CHAPTER 7: INTRODUCTION TO STATISTICAL INFERENCE

1 Introduction

2 Parametric point estimation

2.1 Methods of finding estimates

Method of moments

Maximum likelihood

Other

2.2 Properties of point estimates

Mean Squared Error

Consistency and BAN

Sufficient statistics

2.3 Unbiased estimation

2.4 Highlight: Bayes estimators (postponed)

Posterior distribution - Loss-function invariance -

Minimax estimator

3 Parametric interval estimation

3.1 Estimating with confidence

3.2 Methods of finding confidence intervals

3.3 Highlight: Bayesian interval estimation (postponed)

4 Hypothesis testing

4.1 Tests of significance

4.2 Hypothesis types

Simple hypothesis

Composite hypothesis

4.3 Assumptions about hypothesis tests

4.4 Examples

4.5 Use and abuse of tests

4.6 Multiple testing

5 Inference for distributions

5.1 Inference for the mean of a population

5.2 Comparing two means

5.3 Optional topics in comparing distributions

6 Inference for proportions

6.1 inference for a single proportion

6.2 Comparing two proportions

7 Analysis of two-way tables

7.1 Inference for two-way tables

7.2 Formulas and models for two-way tables and Goodness-of-fit

8 How to select the appropriate test?

1 Introduction

- Statistical inference means drawing conclusions based on data. There are a many contexts in which inference is desirable, and there are many approaches to performing inference.
- One important inferential context is parametric models. For example, if you have noisy (x; y) data that you think follow the pattern $y = \beta_0 + \beta_1 x + \text{error}$, then you might want to estimate β_0, β_1, and the magnitude of the error. This will be the subject of Chapter 8

- In earlier chapters, we indicated that a sample from the distribution of a population is useful in making inferences about the population itself.
- Two important problems in statistical inference are ***estimation*** and ***testing*** (of hypotheses).
- In ***point estimation***, the value of some statistic (i.e. a function of random variables – see later) represents or estimates the unknown parameter of interest.
- In ***interval estimation***, two statistics are defined so that they span an interval for which the probability can be determined that it contains the unknown parameter

Definition of an estimator

Definition Estimator Any statistic (known function of observable random variables that is itself a random variable) whose values are used to estimate $\tau(\theta)$, where $\tau(\cdot)$ is some function of the parameter θ, is defined to be an *estimator* of $\tau(\theta)$. ////

Statistic: characteristic of a sample

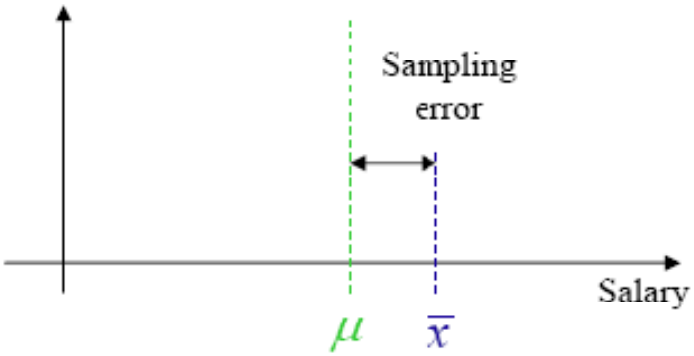
What is the average salary of 2000 people randomly sampled in Spain?

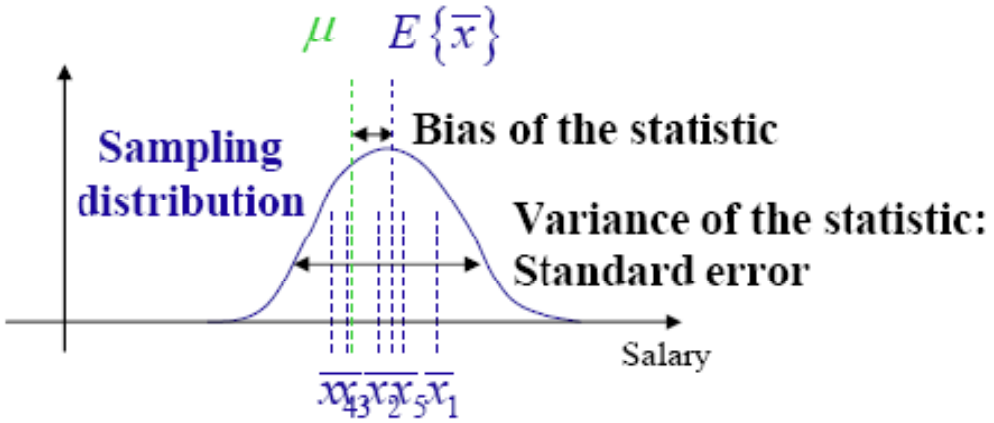
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Parameter: characteristic of a population

What is the average salary of all Spaniards?

μ





2 Parametric point estimation

2.1 Methods of finding estimates

Method of moments

Let $f(\cdot\,; \theta_1, \ldots, \theta_k)$ be a density of a random variable X which has k parameters $\theta_1, \ldots, \theta_k$. As before let μ_r' denote the rth moment about 0; that is, $\mu_r' = \mathcal{E}[X^r]$. In general μ_r' will be a known function of the k parameters $\theta_1, \ldots, \theta_k$. Denote this by writing $\mu_r' = \mu_r'(\theta_1, \ldots, \theta_k)$. Let $X_1, \ldots, X_n$ be a random sample from the density $f(\cdot\,; \theta_1, \ldots, \theta_k)$, and, as before, let M_j' be the jth sample moment; that is,

$$M_j' = \frac{1}{n} \sum_{i=1}^{n} X_i^j.$$

Form the k equations

$$M_j' = \mu_j'(\theta_1, \ldots, \theta_k), \qquad j = 1, \ldots, k, \tag{1}$$

in the k variables $\theta_1, \ldots, \theta_k$, and let $\hat{\Theta}_1, \ldots, \hat{\Theta}_k$ be their solution (we assume that there is a unique solution). We say that the estimator $(\hat{\Theta}_1, \ldots, \hat{\Theta}_k)$, where $\hat{\theta}_j$ estimates θ_j, is the estimator of $(\theta_1, \ldots, \theta_k)$ obtained by the *method of moments*. The estimators were obtained by replacing population moments by sample moments.

Example 1 of the method of moments

Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Let $(\theta_1, \theta_2) = (\mu, \sigma)$. Estimate the parameters μ and σ by the method of moments. Recall that $\sigma^2 = \mu_2' - (\mu_1')^2$ and $\mu = \mu_1'$. The method-of-moments equations become

$$M_1' = \mu_1' = \mu_1'(\mu, \sigma) = \mu$$
$$M_2' = \mu_2' = \mu_2'(\mu, \sigma) = \sigma^2 + \mu^2,$$

and their solution is the following: The method-of-moments estimator of μ is $M_1' = \bar{X}$, and the method-of-moments estimator of σ is $\sqrt{M_2' - \bar{X}^2} = \sqrt{(1/n)\sum X_i^2 - \bar{X}^2} = \sqrt{\sum (X_i - \bar{X})^2/n}$.

////

Example 2 of the method of moments

Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ. Estimate λ. There is only one parameter, hence only one equation, which is

$$M_1' = \mu_1' = \mu_1'(\lambda) = \lambda.$$

Hence the method-of-moments estimator of λ is $M_1' = \bar{X}$, which says estimate the population mean λ with the sample mean $\bar{x}$. ////

Method-of-moment estimators are not unique

Method-of-moments estimators are not uniquely defined. The method-of-moments equations given in Eq. (1) are obtained by using the first k raw moments. Central moments (rather than raw moments) could also be used to obtain equations whose solution would also produce estimators that would be labeled method-of-moments estimators. Also, moments other than the first k could be used to obtain estimators that would be labeled method-of-moments estimators.

Maximum probability

What is the proportion of smokers among statisticians?
In a class of 20 statisticians, 4 of them smoke.

$$\hat{p} = \frac{n}{N} = \frac{4}{20} = 20\%$$

Why? Couldn't it be any other number like 19%, 25%, 50%, 2%?

!!

$$Smokers \sim Binomial(N, p) \Rightarrow \Pr\{Smokers = n\} = \binom{N}{n} p^n (1-p)^{N-n}$$

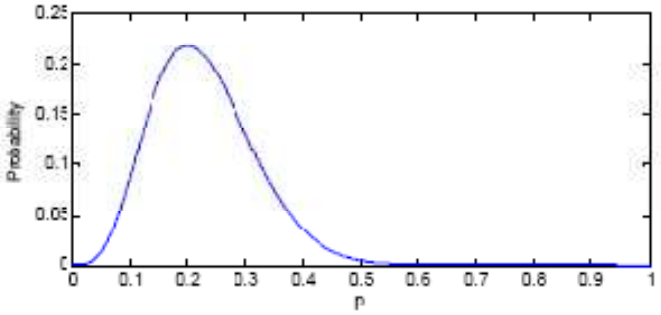
$$E\{Smokers\} = Np \longrightarrow p = \frac{E\{Smokers\}}{N}$$

p	$\Pr\{Smokers = 4\}$
0.20	0.218
0.19	0.217
0.25	0.190
0.50	0.005
0.02	0.001

0.2 is the parameter that maximizes the probability of our observation

$$\hat{\theta} = \arg\max_{\theta} \Pr\{X \mid \theta\}$$

Our data $X \equiv Smokers = 4$



Maximum likelihood

What is the average height of statisticians?
1.73, 1.67, 1.76, 1.76, 1.69, 1.81, 1.81, 1.75, 1.77, 1.76,
1.74, 1.79, 1.72, 1.86, 1.74, 1.76, 1.80, 1.75, 1.75, 1.71

$\bar{x} = 1.76$

Why? Couldn't it be any other number like 1.75, 1.60, 1.78?

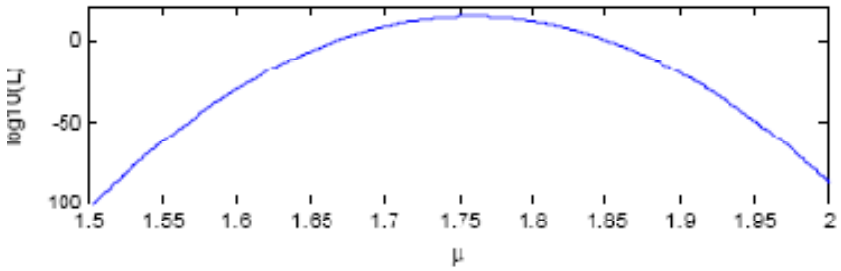
$$\hat{\theta} = \arg\max_{\theta} \Pr\{X \mid \theta\}$$

$$Height \sim N(\mu, 0.05^2) \Rightarrow \Pr\{X \mid \mu\} = \Pr\{X_1 = 1.73 \mid \mu\} \Pr\{X_2 = 1.67 \mid \mu\} \ldots \Pr\{X_{20} = 1.71 \mid \mu\} = 0!!$$

$$\mathrm{L}\{X \mid \mu\} = f_{N(\mu,0.05^2)}(1.73) f_{N(\mu,0.05^2)}(1.67) \ldots f_{N(\mu,0.05^2)}(1.71) \approx 9e13$$

$$f_{N(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$\mu = \bar{x}$



$$\hat{\theta}_{ML} = \arg\max_{\theta} \log L(X \mid \theta) \rightarrow \frac{\partial \mathrm{L}\{X \mid \mu\}}{\partial \mu} = 0 \Rightarrow \hat{\mu} = \bar{x}$$

Definition **Likelihood function** The *likelihood function* of n random variables $X_1, X_2, \ldots, X_n$ is defined to be the joint density of the n random variables, say $f_{X_1, \ldots, X_n}(x_1, \ldots, x_n; \theta)$, which is considered to be a function of θ. In particular, if $X_1, \ldots, X_n$ is a random sample from the density $f(x; \theta)$, then the likelihood function is $f(x_1; \theta) f(x_2; \theta) \cdot \cdots \cdot f(x_n; \theta)$. ////

Notation To remind ourselves to think of the likelihood function as a function of θ, we shall use the notation $L(\theta; x_1, \ldots, x_n)$ or $L(\cdot; x_1, \ldots, x_n)$ for the likelihood function. ////

Intuition behind Maximum Likelihood (ML) estimators

The likelihood function $L(\theta;\ x_1,\ \ldots,\ x_n)$ gives the *likelihood* that the random variables assume a particular value $x_1, x_2, \ldots, x_n$. The *likelihood* is the value of a density function; so for discrete random variables it is a probability. Suppose for a moment that θ is known; denote the value by θ_0. The particular value of the random variables which is "most likely to occur" is that value $x_1', x_2', \ldots, x_n'$ such that $f_{X_1, \ldots, X_n}(x_1, \ldots, x_n;\ \theta_0)$ is a maximum.

Now let us suppose that the joint density of n random variables is $f_{X_1, \ldots, X_n}(x_1, \ldots, x_n; \theta)$, where θ is unknown. Let the particular values which are observed be represented by $x_1', x_2', \ldots, x_n'$. We want to know from which density is this particular set of values most likely to have come. We want to know from which density (what value of θ) is the likelihood largest that the set $x_1', \ldots, x_n'$ was obtained. In other words, we want to find the value of θ in $\underline{\Theta}$, denoted by $\hat{\theta}$, which maximizes the likelihood function $L(\theta; x_1', \ldots, x_n')$. The value $\hat{\theta}$ which maximizes the likelihood function is, in general, a function of $x_1, \ldots, x_n$, say $\hat{\theta} = \hat{\vartheta}(x_1, x_2, \ldots, x_n)$. When this is the case, the random variable $\hat{\Theta} = \hat{\vartheta}(X_1, X_2, \ldots, X_n)$ is called the *maximum-likelihood estimator* of θ. (We are assuming throughout that the maximum

Definition of ML estimator

Definition **Maximum-likelihood estimator** Let

$$L(\theta) = L(\theta; x_1, \ldots, x_n)$$

be the likelihood function for the random variables $X_1, X_2, \ldots, X_n$. If $\hat{\theta}$ [where $\hat{\theta} = \hat{\vartheta}(x_1, x_2, \ldots, x_n)$ is a function of the observations $x_1, \ldots, x_n$] is the value of θ in $\overline{\Theta}$ which maximizes $L(\theta)$, then $\hat{\Theta} = \hat{\vartheta}(X_1, X_2, \ldots, X_n)$ is the *maximum-likelihood estimator* of θ. $\hat{\theta} = \hat{\vartheta}(x_1, \ldots, x_n)$ is the maximum-likelihood estimate of θ for the sample $x_1, \ldots, x_n$. ////

The most important cases which we shall consider are those in which $X_1, X_2, \ldots, X_n$ is a *random sample* from some density $f(x; \theta)$, so that the likelihood function is

$$L(\theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta).$$

Many likelihood functions satisfy regularity conditions; so the maximum-likelihood estimator is the solution of the equation

$$\frac{dL(\theta)}{d\theta} = 0.$$

Also $L(\theta)$ and $\log L(\theta)$ have their maxima at the same value of θ, and it is sometimes easier to find the maximum of the logarithm of the likelihood.

