

Basic Concepts of Inference

Corresponds to Chapter 6 of
Tamhane and Dunlop

Slides prepared by Elizabeth Newton (MIT)
with some slides by Jacqueline Telford
(Johns Hopkins University) and Roy Welsch (MIT).

“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”
H. G. Wells

Statistical Inference

Deals with methods for making statements about a population based on a sample drawn from the population

Point Estimation: Estimate an unknown population parameter

Confidence Interval Estimation: Find an interval that contains the parameter with preassigned probability.

Hypothesis testing: Testing hypothesis about an unknown population parameter

Examples

Point Estimation: estimate the mean package weight of a cereal box filled during a production shift

Confidence Interval Estimation: Find an interval $[L,U]$ based on the data that includes the mean weight of the cereal box with a specified probability

Hypothesis testing: Do the cereal boxes meet the minimum mean weight specification of 16 oz?

Two Levels of Statistical Inference

- Informal, using summary statistics (may only be descriptive statistics)
- Formal, which uses methods of probability and sampling distributions to develop measures of statistical accuracy

Estimation Problems

- Point estimation: estimation of an unknown population parameter by a single statistic calculated from the sample data.
- Confidence interval estimation: calculation of an interval from sample data that includes the unknown population parameter with a pre-assigned probability.

Point Estimation Terminology

Estimator = the random variable (r.v.) $\hat{\theta}$, a function of the X_i 's
(the general formula of the rule to be computed from the data)

Estimate = the numerical value of $\hat{\theta}$ calculated from the
observed sample data $X_1 = x_1, \dots, X_n = x_n$
(the specific value calculated from the data)

Example: $X_i \sim N(\mu, \sigma^2)$

Estimator = $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ is an estimator of $\mu = \hat{\mu}$

Estimate = $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ (= 10.2) is an estimate of μ

Other estimators of μ ?

Methods of Evaluating Estimators

Bias and Variance

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- The bias measures the accuracy of an estimator.
- An estimator whose bias is zero is called unbiased.
- An unbiased estimator may, nevertheless, fluctuate greatly from sample to sample.

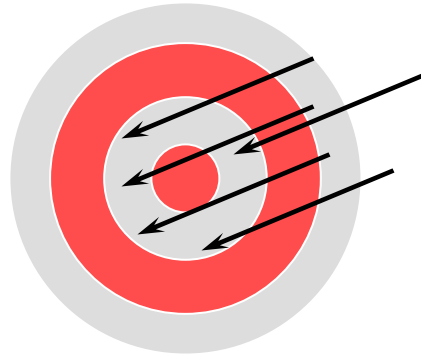
$$\text{Var}(\hat{\theta}) = E\left\{[\hat{\theta} - E(\hat{\theta})]^2\right\}$$

- The lower the variance, the more precise the estimator.
- A low-variance estimator may be biased.
- Among unbiased estimators, the one with the lowest variance should be chosen. “Best”=minimum variance.

Accuracy and Precision



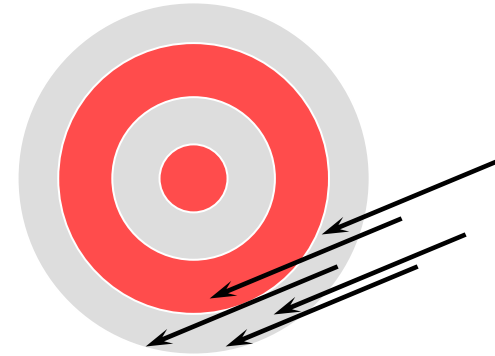
accurate and
precise



accurate,
not precise



precise,
not accurate



not accurate,
not precise

Mean Squared Error

- To choose among all estimators (biased and unbiased), minimize a measure that combines both bias and variance.
- A “good” estimator should have low bias (accurate) AND low variance (precise).

$$\text{MSE}(\hat{\theta}) = E\left\{[\hat{\theta} - \theta]^2\right\} = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2 \text{ (eqn 6.2)}$$

MSE = expected squared error loss function

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

$$\text{Var}(\hat{\theta}) = E\left\{[\hat{\theta} - E(\hat{\theta})]^2\right\}$$

Example: estimators of variance

Two estimators of variance:

$$S_1^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1) \quad \text{is unbiased (Example 6.3)}$$

$$S_2^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / n \quad \text{is biased but has smaller MSE (Example 6.4)}$$

In spite of larger MSE, we almost always use S_1^2

Example - Poisson

(See example in Casella & Berger, page 308)

Standard Error (SE)

- The standard deviation of an estimator is called the standard error of the estimator (SE).
- The estimated standard error is also called standard error (se).
- The precision of an estimator is measured by the SE.

