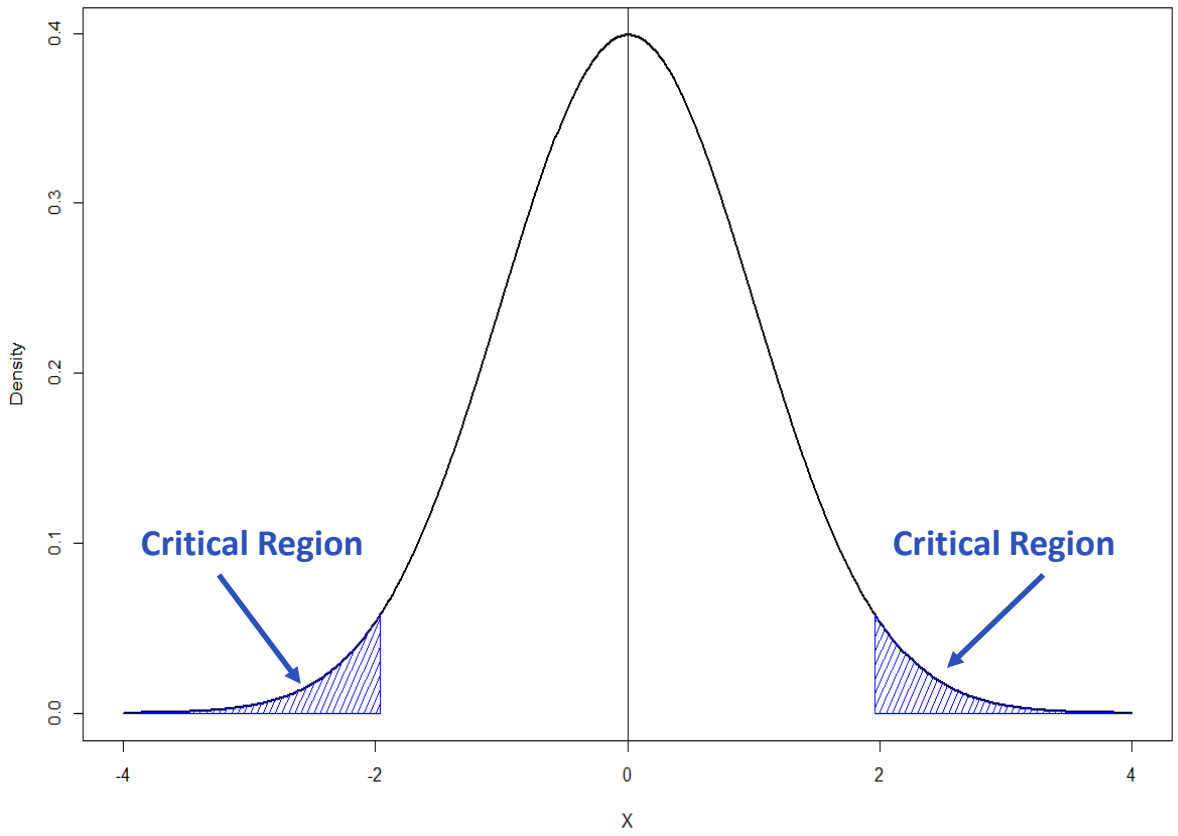
With significance level α pre-specified, compute the test statistic:

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Reject H_0: if $|z| > z_{1-\alpha/2}$

Fail to reject H_0: if $|z| \leq z_{1-\alpha/2}$

$z_{1-\alpha/2}$ is called the critical value and it's 'fixed.'



One-Sample, 1-sided Tests

Tests for the Mean of a Normal Distribution with Unknown Variance

H_0: $\mu = \mu_0$ vs H_1: $\mu < \mu_0$

With significance level α pre-specified, compute the test statistic:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

Reject H_0: if $t < -t_{n-1,\alpha}$

Fail to reject H_0: $t \geq -t_{n-1,\alpha}$

H_0: $\mu = \mu_0$ vs H_1: $\mu > \mu_0$

With significance level α pre-specified, compute the test statistic:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

Reject H_0: if $t > t_{n-1,\alpha}$

Fail to reject H_0: $t \leq t_{n-1,\alpha}$

$t_{n-1,\alpha}$ is called the critical value and it's 'fixed.'

One-Sample, 2-sided Tests

Tests for the Mean of a Normal Distribution with Unknown Variance

H_0: $\mu = \mu_0$ vs H_1: $\mu \neq \mu_0$

With significance level α pre-specified, compute the test statistic:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

Reject H_0: if $|t| > t_{n-1,1-\alpha/2}$

Fail to reject H_0: if $|t| \leq t_{n-1,1-\alpha/2}$

$t_{n-1,1-\alpha/2}$ is called the critical value and it's 'fixed.'

Confidence Intervals vs Hypothesis Test

Suppose we are testing: H_0: $\mu = \mu_0$ vs H_1: $\mu \neq \mu_0$.

H_0 is rejected with a two-sided level α test if and only if the two-sided 100% (1- α) confidence interval for μ does not contain μ_0.

H_0 is not rejected with a two-sided level α test if and only if the two-sided 100% (1- α) confidence interval for μ does contain μ_0.

The CI represents the 'fail to reject' region of the two-sided hypothesis test.

What is the p-value?

A p-value is the probability of observing something as extreme or more extreme when the null hypothesis is true.

- If p-value < α, then we reject the null hypothesis at significance level α.
- A large p-value indicates weak evidence against the null.
  - A p-value=0.4 means that data such as ours would occur 40% of the time if the null is true, so not enough evidence to reject.
- A small p-value indicates strong evidence against the null.
  - A p-value=0.001 means that data such as ours would occur 0.1% of the time if the null is true; so we observe something very unlikely if the null were true.
- P-values should be carefully used and not (ab)used.
  - P=0.05 can be deceiving and have no clinical significance

Defining p-values

The p-values for one-sample, 2-sided, t-test for the mean of a normal distribution is given by:

If the observed value of the test statistic is: $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \leq 0$, then:

$$\text{P-value} = 2 \times \mathrm{P}(t_{n-1} < t | H_0)$$

If the observed value of the test statistic is: $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} > 0$, then:

$$\text{P-value} = 2 \times \mathrm{P}(t_{n-1} \geq t | H_0)$$

Note: same rules apply for the z-test.

Hypothesis Testing: Example

A lab is testing a new drug to reduce the infarct size in patients who have a myocardial infarction (MI) within the past 24h. Suppose we know that in untreated patients, the mean infarct size is 25 units. Further, in 40 patients treated with this drug, the mean infarct size is 16 with a sample standard deviation of 10.

Do treated patients truly have a different infarct size? Assume $\alpha=0.05$.

In class practice.

Readings

- Rosner, *Fundamentals of Biostatistics:* Chapters 6 and 7
- Read ‘P-values and data dredging’ article posted on the course site
- Sample size determination in Recitation 3