If the likelihood function contains k parameters, that is, if

$$L(\theta_1, \theta_2, \ldots, \theta_k) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \ldots, \theta_k),$$

then the maximum-likelihood estimators of the parameters $\theta_1, \theta_2, \ldots, \theta_k$ are the random variables $\hat{\Theta}_1 = \hat{\vartheta}_1(X_1, \ldots, X_n)$, $\hat{\Theta}_2 = \hat{\vartheta}_2(X_1, \ldots, X_n)$, $\ldots$, $\hat{\Theta}_k = \hat{\vartheta}_k(X_1, \ldots, X_n)$, where $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ are the values in $\overline{\Theta}$ which maximize $L(\theta_1, \theta_2, \ldots, \theta_k)$.

If certain regularity conditions are satisfied, the point where the likelihood is a maximum is a solution of the k equations

$$\frac{\partial L(\theta_1, \ldots, \theta_k)}{\partial \theta_1} = 0$$

$$\frac{\partial L(\theta_1, \ldots, \theta_k)}{\partial \theta_2} = 0$$

$$\vdots$$

$$\frac{\partial L(\theta_1, \ldots, \theta_k)}{\partial \theta_k} = 0.$$

In this case it may also be easier to work with the logarithm of the likelihood.

Example of ML estimators

A random sample of size n from the normal distribution has the density

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2\sigma^2)(x_i-\mu)^2} = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right].$$

The logarithm of the likelihood function is

$$L^* = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2,$$

where $\sigma > 0$ and $-\infty < \mu < \infty$.

To find the location of its maximum, we compute

$$\frac{\partial L^*}{\partial \mu} = \frac{1}{\sigma^2}\sum(x_i-\mu)$$

and

$$\frac{\partial L^*}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum (x_i - \mu)^2,$$

and on putting these derivatives equal to 0 and solving the resulting equations for μ and σ^2, we find the estimates

$$\hat{\mu} = \frac{1}{n}\sum x_i = \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \bar{x})^2,$$

which turn out to be the sample moments corresponding to μ and σ^2.

////

The invariance property of ML estimators

Theorem **Invariance property of maximum-likelihood estimators** Let $\hat{\Theta} = \hat{\vartheta}(X_1, X_2, \ldots, X_n)$ be the maximum-likelihood estimator of θ in the density $f(x; \theta)$, where θ is assumed unidimensional. If $\tau(\cdot)$ is a function with a single-valued inverse, then the maximum-likelihood estimator of $\tau(\theta)$ is $\tau(\hat{\Theta})$. ////

For example, in the normal density with μ_0 known the maximum-likelihood estimator of σ^2 is

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2.$$

By the invariance property of maximum-likelihood estimators, the maximum-likelihood estimator of σ is

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2}.$$

Similarly, the maximum-likelihood estimator of, say, $\log \sigma^2$ is

$$\log\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2\right].$$

Bayesian approach

What is the average height of Spanish statisticians knowing that the average should be around 1.70 because that is the average height of Spaniards?
1.73, 1.79, 1.76, 1.76
Now, $\hat{\mu} = \overline{x} = 1.76$ is rather strange. Maybe we were unlucky in our sample

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \log L(X \mid \theta) L(\theta)$$

Now, our parameter is itself a random variable with an *a priori* known distribution

$$\left.\begin{array}{l} Height \sim N(\mu, 0.05^2) \\ \mu \sim N(1.70, 0.05^2) \end{array}\right\} \Rightarrow \hat{\mu} = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\overline{x} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 \longrightarrow \hat{\mu} = 1.75$$

μ_0 σ_0^2

Bayesian parameter estimation is one of the most powerful estimates **IF** you have the right a priori distribution

Other criteria to estimate parameters

Minimum Mean Squared Error

$$\hat{\theta}_{MMSE} = \arg\min_{\hat{\theta}} E\left\{(\theta-\hat{\theta})^2\right\} = \arg\min_{\hat{\theta}} Var\left\{\hat{\theta}\right\} + \left(Bias(\hat{\theta},\theta)\right)^2$$

↑ Depends on something I don't know. Solution: $\hat{\theta}_{SURE}$ Stein's unbiased risk estimator

Minimum risk

$$\hat{\theta}_{risk} = \arg\min_{\hat{\theta}} E\left\{Cost(\hat{\theta},\theta)\right\} \longrightarrow Cost(\hat{\theta},\theta) = \left(\theta-\hat{\theta}\right)^2 \Rightarrow \theta_{risk} = E\{\theta \mid x\}$$

$$Cost(\hat{\theta},\theta) = \left|\theta-\hat{\theta}\right| \Rightarrow \theta_{risk} = Median\{\theta \mid x\}$$

$$Cost(\hat{\theta},\theta) = \begin{cases} 0 & |x| < \Delta \\ \Delta & |x| \geq \Delta \end{cases} \Rightarrow \theta_{risk} = Mode\{\theta \mid x\}$$

Minimum Variance Unbiased Estimator

$$\hat{\theta}_{MVUE} = \arg\min_{\hat{\theta}} Var\left\{\hat{\theta}\right\}$$

Best Linear Unbiased Estimator

$$\hat{\theta}_{BLUE} = \arg\min_{\hat{\theta}} Var\left\{\hat{\theta}\right\} \quad s.t. \quad \hat{\theta}_{BLUE} = \sum_{i=1}^{N} \alpha_i x_i$$

Cramer-Rao Lower Bound

$$Var\left\{\hat{\theta}\right\} \geq \frac{1}{I(\theta)}$$

← Fisher's information

In all of them you need to know the posterior distribution of θ given the data x

2.2 Properties of point estimates

Mean Squared Error

Definition **Mean-squared error** Let $T = \ell(X_1, \ldots, X_n)$ be an estimator of $\tau(\theta)$. $\mathscr{E}_\theta[[T - \tau(\theta)]^2]$ is defined to be the *mean-squared error* of the estimator $T = \ell(X_1, \ldots, X_n)$. ////

Notation Let $\text{MSE}_\ell(\theta)$ denote the mean-squared error of the estimator $T = \ell(X_1, \ldots, X_n)$ of $\tau(\theta)$. ////

Remark The subscript θ on the expectation symbol $\mathscr{E}_\theta$ indicates from which density in the family under consideration the sample came. That is,

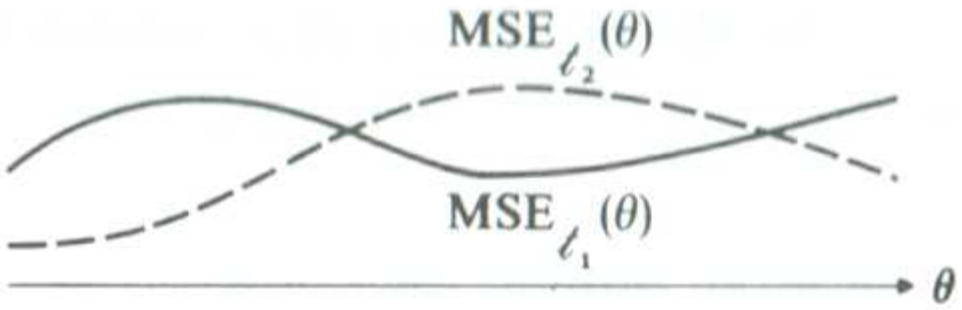
$$\mathscr{E}_\theta[[T - \tau(\theta)]^2]$$

$$= \mathscr{E}_\theta[[\ell(X_1, \ldots, X_n) - \tau(\theta)]^2]$$

$$= \int \cdots \int [\ell(x_1, \ldots, x_n) - \tau(\theta)]^2 f(x_1; \theta) \cdots f(x_n; \theta)\, dx_1 \cdots dx_n,$$

where $f(x; \theta)$ is the probability density function from which the random sample was selected. ////

For any two estimators $T_1 = \ell_1(X_1, \ldots, X_n)$ and $T_2 = \ell_2(X_1, \ldots, X_n)$ of $\tau(\theta)$, their respective mean-squared errors $\text{MSE}_{\ell_1}(\theta)$ and $\text{MSE}_{\ell_2}(\theta)$ as functions of θ are likely to cross; so for some θ, ℓ_1 has smaller MSE, and for others ℓ_2 has smaller MSE. We would then have no basis for preferring one of the estimators over the other.



One reason for being unable to find an estimator with uniformly smallest mean-squared error is that the class of all possible estimators is too large—it includes some estimators that are extremely prejudiced in favor of particular θ.

One could restrict the totality of estimators by considering only estimators that satisfy some other property. One such property is that of *unbiasedness*.

Definition Unbiased An estimator $T = \ell(X_1, \ldots, X_n)$ is defined to be an *unbiased* estimator of $\tau(\theta)$ if and only if

$$\mathscr{E}_\theta[T] = \mathscr{E}_\theta[\ell(X_1, \ldots, X_n)] = \tau(\theta) \qquad \text{for all } \theta \in \underline{\Theta}. \qquad ////$$

Remark

$$\text{MSE}_{\mathcal{A}}(\theta) = \text{var}\,[T] + \{\tau(\theta) - \mathcal{E}_\theta[T]\}^2.$$

So if T is an unbiased estimator of $\tau(\theta)$, then $\text{MSE}_{\mathcal{A}}(\theta) = \text{var}\,[T]$.

PROOF

$$\begin{aligned}\text{MSE}_{\mathcal{A}}(\theta) &= \mathcal{E}_\theta[[T - \tau(\theta)]^2] = \mathcal{E}_\theta[((T - \mathcal{E}_\theta[T]) - \{\tau(\theta) - \mathcal{E}_\theta[T]\})^2] \\ &= \mathcal{E}_\theta[(T - \mathcal{E}_\theta[T])^2] - 2\{\tau(\theta) - \mathcal{E}_\theta[T]\}\mathcal{E}_\theta[T - \mathcal{E}_\theta[T]] \\ &\quad + \mathcal{E}_\theta[\{\tau(\theta) - \mathcal{E}_\theta[T]\}^2] = \text{var}\,[T] + \{\tau(\theta) - \mathcal{E}_\theta[T]\}^2. \qquad ////\end{aligned}$$

The term $\tau(\theta) - \mathcal{E}_\theta[T]$ is called the *bias* of the estimator T and can be either positive, negative, or zero. The remark shows that the mean-squared error is the sum of two nonnegative quantities; it also shows how the mean-squared error, variance, and bias of an estimator are related.

Consistency and BAN

In the previous subsection we defined the mean-squared error of an estimator and the property of unbiasedness. Both concepts were defined for a fixed sample size. In this subsection we will define two concepts that are defined for increasing sample size. In our notation for an estimator of $\tau(\theta)$, let us use $T_n = \ell_n(X_1, \ldots, X_n)$, where the subscript n of ℓ indicates sample size. Actually we will be considering a sequence of estimators, say $T_1 = \ell_1(X_1)$, $T_2 = \ell_2(X_1, X_2)$ $T_3 = \ell_3(X_1, X_2, X_3), \ldots, T_n = \ell_n(X_1, \ldots, X_n), \ldots$. An obvious example is $T_n = \ell_n(X_1, \ldots, X_n) = \overline{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Ordinarily the functions ℓ_n in the sequence will be the same *kind* of function for each n.

When considering a sequence of estimators, it seems that a good sequence of estimators should be one for which the values of the estimators tend to get closer to the quantity being estimated as the sample size increases. The following definitions formalize this intuitively desirable notion of limiting closeness.

Definition **Mean-squared-error consistency** Let $T_1, T_2, \ldots, T_n \ldots$ be a sequence of estimators of $\tau(\theta)$, where $T_n = t_n(X_1, \ldots, X_n)$ is based on a sample of size n. This sequence of estimators is defined to be a *mean-squared-error consistent* sequence of estimators of $\tau(\theta)$, if and only if $\lim_{n \to \infty} \mathscr{E}_\theta[[T_n - \tau(\theta)]^2] = 0$ for all θ in $\underline{\Theta}$. ////

Remark Mean-squared-error consistency implies that both the bias and the variance of T_n approach 0 since $\mathscr{E}_\theta[[T_n - \tau(\theta)]^2] = \text{var}\,[T_n] + \{\tau(\theta) - \mathscr{E}_\theta[T_n]\}^2$. ////

EXAMPLE In sampling from any density having mean μ and variance σ^2, let $\overline{X}_n = (1/n)\sum_{i=1}^{n} X_i$ be a sequence of estimators of μ and $S_n^2 = [1/(n-1)]\sum_{i=1}^{n}(X_i - \overline{X}_n)$ be a sequence of estimators of σ^2. $\mathscr{E}[(\overline{X}_n - \mu)^2] = \text{var}\,[\overline{X}_n] = \sigma^2/n \to 0$ as $n \to \infty$; hence the sequence $\{\overline{X}_n\}$ is a mean-squared-error consistent sequence of estimators of μ.

$$\mathscr{E}[(S_n^2 - \sigma^2)^2] = \text{var}\,[S_n^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right) \to 0$$

as $n \to \infty$ hence the sequence $\{S_n^2\}$ is a mean-squared-error consistent sequence of estimators of σ^2. Note that if $T_n = (1/n)\sum (X_i - \overline{X})^2$, then the sequence $\{T_n\}$ is also a mean-squared-error consistent sequence of estimators of σ^2. ////

Definition 9 Simple consistency Let $T_1, T_2, \ldots, T_n, \ldots$ be a sequence of estimators of $\tau(\theta)$, where $T_n = t_n(X_1, \ldots, X_n)$. The sequence $\{T_n\}$ is defined to be a *simple* (or *weakly*) *consistent* sequence of estimators of $\tau(\theta)$ if for every $\varepsilon > 0$ the following is satisfied:

$$\lim_{n \to \infty} P_\theta[\tau(\theta) - \varepsilon < T_n < \tau(\theta) + \varepsilon] = 1 \qquad \text{for every } \theta \text{ in } \overline{\Theta}. \qquad ////$$

Remark If an estimator is a mean-squared-error consistent estimator, it is also a simple consistent estimator, but not necessarily vice versa.