Examples for the normal and binomial distributions:

1. \bar{X} is an unbiased estimator of μ

$$\left. \begin{aligned} SE(\bar{X}) &= \sigma / \sqrt{n} \\ se(\bar{X}) &= s / \sqrt{n} \end{aligned} \right\} \text{are called the standard error of the mean}$$

2. \hat{p} is an unbiased estimator of p

$$se(\hat{p}) = \sqrt{\hat{p}(1 - \hat{p})/n}$$

Precision and Standard Error

- A precise estimate has a small standard error, but exactly how are the precision and standard error related?
- If the sampling distribution of an estimator is normal with mean equal to the true parameter value (i.e., unbiased). Then we know that about 95% of the time the estimator will be within two SE's from the true parameter value.

Methods of Point Estimation

- Method of Moments (Chapter 6)
- **Maximum Likelihood Estimation (Chapter 15)**
- Least Squares (Chapter 10 and 11)

Method of Moments

- Equate sample moments to population moments (as we did with Poisson).
- Example: for the continuous uniform distribution, $f(x|a,b)=1/(b-a)$, $a \leq x \leq b$
- $E(X) = (b+a)/2$, $\text{Var}(X)=(b-a)^2/12$
- Set $\bar{X} = (b+a)/2$
- $S^2 = (b-a)^2/12$
- Solve for a and b (can be a bit messy).

Maximum Likelihood Parameter Estimation

- By far the most popular estimation method! (Casella & Berger).
- MLE is the parameter point for which observed data is most likely under the assumed probability model.
- Likelihood function: $L(\theta | \mathbf{x}) = f(\mathbf{x} | \theta)$, where \mathbf{x} is the vector of sample values, θ also a vector possibly.
- When we consider $f(\mathbf{x} | \theta)$, we consider θ as fixed and \mathbf{x} as the variable.
- When we consider $L(\theta | \mathbf{x})$, we are considering \mathbf{x} to be the fixed observed sample point and θ to be varying over all possible parameter values.

MLE (continued)

- If X_1, \dots, X_n are iid then

$$L(\theta|\mathbf{x}) = f(x_1, \dots, x_n | \theta) = \prod f(x_i | \theta)$$

- The MLE of θ is the value which maximizes the likelihood function (assuming it has a global maximum).
- Found by differentiating when possible.
- Usually work with log of likelihood function ($\prod \rightarrow \sum$).
- Equations obtained by setting partial derivatives of $\ln L(\theta) = 0$ are called the likelihood equations.
- See text page 616 for example – normal distribution.

Confidence Interval Estimation

We want an interval $[L, U]$ where L and U are two statistics calculated from X_1, X_2, \dots, X_n such that

$$P[L \leq \theta \leq U] = 1 - \alpha \quad \left\{ \begin{array}{l} \text{Note: } L \text{ and } U \text{ are random} \\ \text{and } \theta \text{ is fixed but unknown} \end{array} \right.$$

regardless of the true value of θ .

- $[L, U]$ is called a $100(1 - \alpha)\%$ confidence interval (CI).
- $1 - \alpha$ is called the confidence level of the interval.
- After the data is observed $X_1 = x_1, \dots, X_n = x_n$, the confidence limits $L = l$ and $U = u$ can be calculated.