PROOF

$$P_\theta[\tau(\theta) - \varepsilon < T_n < \tau(\theta) + \varepsilon] = P[|T_n - \tau(\theta)| < \varepsilon]$$

$$= P_\theta[[T_n - \tau(\theta)]^2 < \varepsilon^2] \geq 1 - \frac{\mathscr{E}_\theta[[T_n - \tau(\theta)]^2]}{\varepsilon^2}$$

by the Chebyshev inequality. As n approaches infinity, $\mathscr{E}_\theta[[T_n - \tau(\theta)]^2]$ approaches 0; hence $\lim_{n \to \infty} P_\theta[\tau(\theta) - \varepsilon < T_n < \tau(\theta) + \varepsilon] = 1$. ////

Definition **Best asymptotically normal estimators (BAN estimators)** A sequence of estimators $T_1^*, \ldots, T_n^*, \ldots$ of $\tau(\theta)$ is defined to be *best asymptotically normal* (BAN) if and only if the following four conditions are satisfied:

(i) The distribution of $\sqrt{n}[T_n^* - \tau(\theta)]$ approaches the normal distribution with mean 0 and variance $\sigma^{*2}(\theta)$ as n approaches infinity.

(ii) For every $\varepsilon > 0$,

$$\lim_{n \to \infty} P_\theta[|T_n^* - \tau(\theta)| > \varepsilon] = 0 \qquad \text{for each } \theta \text{ in } \overline{\Theta}.$$

(iii) Let $\{T_n\}$ be any other sequence of simple consistent estimators for which the distribution of $\sqrt{n}[T_n - \tau(\theta)]$ approaches the normal distribution with mean 0 and variance $\sigma^2(\theta)$.

(iv) $\sigma^2(\theta)$ is not less than $\sigma^{*2}(\theta)$ for all θ in any open interval. ////

Remark The abbreviation BAN is sometimes replaced by CANE, standing for *consistent asymptotically normal efficient.* ////

The usefulness of this definition derives partially from theorems proving the existence of BAN estimators and from the fact that ordinarily reasonable estimators are asymptotically normally distributed.

It can be shown that for samples drawn from a normal density with mean μ and variance σ^2 the sequence $T_n^* = (1/n)\sum_{i=1}^{n} X_i = \bar{X}_n$ for $n = 1, 2, \ldots$ is a BAN estimator of μ. In fact, the limiting distribution of $\sqrt{n}(\bar{X}_n - \mu)$ is normal with mean 0 and variance σ^2, and no other estimator can have smaller limiting variance in any interval of μ values.

Loss and Risk function

we used *mean-squared error* of an estimator as a measure of the closeness of the estimator to $\tau(\theta)$. Other measures are possible, for example,

$$\mathcal{E}_\theta[|T - \tau(\theta)|],$$

called the *mean absolute deviation*. In order to exhibit and consider still other measures of closeness, we will borrow and rely on the language of decision theory. On the basis of an observed random sample from some density function, the statistician has to *decide* what to estimate $\tau(\theta)$ to be. One might then call the value of some estimator $T = \ell(X_1, \ldots, X_n)$ a *decision* and call the estimator itself a *decision function* since it tells us what *decision* to make. Now the estimate t of $\tau(\theta)$ might be in error; if so, some measure of the severity of the error seems appropriate. The word "loss" is used in place of "error," and "loss function" is used as a measure of the "error." A formal definition follows.

Definition **Loss function** Consider estimating $\tau(\theta)$. Let t denote an estimate of $\tau(\theta)$. The *loss function*, denoted by $\ell(t; \theta)$, is defined to be a real-valued function satisfying (i) $\ell(t; \theta) \geq 0$ for all possible estimates t and all θ in $\overline{\Theta}$ and (ii) $\ell(t; \theta) = 0$ for $t = \tau(\theta)$. $\ell(t; \theta)$ equals the *loss* incurred if one estimates $\tau(\theta)$ to be t when θ is the true parameter value. ////

In a given estimation problem one would have to define an appropriate loss function for the particular problem under study. It is a measure of the error and presumably would be greater for large error than for small error. We would want the loss to be small; or, stated another way, we want the error in estimation to be small, or we want the estimate to be close to what it is estimating.

Several possible loss functions are:

(i) $\ell_1(t;\theta) = [t - \tau(\theta)]^2$.

(ii) $\ell_2(t;\theta) = |t - \tau(\theta)|$.

(iii) $\ell_3(t;\theta) = \begin{cases} A & \text{if } |t - \tau(\theta)| > \varepsilon \\ 0 & \text{if } |t - \tau(\theta)| \leq \varepsilon, \text{ where } A > 0. \end{cases}$

(iv) $\ell_4(t;\theta) = \rho(\theta)|t - \tau(\theta)|^r$ for $\rho(\theta) \geq 0$ and $r > 0$.

ℓ_1 is called the *squared-error* loss function, and ℓ_2 is called the *absolute-error* loss function. Note that both ℓ_1 and ℓ_2 increase as the error $t - \tau(\theta)$ increases in magnitude. ℓ_3 says that you lose nothing if the estimate t is within ε units of $\tau(\theta)$ and otherwise you lose amount A. ℓ_4 is a general loss function that includes both ℓ_1 and ℓ_2 as special cases. ////

We assume now that an appropriate loss function has been defined for our estimation problem, and we think of the loss function as a measure of error or loss. Our object is to select an estimator $T = \ell(X_1, \ldots, X_n)$ that makes this error or loss small. (Admittedly, we are not considering a very important, substantive problem by assuming that a suitable loss function is given. In general, selection of an appropriate loss function is not trivial.) The loss function in its first argument depends on the estimate t, and t is a value of the estimator T; that is, $t = \ell(x_1, \ldots, x_n)$. Thus, our loss depends on the sample $X_1, \ldots, X_n$. We cannot hope to make the loss small for every possible sample, but we can try to make the loss small on the average. Hence, if we alter our objective of picking that estimator that makes the loss small to picking that estimator that makes the *average* loss small, we can remove the dependence of the loss on the sample $X_1, \ldots, X_n$. This notion is embodied in the following definition.

Definition **Risk function** For a given loss function $\ell(\cdot\,;\,\cdot)$, the *risk function*, denoted by $\mathcal{R}_t(\theta)$, of an estimator $T = t(X_1, \ldots, X_n)$ is defined to be

$$\mathcal{R}_t(\theta) = \mathcal{E}_\theta[\ell(T\,;\,\theta)].$$

////

Interpretation

The risk function is the *average loss*. The expectation in Eq. (10) can be taken in two ways. For example, if the density $f(x; \theta)$ from which we sampled is a probability density function, then

$$\mathscr{E}_\theta[\ell(T; \theta)] = \mathscr{E}_\theta[\ell(\ell(X_1, \ldots, X_n); \theta)]$$
$$= \int \cdots \int \ell(\ell(x_1, \ldots, x_n); \theta) \prod_{i=1}^{n} f(x_i; \theta)\, dx_i.$$

Or we can consider the random variable T and the density of T. We get

$$\mathscr{E}_\theta[\ell(T; \theta)] = \int \ell(\ ; \theta) f_T(t)\, dt,$$

where $f_T(t)$ is the density of the estimator T. In either case, the expectation averages out the values of $x_1, \ldots, x_n$.

EXAMPLE Consider the same loss functions given before The corresponding risks are given by:

(i) $\mathcal{E}_\theta[[T - \tau(\theta)]^2]$, our familiar mean-squared error.
(ii) $\mathcal{E}_\theta[|T - \tau(\theta)|]$, the mean absolute error.
(iii) $A \cdot P_\theta[|T - \tau(\theta)| > \varepsilon]$.
(iv) $\rho(\theta)\mathcal{E}_\theta[|T - \tau(\theta)|^r]$. ////

Our object now is to select an estimator that makes the average loss (risk) small and ideally select an estimator that has the smallest risk. To help meet this objective, we use the concept of admissible estimators.

Definition **Admissible estimator** For two estimators $T_1 = t_1(X_1, \ldots, X_n)$ and $T_2 = t_2(X_1, \ldots, X_n)$, estimator t_1 is defined to be a *better* estimator than t_2 if and only if

$$\mathscr{R}_{t_1}(\theta) \leq \mathscr{R}_{t_2}(\theta) \qquad \text{for all } \theta \text{ in } \overline{\Theta}$$

and

$$\mathscr{R}_{t_1}(\theta) < \mathscr{R}_{t_2}(\theta) \qquad \text{for at least one } \theta \text{ in } \overline{\Theta}.$$

An estimator $T = t(X_1, \ldots, X_n)$ is defined to be *admissible* if and only if there is no better estimator. ////

In general, given two estimators t_1 and t_2 neither is better than the other; that is, their respective risk functions as functions of θ, cross. We observed this same phenomenon when we studied the mean-squared error. Here, as there, there will not, in general, exist an estimator with uniformly smallest risk.

The problem is the dependence of the risk function on θ. What we might do is average out θ, just as we average out the dependence on $x_1, \ldots, x_n$ when going from the loss function to the risk function. The question then is: Just how should θ be averaged out? We will consider just this problem in Sec. 7 on the Bayes estimators. Another way of removing the dependence of the risk function on θ is to replace the risk function by its maximum value and compare estimators by looking at their respective maximum risks, naturally preferring that estimator with smallest maximum risk. Such an estimator is said to be *minimax*.

Definition **Minimax** An estimator t^* is defined to be a *minimax* estimator if and only if $\sup_\theta \mathscr{R}_{t^*}(\theta) \leq \sup_\theta \mathscr{R}_t(\theta)$ for every estimator t. ////

Sufficient statistics

Sufficient statistics are of interest in themselves, as well as being useful in statistical inference problems such as estimation or testing of hypotheses. Because the concept of sufficiency is widely applicable, possibly the notion should have been isolated in a chapter by itself rather than buried in this chapter on estimation.

Definition **Sufficient statistic** Let $X_1, \ldots, X_n$ be a random sample from the density $f(\cdot\,; \theta)$, where θ may be a vector. A statistic $S = \mathcal{S}(X_1, \ldots, X_n)$ is defined to be a *sufficient statistic* if and only if the conditional distribution of $X_1, \ldots, X_n$ given $S = s$ does not depend on θ for any value s of S. ////

A sufficient statistic is a particular kind of statistic. It is a statistic that condenses $\mathfrak{X}$ in such a way that no "information about θ" is lost. The only information about the parameter θ in the density $f(\cdot\,;\theta)$ from which we sampled is contained in the sample $X_1, \ldots, X_n$; so, when we say that a statistic loses no information, we mean that it contains all the information about θ that is contained in the sample. We emphasize that the type of information of which we are speaking is that information about θ contained in the sample given that we know the form of the density; that is, we know the function $f(\cdot\,;\,\cdot)$ in $f(\cdot\,;\theta)$, and the parameter θ is the only unknown. We are not speaking of information in the sample that might be useful in checking the validity of our assumption that the density does indeed have form $f(\cdot\,;\,\cdot)$.

The definition says that a statistic $S = s(X_1, \ldots, X_n)$ is sufficient if the conditional distribution of the sample given the value of the statistic does not depend on θ. The idea is that if you know the value of the sufficient statistic, then the sample values themselves are not needed and can tell you nothing more about θ, and this is true since the distribution of the sample given the sufficient statistic does not depend on θ. One cannot hope to learn anything about θ by sampling from a distribution that does not depend on θ.

EXAMPLE Let X_1, X_2, X_3 be a sample of size 3 from the Bernoulli distribution. Consider the two statistics $S = \mathcal{A}(X_1, X_2, X_3) = X_1 + X_2 + X_3$ and $T = \mathcal{U}(X_1, X_2, X_3) = X_1X_2 + X_3$. We will show that $\mathcal{A}(\cdot, \cdot, \cdot)$ is sufficient and $\mathcal{U}(\cdot, \cdot, \cdot)$ is not.

The conditional densities given in the last two columns are routinely calculated. For instance,

$$f_{X_1, X_2, X_3|S=1}(0, 1, 0|1) = P[X_1 = 0; X_2 = 1; X_3 = 0|S = 1]$$

$$= \frac{P[X_1 = 0; X_2 = 1; X_3 = 0; S = 1]}{P[S = 1]}$$

$$= \frac{(1-p)p(1-p)}{\binom{3}{1}p(1-p)^2} = \frac{1}{3};$$

	Values of S	Values of T	$f_{X_1,X_2,X_3\|S}$	$f_{X_1,X_2,X_3\|T}$
(0, 0, 0)	0	0	1	$\frac{1-p}{1+p}$
(0, 0, 1)	1	1	$\frac{1}{3}$	$\frac{1-p}{1+2p}$
(0, 1, 0)	1	0	$\frac{1}{3}$	$\frac{p}{1+p}$
(1, 0, 0)	1	0	$\frac{1}{3}$	$\frac{p}{1+p}$
(0, 1, 1)	2	1	$\frac{1}{3}$	$\frac{p}{1+2p}$
(1, 0, 1)	2	1	$\frac{1}{3}$	$\frac{p}{1+2p}$
(1, 1, 0)	2	1	$\frac{1}{3}$	$\frac{p}{1+2p}$
(1, 1, 1)	3	2	1	1

and

$$f_{X_1, X_2, X_3|T=0}(0, 1, 0|0) = \frac{P[X_1 = 0; X_2 = 1; X_3 = 0; T = 0]}{P[T = 0]}$$

$$= \frac{(1-p)^2 p}{(1-p)^3 + 2(1-p)^2 p} = \frac{p}{1-p+2p} = \frac{p}{1+p}.$$

The conditional distribution of the sample given the values of S is independent of p; so S is a sufficient statistic; however, the conditional distribution of the sample given the values of T depends on p; so T is not sufficient.

Hence …

Definition **Sufficient statistic** Let $X_1, \ldots, X_n$ be a random sample from the density $f(\cdot\,;\,\theta)$. A statistic $S = \mathcal{S}(X_1, \ldots, X_n)$ is defined to be a *sufficient statistic* if and only if the conditional distribution of T given S does not depend on θ for *any* statistic $T = \mathcal{t}(X_1, \ldots, X_n)$. ////

For instance, to prove that a statistic $T' = \mathcal{t}'(X_1, \ldots, X_n)$ is not sufficient, one needs only to find another statistic $T = \mathcal{t}(X_1, \ldots, X_n)$ for which the conditional distribution of T given T' depends on θ.

For some problems, no single sufficient statistic exists. However, there will always exist jointly sufficient statistics.

Definition **Jointly sufficient statistics** Let $X_1, \ldots, X_n$ be a random sample from the density $f(\cdot\,; \theta)$. The statistics $S_1, \ldots, S_r$ are defined to be *jointly sufficient* if and only if the conditional distribution of $X_1, \ldots, X_n$ given $S_1 = s_1, \ldots, S_r = s_r$ does not depend on θ. ////

Definition 18 **Minimal sufficient statistic** A set of jointly sufficient statistics is defined to be *minimal sufficient* if and only if it is a function of every other set of sufficient statistics. ////

Like many definitions, Definition is of little use in finding minimal sufficient statistics. A technique for finding minimal sufficient statistics has been devised by Lehmann and Scheffé , but we will not present it. If the joint density is properly factored, the factorization criterion will give us minimal sufficient statistics.

2.3 Unbiased estimation

the mean-squared error of an estimator T of $\tau(\theta)$ can be written as

$$\mathscr{E}_\theta[[T - \tau(\theta)]^2] = \text{var}_\theta[T] + \{\tau(\theta) - \mathscr{E}_\theta[T]\}^2,$$

and if T is an unbiased estimator of $\tau(\theta)$, then $\mathscr{E}_\theta[T] = \tau(\theta)$, and so $\mathscr{E}_\theta[[T - \tau(\theta)]^2] = \text{var}_\theta[T]$. Hence, seeking an estimator with uniformly minimum mean-squared error among unbiased estimators is tantamount to seeking an estimator with uniformly minimum variance among unbiased estimators.

Definition **Uniformly minimum-variance unbiased estimator (UMVUE)**

Let $X_1, \ldots, X_n$ be a random sample from $f(\cdot\,; \theta)$. An estimator $T^* = t^*(X_1, \ldots, X_n)$ of $\tau(\theta)$ is defined to be a *uniformly minimum-variance unbiased estimator* of $\tau(\theta)$ if and only if (i) $\mathcal{E}_\theta[T^*] = \tau(\theta)$, that is, T^* is unbiased, and (ii) $\text{var}_\theta[T^*] \leq \text{var}_\theta[T]$ for any other estimator $T = t(X_1, \ldots, X_n)$ of $\tau(\theta)$ which satisfies $\mathcal{E}_\theta[T] = \tau(\theta)$. ////

How to find UMVUEs?

Lower bound for the variance

Let $X_1, \ldots, X_n$ be a random sample from $f(\cdot\,;\theta)$, where θ belongs to $\underline{\Theta}$. Assume that $\underline{\Theta}$ is a subset of the real line. Let $T = \ell(X_1, \ldots, X_n)$ be an unbiased estimator of $\tau(\theta)$. We will consider the case where $f(\cdot\,;\theta)$ is a probability density function; the development for discrete density functions is analogous. We make the following assumptions, called *regularity conditions:*

(i) $\dfrac{\partial}{\partial\theta} \log f(x;\theta)$ exists for all x and all θ.

(ii) $$\frac{\partial}{\partial\theta}\int\cdots\int\prod_{i=1}^{n} f(x_i;\theta)\,dx_1\cdots dx_n$$

$$=\int\cdots\int\frac{\partial}{\partial\theta}\prod_{i=1}^{n} f(x_i;\theta)\,dx_1\cdots dx_n.$$

(iii) $$\frac{\partial}{\partial\theta}\int\cdots\int t(x_1,\ldots,x_n)\prod_{i=1}^{n} f(x_i;\theta)\,dx_1\cdots dx_n$$

$$=\int\cdots\int t(x_1,\ldots,x_n)\frac{\partial}{\partial\theta}\prod_{i=1}^{n} f(x_i;\theta)\,dx_1\cdots dx_n.$$

(iv) $$0<\mathscr{E}_\theta\left[\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right]<\infty \text{ for all } \theta \text{ in } \underline{\Theta}.$$

Theorem **Cramér-Rao inequality** Under assumptions (i) to (iv) above

$$\operatorname{var}_\theta [T] \geq \frac{[\tau'(\theta)]^2}{n\mathscr{E}_\theta\left[\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right]} \tag{15}$$

where $T = \ell(X_1, \ldots, X_n)$ is an unbiased estimator of $\tau(\theta)$. Equality prevails in Eq. (15) if and only if there exists a function, say $K(\theta, n)$, such that

$$\sum_1^n \frac{\partial}{\partial\theta}\log f(x_i;\theta) = K(\theta, n)[\ell(x_1, \ldots, x_n) - \tau(\theta)]. \tag{16}$$

Equation (15) is called the *Cramér-Rao inequality*, and the right-hand side is called the *Cramér-Rao lower bound* for the variance of unbiased estimators of $\tau(\theta)$.

(proof for your information)

PROOF

$$\tau'(\theta) = \frac{\partial}{\partial\theta}\tau(\theta) = \frac{\partial}{\partial\theta}\int\cdots\int t(x_1, \ldots, x_n)\prod_{i=1}^{n} f(x_i; \theta)\, dx_1 \cdots dx_n$$

$$= \int\cdots\int t(x_1, \ldots, x_n)\frac{\partial}{\partial\theta}\left[\prod_{i=1}^{n} f(x_i; \theta)\right] dx_1 \cdots dx_n$$

$$- \tau(\theta)\ \frac{\partial}{\partial\theta}\int\cdots\int\prod_{i=1}^{n}[f(x_i; \theta)\, dx_i]$$

$$= \int\cdots\int t(x_1, \ldots, x_n)\frac{\partial}{\partial\theta}\left[\prod_{i=1}^{n} f(x_i; \theta)\right] dx_1 \cdots dx_n$$

$$- \tau(\theta)\int\cdots\int\frac{\partial}{\partial\theta}\left[\prod_{i=1}^{n} f(x_i; \theta)\right] dx_1 \cdots dx_n$$

$$= \int \cdots \int [t(x_1, \ldots, x_n) - \tau(\theta)] \frac{\partial}{\partial \theta} \left[\prod_{i=1}^{n} f(x_i; \theta \right] dx_1 \cdots dx_n$$

$$= \int \cdots \int [t(x_1, \ldots, x_n) - \tau(\theta)] \left[\frac{\partial}{\partial \theta} \log \prod_{i=1}^{n} f(x_i; \theta) \right]$$

$$\times \prod_{i=1}^{n} f(x_i; \theta)\, dx_1 \cdots dx_n$$

$$= \mathscr{E}_\theta \left[[t(X_1, \ldots, X_n) - \tau(\theta)] \left[\frac{\partial}{\partial \theta} \log \prod_{i=1}^{n} f(X_i; \theta) \right] \right].$$

Now by the Cauchy-Schwarz inequality

$$[\tau'(\theta)]^2 \leq \mathscr{E}_\theta[[t(X_1, \ldots, X_n) - \tau(\theta)]^2] \mathscr{E}_\theta \left[\left[\frac{\partial}{\partial \theta} \log \prod_{i=1}^{n} f(X_i; \theta) \right]^2 \right],$$

or

$$\operatorname{var}_\theta [T] \geq \frac{[\tau'(\theta)]^2}{\mathscr{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^{n} f(X_i;\theta)\right)^2\right]};$$

but

$$\begin{aligned}\mathscr{E}_\theta\left[\left[\frac{\partial}{\partial\theta}\log\prod_{i=1}^{n} f(X_i;\theta)\right]^2\right] &= \mathscr{E}_\theta\left[\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i;\theta)\right]^2\right] \\ &= \sum_i\sum_j \mathscr{E}_\theta\left[\left[\frac{\partial}{\partial\theta}\log f(X_i;\theta)\right]\left[\frac{\partial}{\partial\theta}\log f(X_j;\theta)\right]\right] \\ &= n\mathscr{E}_\theta\left[\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right],\end{aligned}$$

using the independence of X_i and X_j and noting that

$$\mathscr{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] = \int\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right]f(x;\theta)\,dx$$

$$= \int\frac{\partial}{\partial\theta}f(x;\theta)\,dx = \frac{\partial}{\partial\theta}\int f(x;\theta)\,dx = \frac{\partial}{\partial\theta}(1) = 0.$$

The inequality in the Cauchy-Schwarz inequality becomes an equality if and only if one function is proportional to the other; in our case this requires that $\dfrac{\partial}{\partial\theta}\log\prod_{i=1}^{n} f(x_i;\theta)$ be proportional to $t(x_1,\ldots,x_n) - \tau(\theta)$ or that there exists $K = K(\theta, n)$ such that

$$\frac{\partial}{\partial\theta}\log\left[\prod_{i=1}^{n} f(x_i;\theta)\right] = K(\theta,n)[t(x_1,\ldots,x_n) - \tau(\theta)]. \qquad ////$$

Uses of the theorem

The theorem has two uses: First, it gives a lower bound for the variance of unbiased estimators. An experimenter using an unbiased estimator whose variance was close to the Cramér-Rao lower bound would know that he was using a good unbiased estimator. Second, if an unbiased estimator whose variance coincides with the Cramér-Rao lower bound can be found, then this estimator is an UMVUE. Equation (16) aids in finding an estimator whose variance coincides with the Cramér-Rao lower bound. In fact, if there exists a $T^* = \ell^*(X_1, \ldots, X_n)$ such that

$$\sum_1^n \frac{\partial}{\partial\theta} \log f(x_i; \theta) = K(\theta, n)[\ell^*(x_1, \ldots, x_n) - \tau^*(\theta)]$$

for some functions $K(\theta, n)$ and $\tau^*(\theta)$, then T^* is an UMVUE of $\tau^*(\theta)$.

EXAMPLE Let $X_1, \ldots, X_n$ be a random sample from $f(x; \theta) = \theta e^{-\theta x} I_{(0,\infty)}(x)$. Take $\tau(\theta) = \theta$. It can be shown that the regularity conditions are satisfied. $\tau'(\theta) = 1$; hence

$$\operatorname{var}_\theta [T] \geq \frac{1}{n\mathscr{E}_\theta\left[\left[\frac{\partial}{\partial\theta} \log f(X; \theta)\right]^2\right]}.$$

Note that $\frac{\partial}{\partial\theta} \log f(x; \theta) = \frac{\partial}{\partial\theta} (\log\theta - \theta x) = 1/\theta - x$, and so

$$\mathscr{E}_\theta\left[\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right] = \mathscr{E}_\theta\left[\left(\frac{1}{\theta} - X\right)^2\right] = \text{var}\,[X] = \frac{1}{\theta^2}.$$

Hence, the Cramér-Rao lower bound for the variance of unbiased estimators of θ is given by

$$\text{var}_\theta\,[T] \geq \frac{1}{n(1/\theta^2)} = \frac{\theta^2}{n}.$$

Similarly the Cramér-Rao lower bound for the variance of unbiased estimators of $\tau(\theta) = 1/\theta$ is given by

$$\text{var}_\theta\,[T] \geq \frac{[\tau'(\theta)]^2}{n(1/\theta^2)} = \frac{1}{n\theta^2}.$$

The left-hand side of Eq. (16) is

$$\sum_1^n \frac{\partial}{\partial\theta} \log f(x_i;\theta) = \sum_1^n \frac{\partial}{\partial\theta}(\log\theta - \theta x_i) = \sum_1^n \left(\frac{1}{\theta} - x_i\right) = -n\left(\bar{x}_n - \frac{1}{\theta}\right).$$

By taking $K(\theta, n) = -n$ and utilizing the result of Eq. (16), we see that $\bar{X}_n$ is an UMVUE of $1/\theta$ since its variance coincides with the Cramér-Rao lower bound. ////

In this subsection we will continue our search for UMVUEs. Our first result will show how sufficiency aids in this search. Loosely speaking, an unbiased estimator which is a function of sufficient statistics has smaller variance than an unbiased estimator which is not based on sufficient statistics. In fact, let $f(\cdot\,; \theta)$ be the density from which we can sample, and suppose that we want to estimate $\tau(\theta)$. Let us assume that $T = \ell(X_1, \ldots, X_n)$ is an unbiased estimator of $\tau(\theta)$ and that $S = s(X_1, \ldots, X_n)$ is a sufficient statistic. It can be shown that another unbiased estimator, denoted by T', can be derived from T such that (i) T' is a function of the sufficient statistic S and (ii) T' is an unbiased estimator of $\tau(\theta)$ with variance less than or equal to the variance of T. Therefore, in our search for UMVUEs we need to consider only unbiased estimators that are functions of sufficient statistics. We shall formalize these ideas in the following theorem.

Theorem Rao-Blackwell Let $X_1, \ldots, X_n$ be a random sample from the density $f(\cdot\,; \theta)$, and let $S_1 = \mathcal{s}_1(X_1, \ldots, X_n), \ldots, S_k = \mathcal{s}_k(X_1, \ldots, X_n)$ be a set of jointly sufficient statistics. Let the statistic $T = \mathcal{t}(X_1, \ldots, X_n)$ be an unbiased estimator of $\tau(\theta)$. Define T' by $T' = \mathcal{E}[T \mid S_1, \ldots, S_k]$. Then,

(i) T' is a statistic, and it is a function of the sufficient statistics $S_1, \ldots, S_k$. Write $T' = \mathcal{t}'(S_1, \ldots, S_k)$.

(ii) $\mathcal{E}_\theta[T'] = \tau(\theta)$; that is, T' is an unbiased estimator of $\tau(\theta)$.

(iii) $\text{var}_\theta\,[T'] \leq \text{var}_\theta\,[T]$ for every θ, and $\text{var}_\theta\,[T'] < \text{var}_\theta\,[T]$ for some θ unless T is equal to T' with probability 1.

For many applications (particularly where the density involved has only one unknown parameter) there will exist a single sufficient statistic, say $S = s(X_1, \ldots, X_n)$, which would then be used in place of the jointly sufficient set of statistics $S_1, \ldots, S_k$. What the theorem says is that, given an unbiased estimator, another unbiased estimator that is a function of sufficient statistics can be derived and it will not have larger variance. To find the derived statistic, the calculation of a conditional expectation, which may or may not be easy, is required.