95% Confidence Interval: Normal σ^2 known

Consider a random sample $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$ where σ^2 is assumed to be known and μ is an unknown parameter to be estimated. Then

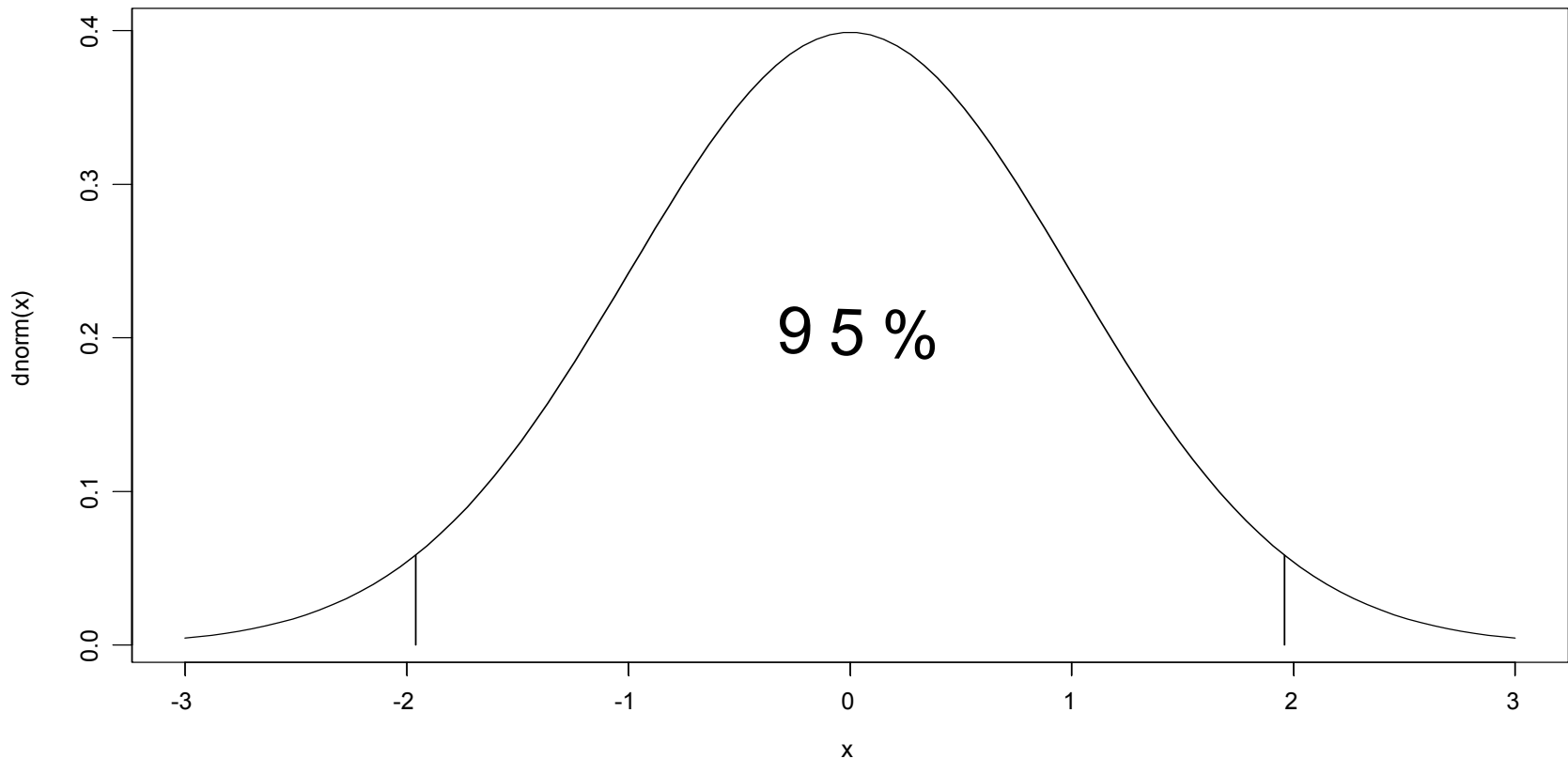
$$P\left[-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right] = 0.95 \left\{ \begin{array}{l} \text{By the CLT even if the sample} \\ \text{is not normal, this result is} \\ \text{approximately correct.} \end{array} \right.$$

$$\Rightarrow P\left[L = \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq U = \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right] = 0.95$$

$$\Rightarrow l = \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} = u \text{ is a 95\% CI for } \mu \text{ (two-sided)}$$