2.4 Highlight: Bayes estimators (postponed)

3 Parametric interval estimation

3.1 Estimating with confidence

Overview of inference

- Methods for drawing conclusions about a population from sample data are called statistical inference

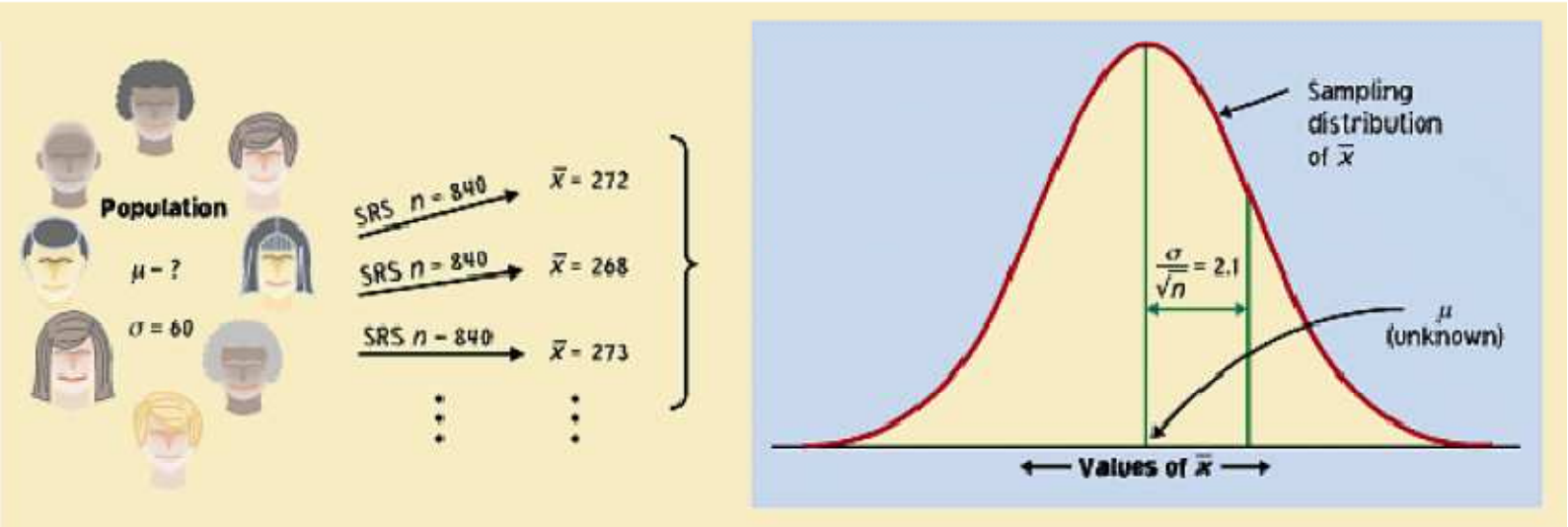
- Methods
  - Confidence Intervals - estimating a value of a population parameter
  - Tests of significance - assess evidence for a claim about a population

- Inference is appropriate when data are produced by either
  - a random sample or
  - a randomized experiment

Statistical confidence

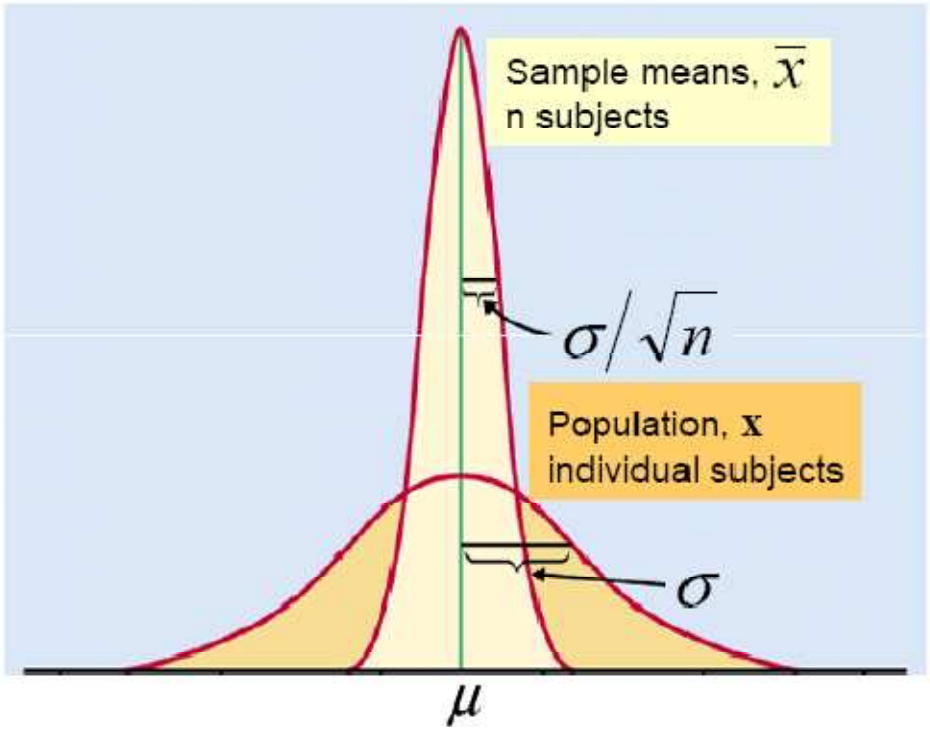
Although the sample mean, $\overline{x}$, is a unique number for any particular sample, if you pick a different sample you will probably get a different sample mean.

In fact, you could get many different values for the sample mean, and virtually none of them would actually equal the true population mean, μ.



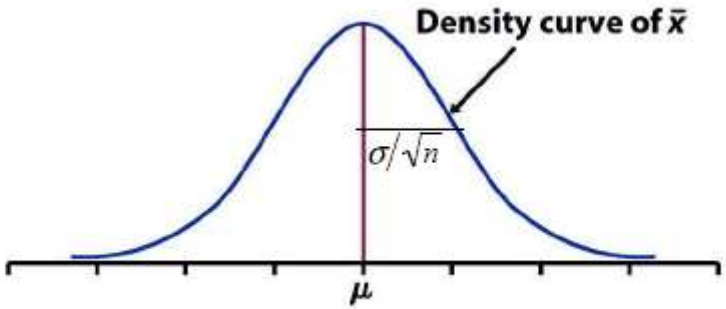
But the sample distribution is narrower than the population distribution, by a factor of $\sqrt{n}$.

Thus, the estimates $\bar{x}$ gained from our samples are always relatively close to the population parameter μ.

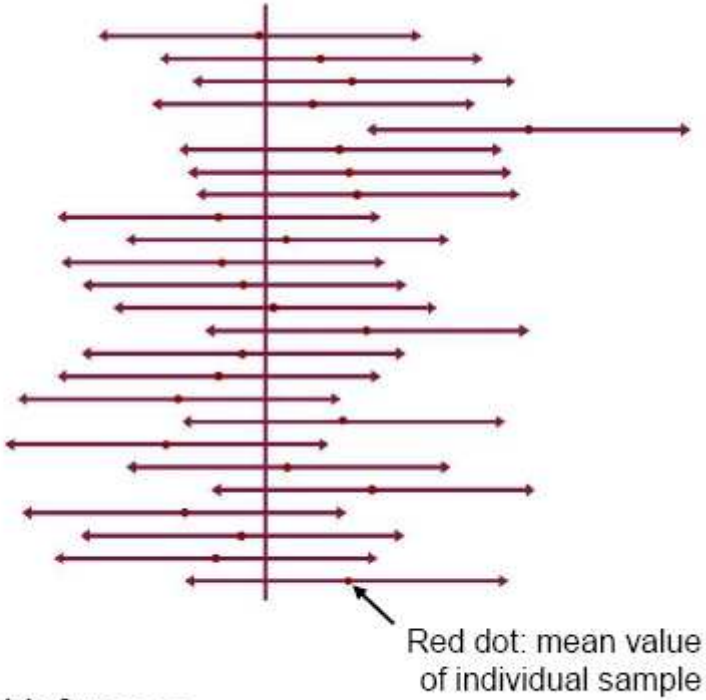


If the population is normally distributed $N(\mu,\sigma)$, so will the sampling distribution $N(\mu,\sigma/\sqrt{n})$,

95% of all sample means will be within roughly 2 standard deviations ($2^*\sigma/\sqrt{n}$) of the population parameter μ.



Distances are symmetrical which implies that **the population parameter μ must be within roughly 2 standard deviations from the sample average $\bar{x}$, in 95% of all samples.**



This reasoning is the essence of statistical inference.

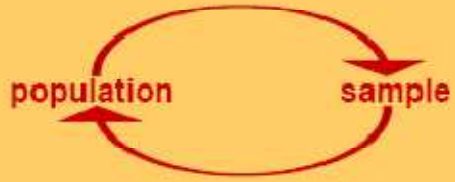
The weight of single eggs of the brown variety is normally distributed *N*(65 g,5 g). Think of a carton of 12 brown eggs as an SRS of size 12.

- What is the distribution of the sample means $\overline{x}$?

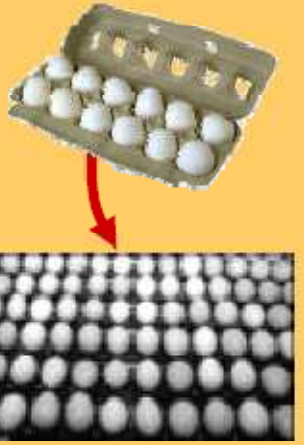
  Normal (mean μ, standard deviation $\sigma/\sqrt{n}$) = N(65 g,1.44 g).

- Find the middle 95% of the sample means distribution.

  Roughly ± 2 standard deviations from the mean, or 65g ± 2.88g.



You buy a carton of 12 white eggs instead. The box weighs 770 g. The average egg weight from that SRS is thus $\overline{x}$ = 64.2 g.

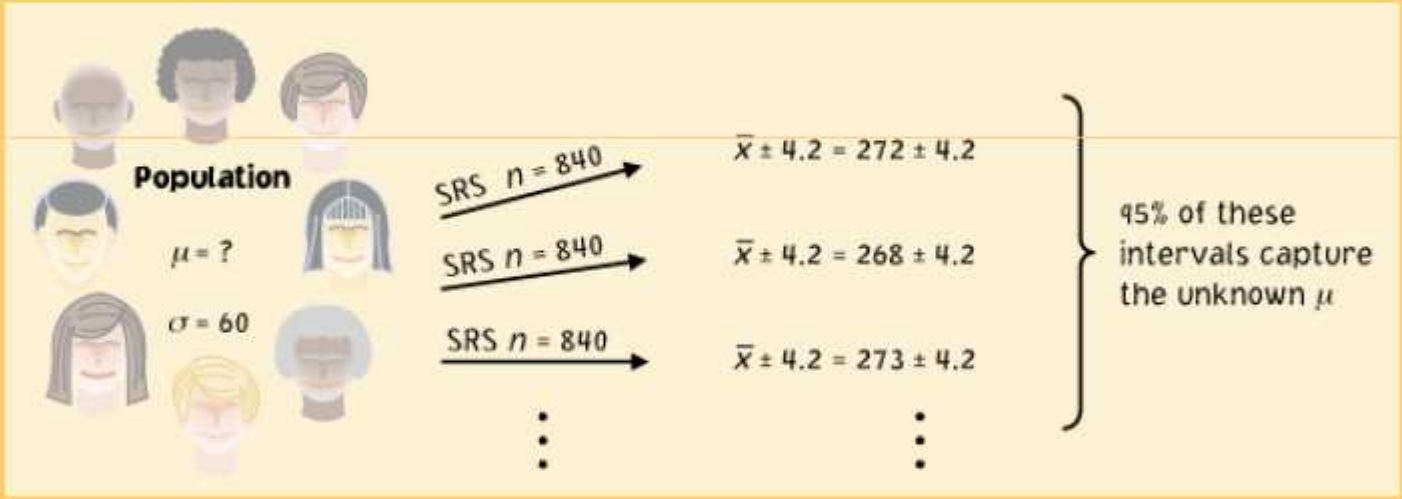


- Knowing that the standard deviation of egg weight is 5 g, what can you infer about the mean μ of the white egg population?

  There is a 95% chance that the population mean μ is roughly within ± $2\sigma/\sqrt{n}$ of $\overline{x}$, or 64.2 g ± 2.88 g.

Confidence intervals

The **confidence interval** is a range of values with an associated probability or **confidence level *C***. The probability quantifies the chance that the interval contains the true population parameter.

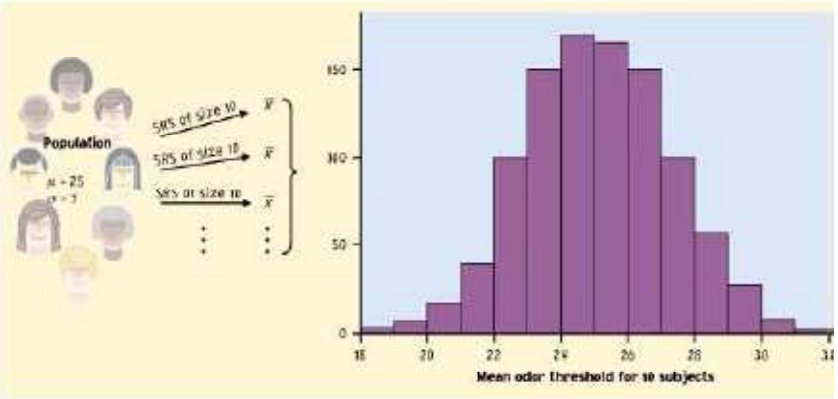


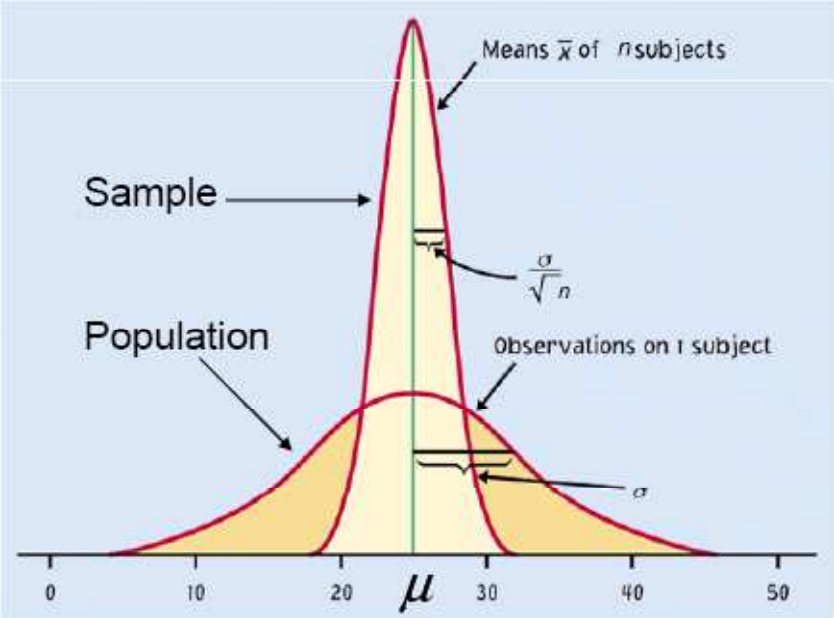
$\bar{x} \pm 4.2$ is a 95% confidence interval for the population parameter μ.

This equation says that in 95% of the cases, the actual value of μ will be within 4.2 units of the value of $\bar{x}$.

Implications

We don't need to take a lot of random samples to "rebuild" the sampling distribution and find μ at its center.