- See Example 6.7, Airline Revenues, p. 204

Normal Distribution, 95% of area under curve is between -1.96 and 1.96



Frequentist Interpretation of CI's

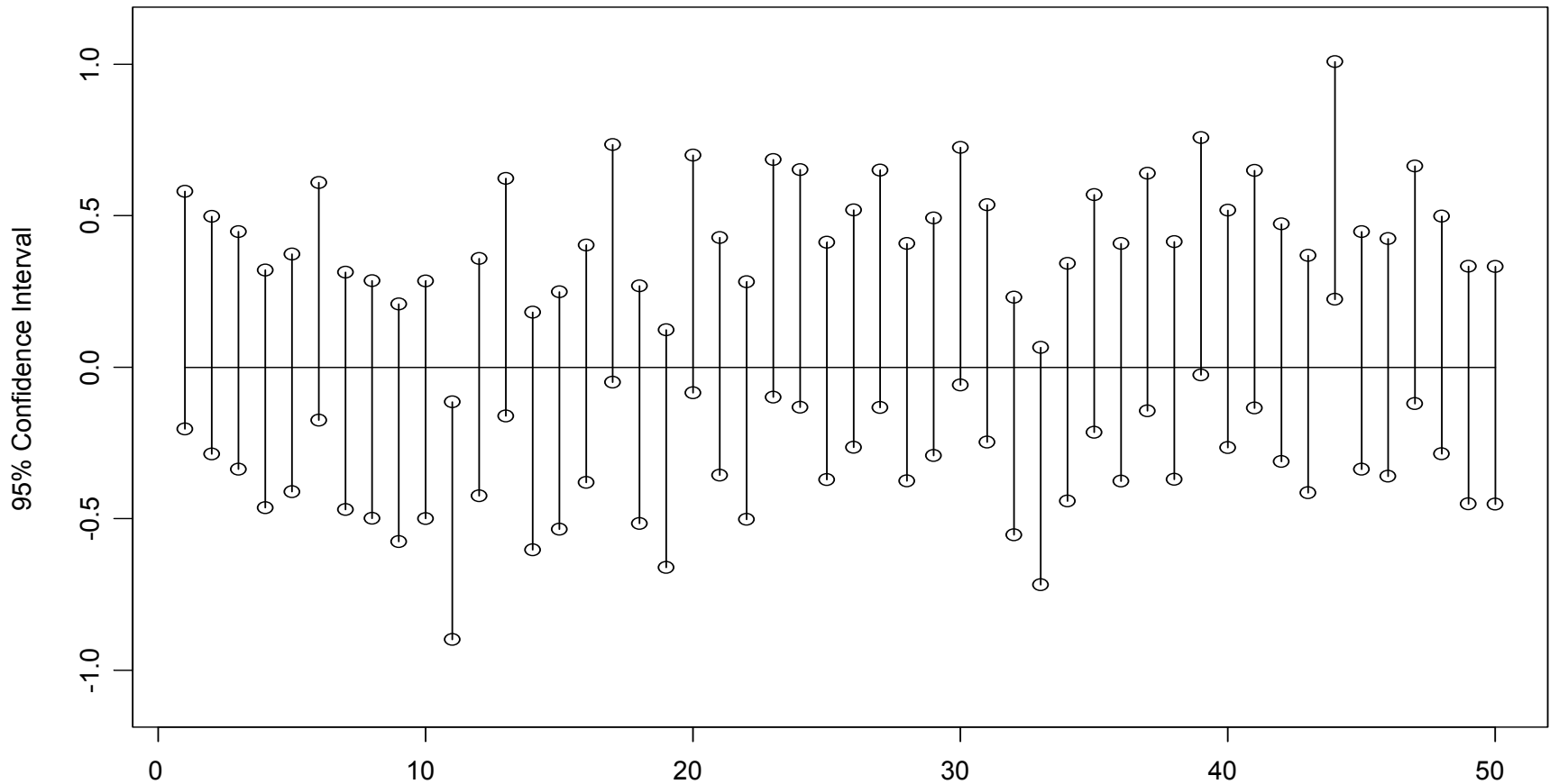
In an infinitely long series of trials in which repeated samples of size n are drawn from the same population and 95% CI's for μ are calculated using the same method, the proportion of intervals that actually include μ will be 95% (coverage probability).