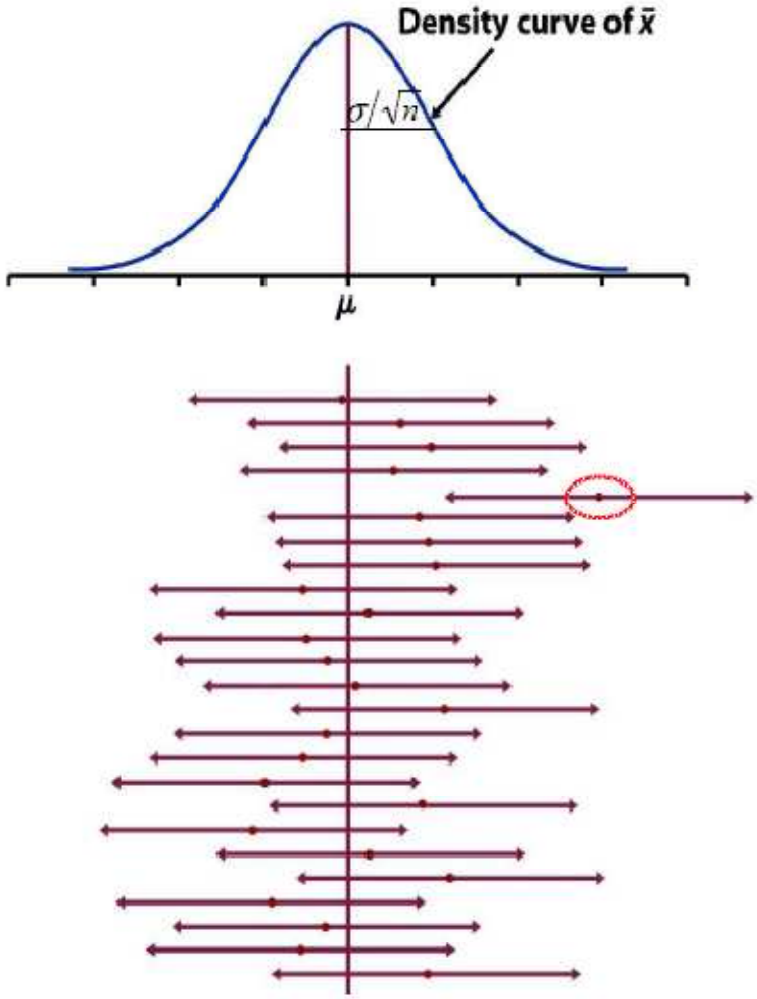




All we need is one SRS of size *n* and rely on the properties of the sample means distribution to infer the population mean μ.

With 95% confidence, we can say that μ should be within roughly 2 standard deviations ($2^*\sigma/\sqrt{n}$) from our sample mean $\bar{x}$.

- In 95% of all possible samples of this size n, μ will indeed fall in our confidence interval.
- In only 5% of samples would $\bar{x}$ be farther from μ.

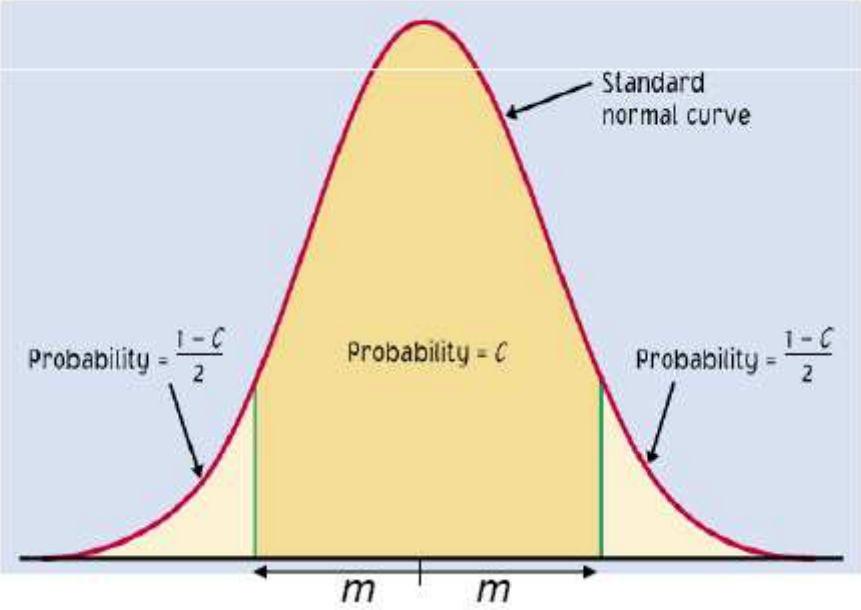


A **confidence interval** can be expressed as:

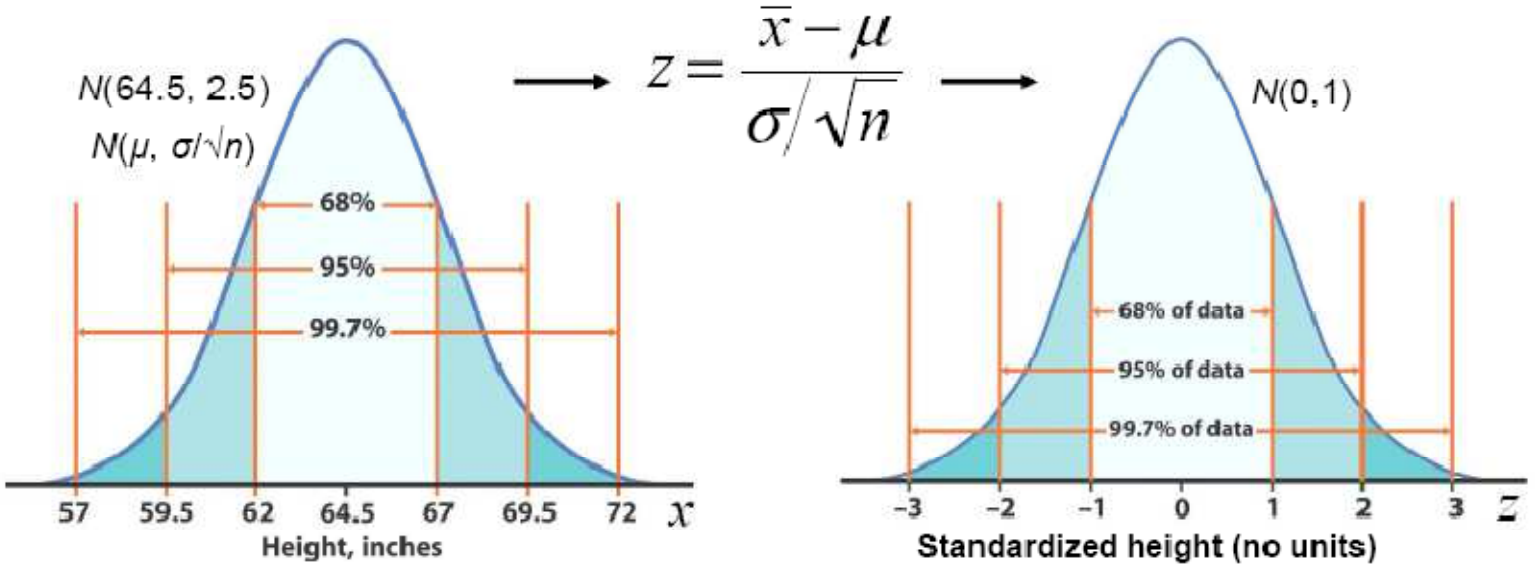
- Mean ± *m*
  m is called the **margin of error**
  μ within $\bar{x} \pm m$
  Example: 120 ± 6
- Two endpoints of an interval
  μ within $(\bar{x} - m)$ to $(\bar{x} + m)$
  ex. 114 to 126

A **confidence level *C*** (in %) indicates the probability that the μ falls within the interval.

It represents the area under the normal curve within ± *m* of the center of the curve.



Standardizing the normal curve using z



Here, we work with the sampling distribution, and $\sigma/\sqrt{n}$ is its standard deviation (spread).

Remember that σ is the standard deviation of the original population.

Varying confidence levels

Confidence intervals contain the population mean μ in C% of samples.

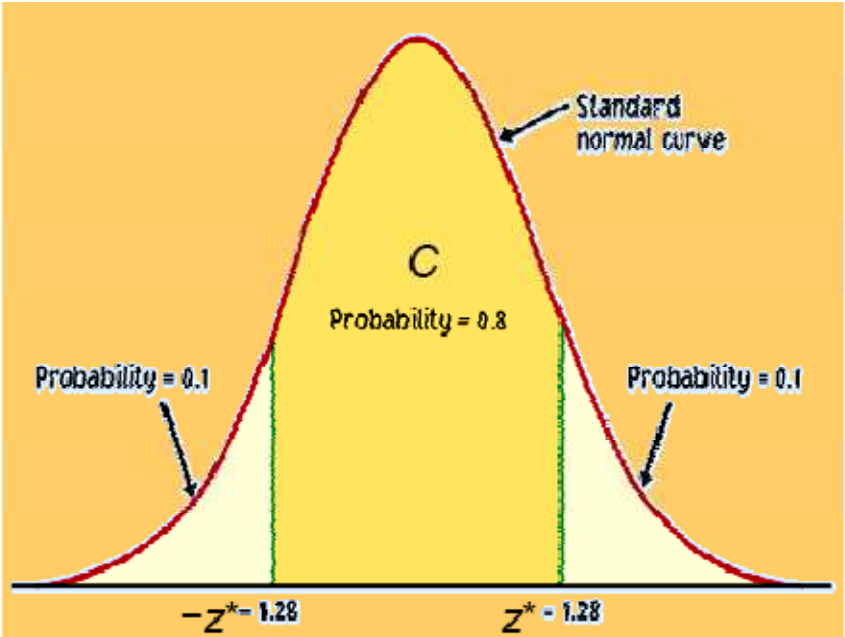
Different areas under the curve give different confidence levels C.

Practical use of z: z*

- z^* is related to the chosen confidence level C.
- C is the area under the standard normal curve between $-z^*$ and z^*.

The confidence interval is thus:

$$\bar{x} \pm z^* \sigma / \sqrt{n}$$



Example: For an 80% confidence level C, 80% of the normal curve's area is contained in the interval.

How do we find specific z* values?

We can use a table of *z/t* values (Table D). For a particular confidence level, *C*, the appropriate *z** value is just above it.

z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level *C*											

Example: For a 98% confidence level, *z**=2.326

We can use software. In **Excel:**

=NORMINV(probability,mean,standard_dev)

gives z for a given cumulative probability.

Since we want the middle *C* probability, the probability we require is (1 - C)/2

Example: For a 98% confidence level, =NORMINV(.01,0,1) = −2.32635 (= neg. *z**)

Link between confidence level and margin of error

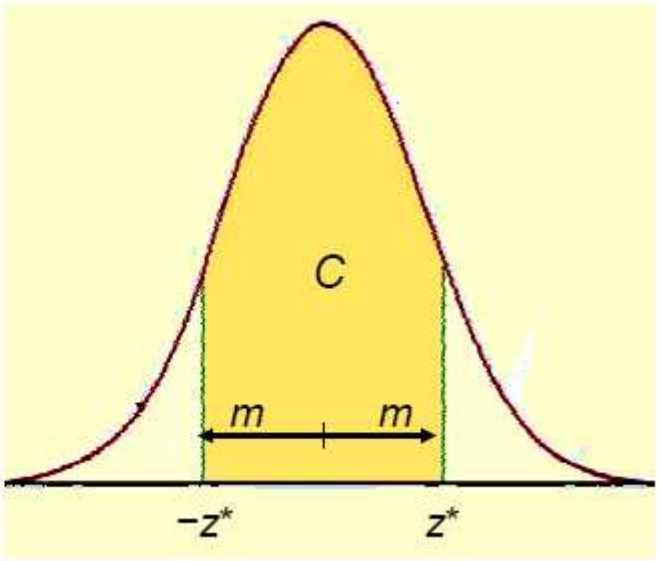
The confidence level C determines the value of z^* (in table C).

The margin of error also depends on z^*.

$$m = z^* \sigma / \sqrt{n}$$

Higher confidence ***C*** implies a larger margin of error ***m*** (thus less precision in our estimates).

A lower confidence level ***C*** produces a smaller margin of error ***m*** (thus better precision in our estimates).



Different confidence intervals for the same set of measurements

Density of bacteria in solution:

Measurement equipment has standard deviation $\sigma = 1 * 10^6$ bacteria/ml fluid.

Three measurements: 24, 29, and 31 * 10^6 bacteria/ml fluid

Mean: $\bar{x} = 28 * 10^6$ bacteria/ml. Find the 96% and 70% CI.

- 96% confidence interval for the true density, $z^* = 2.054$, and write

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 28 \pm 2.054(1/\sqrt{3})$$

$$= 28 \pm 1.19 \times 10^6 \text{ bacteria/ml}$$

- 70% confidence interval for the true density, $z^* = 1.036$, and write

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 28 \pm 1.036(1/\sqrt{3})$$

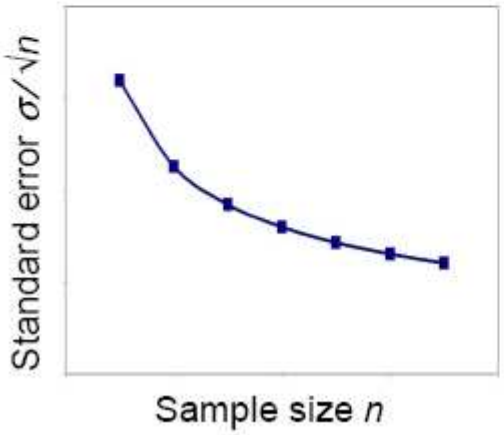
$$= 28 \pm 0.60 \times 10^6 \text{ bacteria/ml}$$

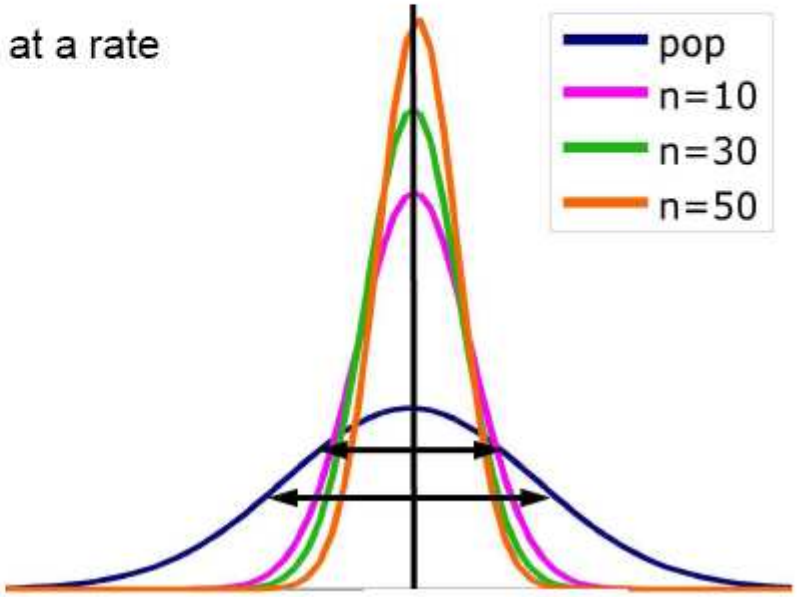
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Impact of sample size

The spread in the sampling distribution of the mean is a function of the number of individuals per sample.

- The larger the sample size, the smaller the standard deviation (spread) of the sample mean distribution.
- But the spread only decreases at a rate equal to $\sqrt{n}$.





Sample size and experimental design

You may need a certain margin of error (e.g., drug trial, manufacturing specs). In many cases, the population variability (σ) is fixed, but we can choose the number of measurements (n).

So plan ahead what sample size to use to achieve that margin of error.

$$m = z^* \frac{\sigma}{\sqrt{n}} \quad \Leftrightarrow \quad n = \left(\frac{z^* \sigma}{m} \right)^2$$

Remember, though, that sample size is not always stretchable at will. There are typically costs and constraints associated with large samples. The best approach is to use the smallest sample size that can give you useful results.

What sample size for a given margin of error?

Density of bacteria in solution:

Measurement equipment has standard deviation $\sigma = 1 * 10^6$ bacteria/ml fluid.

How many measurements should you make to obtain a margin of error of at most $0.5 * 10^6$ bacteria/ml with a confidence level of 90%?

For a 90% confidence interval, $z^* = 1.645$.

$$n = \left(\frac{z^*\sigma}{m}\right)^2 \Rightarrow n = \left(\frac{1.645*1}{0.5}\right)^2 = 3.29^2 = 10.8241$$

Using only 10 measurements will not be enough to ensure that m is no more than 0.5 * 106. Therefore, we need at least 11 measurements.

z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

3.2 Methods of finding confidence intervals

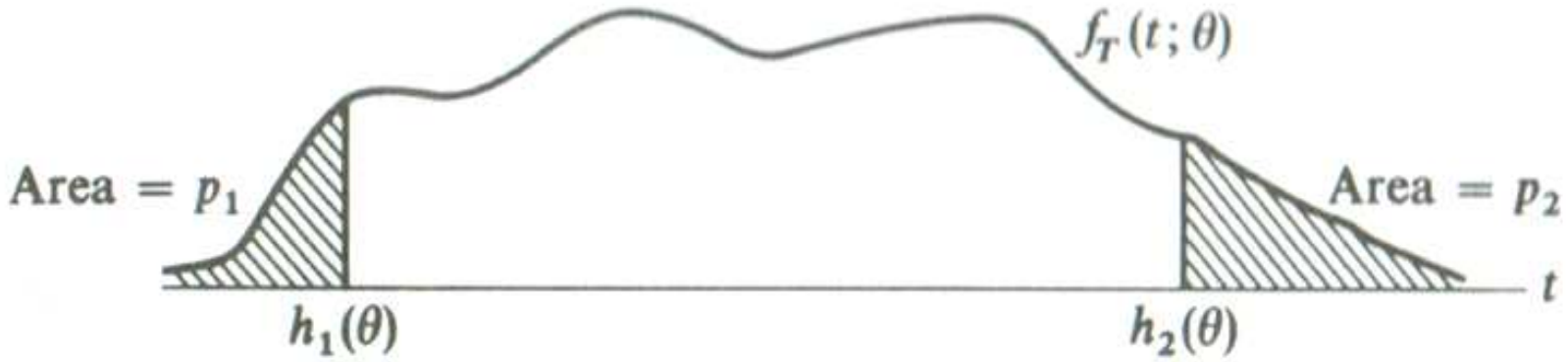
Statistical method

As usual, we assume that we have a random sample $X_1, \ldots, X_n$ from the density $f(\cdot\,; \theta_0)$. We further assume that the parameter θ_0 is real and that the parameter space $\overline{\Theta}$ is some interval. (In this subsection, we will let θ_0 denote the true parameter value.) We seek an interval estimate of θ_0 itself. Let $T = t(X_1, \ldots, X_n)$ be some statistic. The statistic T can be selected in several ways. For instance, if a sufficient statistic (unidimensional) exists, then T could be taken to be a sufficient statistic; or if no sufficient statistic exists, T could be taken to be a point estimator, possibly the maximum-likelihood estimator, of θ_0. The actual choice of T might depend on the ease with which the operations that need to be performed to obtain the confidence interval can be performed. One of those operations will be the determination of the density of T.

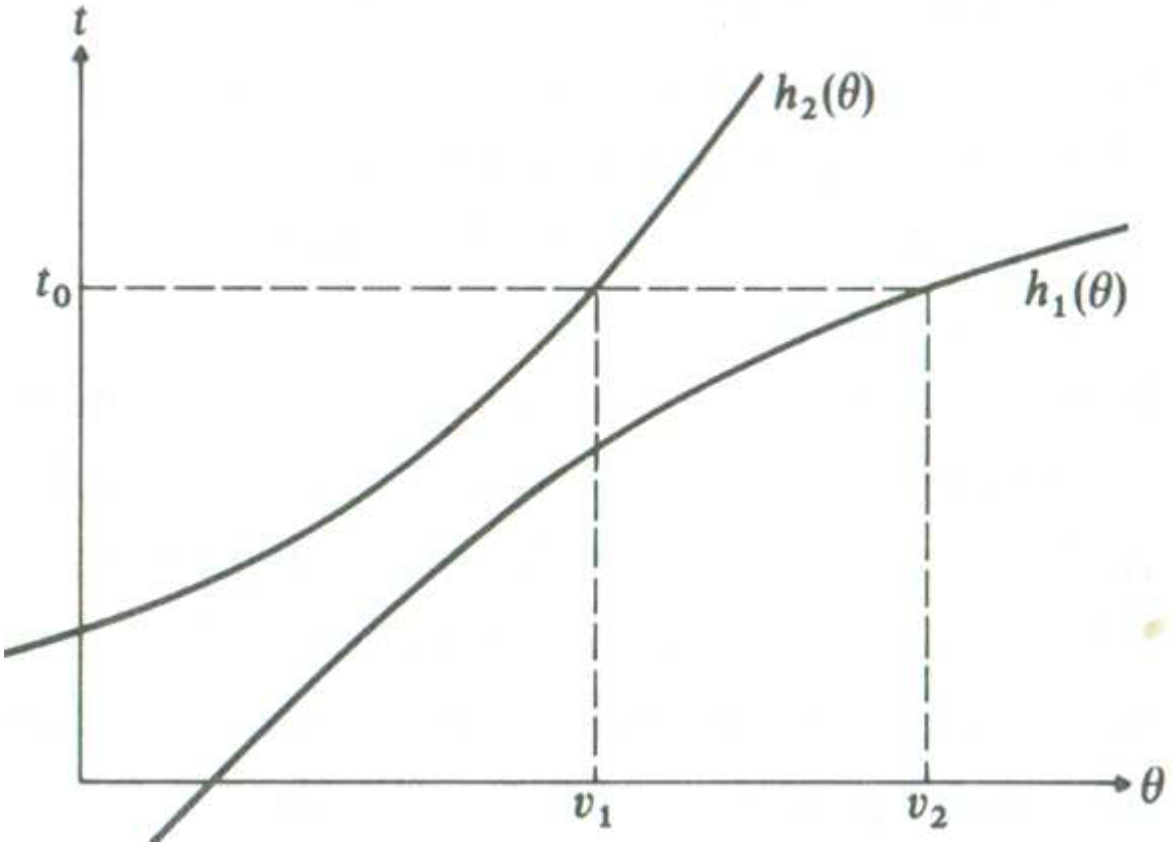
Let $f_T(t;\theta)$ denote the density of T. We will proceed as though T is a continuous random variable; although the technique will also work for T as a discrete random variable. We can define two functions, say $h_1(\theta)$ and $h_2(\theta)$, as follows:

$$\int_{-\infty}^{h_1(\theta)} f_T(t;\theta)\,dt = p_1 \qquad \text{and} \qquad \int_{h_2(\theta)}^{\infty} f_T(t;\theta)\,dt = p_2,$$

where p_1 and p_2 are two fixed numbers satisfying $0 < p_1, 0 < p_2$, and $p_1 + p_2 < 1$.

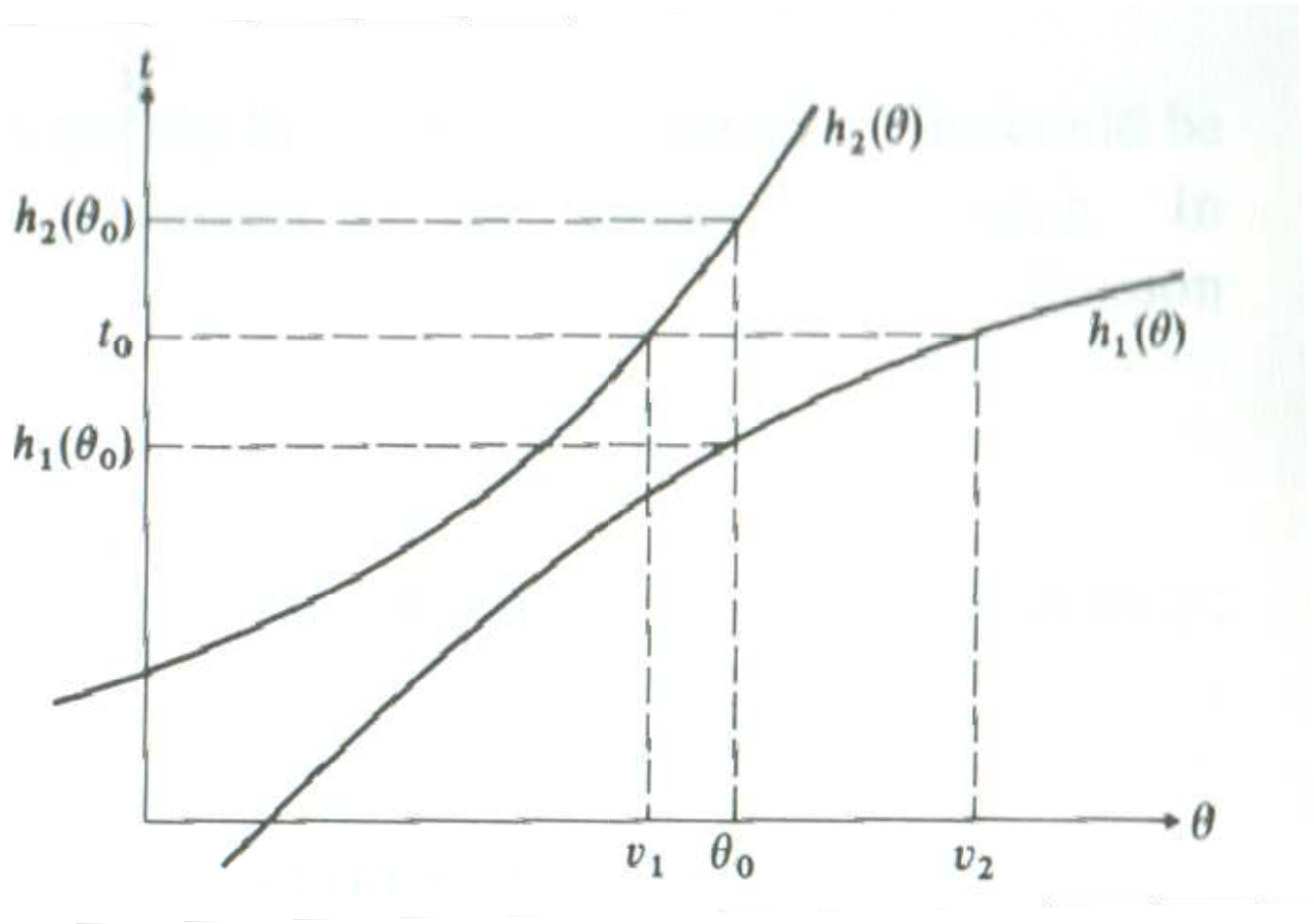


$h_1(\theta)$ and $h_2(\theta)$ can be plotted as functions of θ. We will assume that both $h_1(\cdot)$ and $h_2(\cdot)$ are strictly monotone, and for our sketch we will assume that they are monotone, increasing functions. We know that $h_1(\theta) < h_2(\theta)$.



(Figure A)

Let t_0 denote an observed value of T; that is, $t_0 = t(x_1, \ldots, x_n)$ for an observed random sample $x_1, \ldots, x_n$. Plot the value of t_0 on the vertical axis in Fig. A and then find v_1 and v_2 as indicated. For any possible value of t_0, a corresponding v_1 and v_2 can be obtained, so v_1 and v_2 are functions of t_0; denote these by $v_1 = v_1(t_0)$ and $v_2 = v_2(t_0)$. The interval (V_1, V_2) will turn out to be a $100(1 - p_1 - p_2)$ percent confidence interval for θ_0.



(Figure B)

We see from Fig. B that $h_1(\theta_0) < t_0 = \ell(x_1, \ldots, x_n) < h_2(\theta_0)$ if and only if $v_1 = v_1(x_1, \ldots, x_n) < \theta_0 < v_2 = v_2(x_1, \ldots, x_n)$ for any possible observed sample $(x_1, \ldots, x_n)$. But by definition of $h_1(\cdot)$ and $h_2(\cdot)$,

$$P_{\theta_0}[h_1(\theta_0) < \ell(X_1, \ldots, X_n) < h_2(\theta_0)] = 1 - p_1 - p_2;$$

so

$$P_{\theta_0}[v_1(X_1, \ldots, X_n) < \theta_0 < v_2(X_1, \ldots, X_n)] = 1 - p_1 - p_2;$$

that is, as stated, (V_1, V_2) is a $100(1 - p_1 - p_2)$ percent confidence interval for θ_0, where $V_i = v_i(X_1, \ldots, X_n)$ for $i = 1, 2$.

4 Hypothesis testing

4.1 Tests of significance

Reasoning of significance tests

We have seen that the properties of the sampling distribution of $\bar{x}$ help us estimate a range of likely values for population mean μ.

We can also rely on the properties of the sample distribution to test hypotheses.

Example: You are in charge of quality control in your food company. You sample randomly four packs of cherry tomatoes, each labeled 1/2 lb. (227 g).

The average weight from your four boxes is 222 g. Obviously, we cannot expect boxes filled with whole tomatoes to all weigh exactly half a pound. Thus,

- Is the somewhat smaller weight simply due to chance variation?
- Is it evidence that the calibrating machine that sorts cherry tomatoes into packs needs revision?

Stating hypotheses

A test of statistical significance tests a specific hypothesis using sample data to decide on the validity of the hypothesis.

In statistics, a **hypothesis** is an assumption or a theory about the characteristics of one or more variables in one or more populations.

What you want to know: Does the calibrating machine that sorts cherry tomatoes into packs need revision?

The same question reframed statistically: Is the population mean μ for the distribution of weights of cherry tomato packages equal to 227 g (i.e., half a pound)?