However, for any particular CI, it is not known whether or not the CI includes μ , but the probability that it includes μ is either 0 or 1, that is, either it does or it doesn't.

It is incorrect to say that the probability is 0.95 that the true μ is in a particular CI.

- See Figure 6.2, p. 205

95% CI, 50 samples from unit normal distribution



Arbitrary Confidence Level for CI: σ^2 known

100(1- α)% two-sided CI for μ based on the observed sample mean

$$\bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

For 99% confidence,
 $Z_{\alpha/2} = 2.576$

The price paid for higher confidence level is a wider interval.

For large samples, these CI can be used for data from any distribution, since by CLT $\bar{x} \approx N(\mu, \sigma^2/n)$.

One-sided Confidence Intervals

$$\mu \geq \bar{x} - Z_{\alpha} \frac{\sigma}{\sqrt{n}} \quad \text{Lower one-sided CI}$$

$$\mu \leq \bar{x} + Z_{\alpha} \frac{\sigma}{\sqrt{n}} \quad \text{Upper one-sided CI}$$

For 95% confidence,
 $Z_{\alpha} = 1.645$ vs.
 $Z_{\alpha/2} = 1.96$

One-sided CIs are tighter for the same confidence level.

Hypothesis Testing

The objective of hypothesis testing is to assess the validity of a claim against a counterclaim using sample data.

- The claim to be “proved” is the alternative hypothesis (H_1).
- The competing claim is called the null hypothesis (H_0).
- One begins by assuming that H_0 is true. If the data fails to contradict H_0 beyond a reasonable doubt, then H_0 is not rejected. However, failing to reject H_0 does not mean that we accept it as true. It simply means that H_0 cannot be ruled out as a possible explanation for the observed data. A proof by insufficient data is not a proof at all.

Testing Hypotheses

“The process by which we use data to answer questions about parameters is very similar to how juries evaluate evidence about a defendant.” – from Geoffrey Vining, *Statistical Methods for Engineers*, Duxbury, 1st edition, 1998. For more information, see that textbook.

Hypothesis Tests

- A hypothesis test is a data-based rule to decide between H_0 and H_1 .
- A test statistic calculated from the data is used to make this decision.
- The values of the test statistics for which the test rejects H_0 comprise the rejection region of the test.
- The complement of the rejection region is called the acceptance region.
- The boundaries of the rejection region are defined by one or more critical constants (critical values).
- See Examples 6.13(acc. sampling) and 6.14(SAT coaching), pp. 210-211.

Hypothesis Testing as a Two-Decision Problem

Framework developed by Neyman and Pearson in 1933.

When a hypothesis test is viewed as a decision procedure, two types of errors are possible:

		<u>Decision</u>		
		Do not reject H_0	Reject H_0	
<u>Reality</u>	H_0 True	Correct Decision “Confidence” $1 - \alpha$	Type I Error “Significance Level” α	=1
	H_0 False	Type II Error “Failure to Detect” β	Correct Decision “Prob. of Detection” $1 - \beta$	=1
Column Total		$\neq 1$	$\neq 1$	

Probabilities of Type I and II Errors

$\alpha = P\{\text{Type I error}\} = P\{\text{Reject } H_0 \text{ when } H_0 \text{ is true}\} = P\{\text{Reject } H_0 | H_0\}$
also called α -risk or producer's risk or false alarm rate

$\beta = P\{\text{Type II error}\} = P\{\text{Fail to reject } H_0 \text{ when } H_1 \text{ is true}\} = P\{\text{Fail to reject } H_0 | H_1\}$
also called β -risk or consumer's risk or prob. of not detecting

$\pi = 1 - \beta = P\{\text{Reject } H_0 | H_1\}$ is prob. of detection or power of the test

We would like to have low α and low β (or equivalently, high power).

α and $1 - \beta$ are directly related, can increase power by increasing α .

These probabilities are calculated using the sampling distributions from either the null hypothesis (for α) or alternative hypothesis (for β).

Example 6.17 (SAT Coaching)

See Example 6.17, “SAT Coaching,” in the course textbook.

Power Function and OC Curve

The operating characteristic function of a test is the probability that the test fails to reject H_0 as a function of θ , where θ is the test parameter.

$$OC(\theta) = P\{\text{test fails to reject } H_0 \mid \theta\}$$

For θ values included in H_1 the OC function is the β –risk.

The power function is:

$$\pi(\theta) = P\{\text{Test rejects } H_0 \mid \theta\} = 1 - OC(\theta)$$

Example: In SAT coaching, for the test that rejects the null hypothesis when mean change is 25 or greater, the power = `1-pnorm(25,mean=0:50,sd=40/sqrt(20))`

Level of Significance

The practice of test of hypothesis is to put an upper bound on the P(Type I error) and, subject to that constraint, find a test with the lowest possible P(Type II error).

The upper bound on P(Type I error) is called the level of significance of the test and is denoted by α (usually some small number such as 0.01, 0.05, or 0.10).

The test is required to satisfy:

$$P\{ \text{Type I error} \} = P\{ \text{Test Rejects } H_0 \mid H_0 \} \leq \alpha$$

Note that α is now used to denote an upper bound on P(Type I error).

Motivated by the fact that the Type I error is usually the more serious.

A hypothesis test with a significance level α is called an a α -level test.

Choice of Significance Level

What α level should one use?

Recall that as $P(\text{Type I error})$ decreases $P(\text{Type II error})$ increases.

A proper choice of α should take into account the relative costs of Type I and Type II errors. (These costs may be difficult to determine in practice, but must be considered!)

Fisher said: $\alpha = 0.05$

Today $\alpha = 0.10, 0.05, 0.01$ depending on how much proof against the null hypothesis we want to have before rejecting it.

P-values have become popular with the advent of computer programs.

Observed Level of Significance or P-value

Simply rejecting or not rejecting H_0 at a specified α level does not fully convey the information in the data.

Example: $H_0 : \mu = 15$ vs $H_1 : \mu > 15$ is rejected at the $\alpha = 0.05$

$$\text{when } \bar{x} > 15 + 1.645 \times \frac{40}{\sqrt{20}} = 29.71$$

Is a sample with a mean of 30 equivalent to a sample with a mean of 50? (Note that both lead to rejection at the α -level of 0.05.)

More useful to report the *smallest* α -level for which the data would reject (this is called the observed level of significance or P-value).

Reject H_0 if P-value $< \alpha$

Example 6.23 (SAT Coaching: P-Value)

See Example 6.23, “SAT Coaching,” on page 220 of the course textbook.

One-sided and Two-sided Tests

$H_0 : \mu = 15$ can have three possible alternative hypotheses:

$H_1 : \mu > 15$, $H_1 : \mu < 15$, or $H_1 : \mu \neq 15$
(upper one-sided) (lower one-sided) (two-sided)

Example 6.27 (SAT Coaching: Two-sided testing)

See Example 6.27 in the course textbook.

Example 6.27 continued

See Example 6.27, “SAT Coaching,” on page 223 of the course textbook.

Relationship Between Confidence Intervals and Hypothesis Tests

An α -level two-sided test rejects a hypothesis $H_0 : \mu = \mu_0$ if and only if the $(1 - \alpha)100\%$ confidence interval does not contain μ_0 .

Example 6.7 (Airline Revenues)

See Example 6.7, “Airline Revenues,” on page 207 of the course textbook.

Use/Misuse of Hypothesis Tests in Practice

- Difficulties of Interpreting Tests on Non-random samples and observational data
- Statistical significance versus Practical significance
 - Statistical significance is a function of sample size
- Perils of searching for significance
- Ignoring lack of significance
- Confusing confidence $(1 - \alpha)$ with probability of detecting a difference $(1 - \beta)$

Jerzy Neyman
(1894-1981)

Egon Pearson
(1895-1980)

Carried on a decades-long feud with Fisher over the foundations of statistics (hypothesis testing and confidence limits)

- Fisher never recognized Type II error & developed fiducial limits