The **null hypothesis** is a very specific statement about a parameter of the population(s). It is labeled H_0.

The **alternative hypothesis** is a more general statement about a parameter of the population(s) that is exclusive of the null hypothesis. It is labeled H_a.

Weight of cherry tomato packs:

H_0 : μ = 227 g (μ is the average weight of the population of packs)

H_a : $\mu \neq$ 227 g (μ is either larger or smaller)

One-sided and two-sided tests

- A **two-tail** or **two-sided test** of the population mean has these null and alternative hypotheses:

H_0: μ = [a specific number] H_a: $\mu \neq$ [a specific number]

- A **one-tail** or **one-sided test** of a population mean has these null and alternative hypotheses:

H_0: μ = [a specific number] H_a: $\mu <$ [a specific number] OR

H_0: μ = [a specific number] H_a: $\mu >$ [a specific number]

The FDA tests whether a generic drug has an absorption extent similar to the known absorption extent of the brand-name drug it is copying. Higher or lower absorption would both be problematic, thus we test:

$H_0: \mu_{generic} = \mu_{brand}$ $H_a: \mu_{generic} \neq \mu_{brand}$ two-sided

How to choose?

What determines the choice of a one-sided versus a two-sided test is what we know about the problem before we perform a test of statistical significance.

A health advocacy group tests whether the mean nicotine content of a brand of cigarettes is greater than the advertised value of 1.4 mg.

Here, the health advocacy group suspects that cigarette manufacturers sell cigarettes with a nicotine content higher than what they advertise in order to better addict consumers to their products and maintain revenues.

Thus, this is a one-sided test: $H_0 : \mu = 1.4$ mg $H_a : \mu > 1.4$ mg

It is important to make that choice before performing the test or else you could make a choice of "convenience" or fall into circular logic.

The p-value

The packaging process has a known standard deviation $\sigma = 5$ g.

$H_0 : \mu = 227$ g versus $H_a : \mu \neq 227$ g

The average weight from your four random boxes is 222 g.

What is the probability of drawing a random sample such as yours if H_0 is true?

Tests of statistical significance quantify the chance of obtaining a particular random sample result if the null hypothesis were true. This quantity is the **P-value.**

This is a way of assessing the "believability" of the null hypothesis, given the evidence provided by a random sample.

Interpreting a p-value

Could random variation alone account for the difference between the null hypothesis and observations from a random sample?

- A small P-value implies that random variation due to the sampling process alone is not likely to account for the observed difference.

- With a small p-value we **reject H_0.** The true property of the population is **significantly** different from what was stated in H_0.

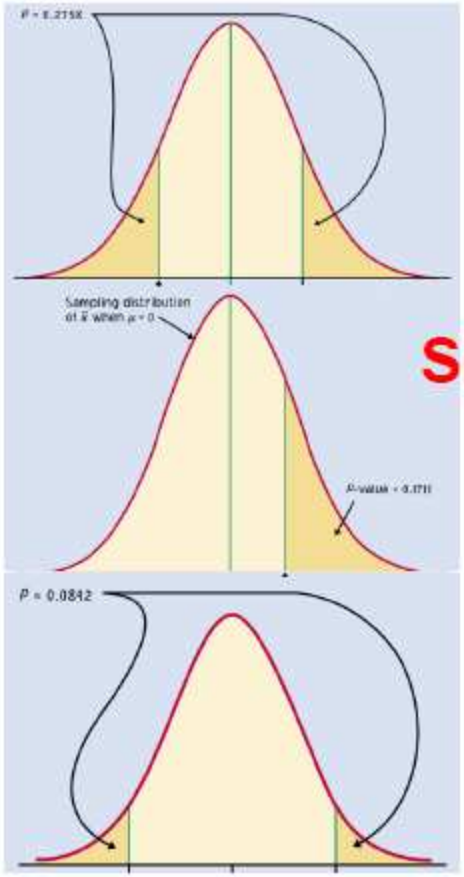
Thus, small P-values are strong evidence AGAINST H_0.

But how small is small…?

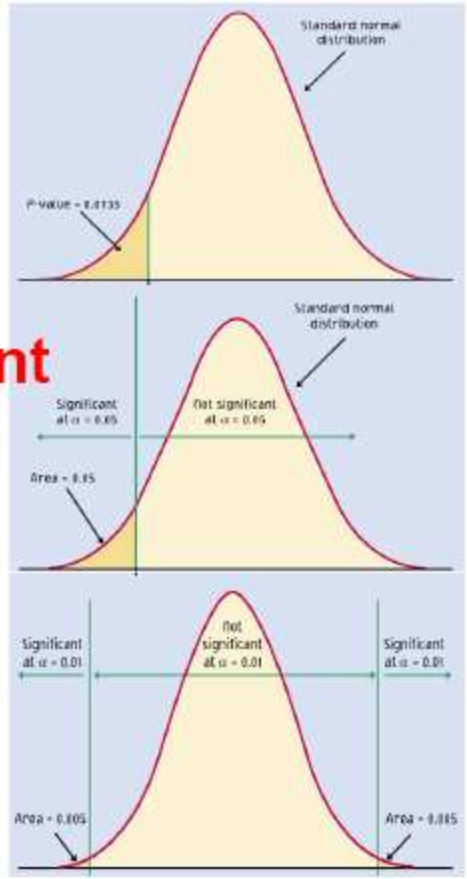
P = 0.2758

P = 0.1711

P = 0.0892



Significant P-value ???



P = 0.0735

P = 0.05

P = 0.01

When the shaded area becomes very small, the probability of drawing such a sample at random gets very slim. Oftentimes, a P-value of 0.05 or less is considered **significant**: The phenomenon observed is unlikely to be entirely due to chance event from the random sampling.

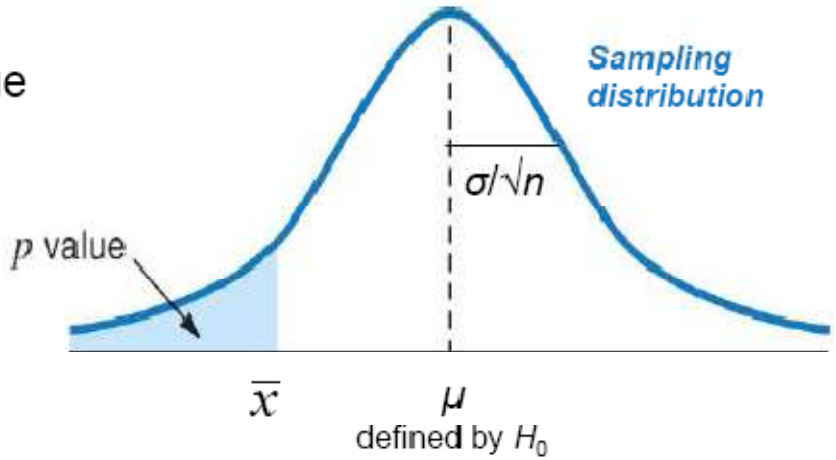
Tests for a population mean

To test the hypothesis $H_0 : \mu = \mu_0$ based on an SRS of size n from a Normal population with unknown mean μ and known standard deviation σ, we rely on the properties of the sampling distribution $N(\mu, \sigma/\sqrt{n})$.

The P-value is the area under the sampling distribution for values at least as extreme, in the direction of H_a, as that of our random sample.

Again, we first calculate a z-value and then use Table A.

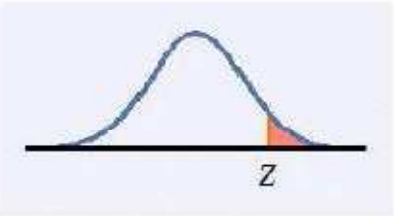
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$



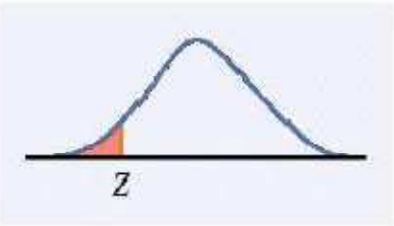
P-value in one-sided and two-sided tests

One-sided (one-tailed) test

$H_a: \mu > \mu_0$ is $P(Z \geq z)$

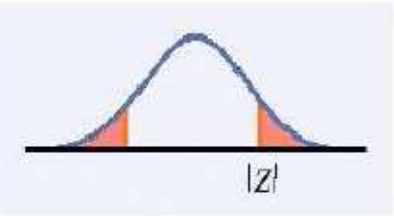


$H_a: \mu < \mu_0$ is $P(Z \leq z)$



Two-sided (two-tailed) test

$H_a: \mu \neq \mu_0$ is $2P(Z \geq |z|)$



To calculate the P-value for a two-sided test, use the symmetry of the normal curve. Find the P-value for a one-sided test and double it.

Does the packaging machine need revision?

- $H_0: \mu = 227$ g versus $H_a: \mu \neq 227$ g
- What is the probability of drawing a random sample such as yours if H_0 is true?

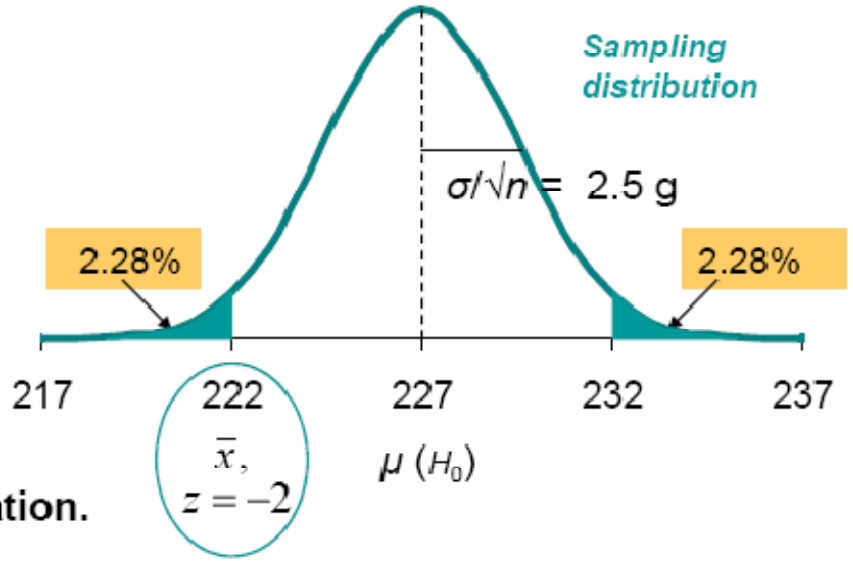
$$\bar{x} = 222\text{g} \quad \sigma = 5\text{g} \quad n = 4 \qquad z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{222 - 227}{5/\sqrt{4}} = -2$$

From table A, the area under the standard normal curve to the left of z is 0.0228.

Thus, P-value = 2*0.0228 = 4.56%.

The probability of getting a random sample average so different from μ is so low that we reject H_0.

→The machine does need recalibration.



The significance level α

The significance level, α, is the largest P-value tolerated for rejecting a true null hypothesis (how much evidence against H_0 we require). This value is decided arbitrarily before conducting the test.

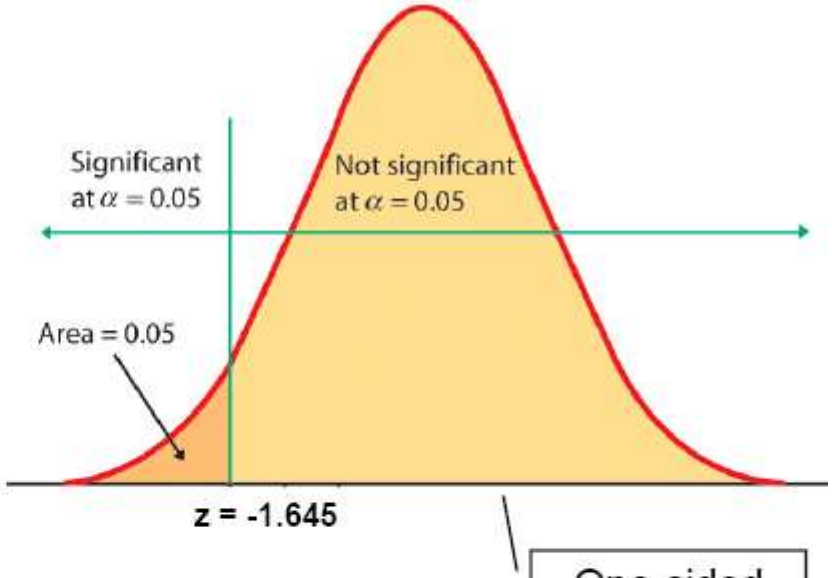
- If the P-value is equal to or less than α ($P \leq \alpha$), then we **reject H_0**.
- If the P-value is greater than α ($P > \alpha$), then we **fail to reject H_0**.

Does the packaging machine need revision?

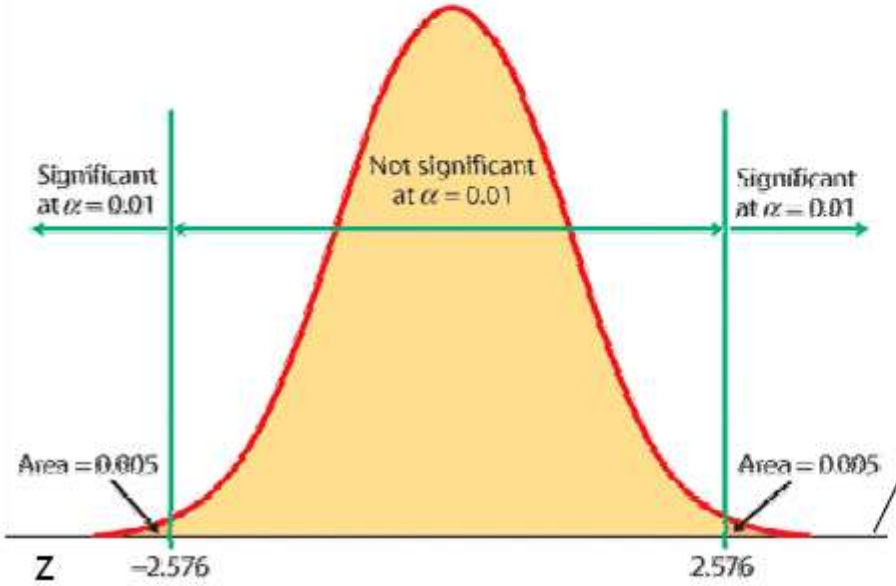
Two-sided test. The P-value is 4.56%.

* If α had been set to 5%, then the P-value would be significant.
* If α had been set to 1%, then the P-value would not be significant.

When the z score falls within the rejection region (shaded area on the tail-side), the p-value is smaller than α and you have shown statistical significance.



One-sided test, α = 5%



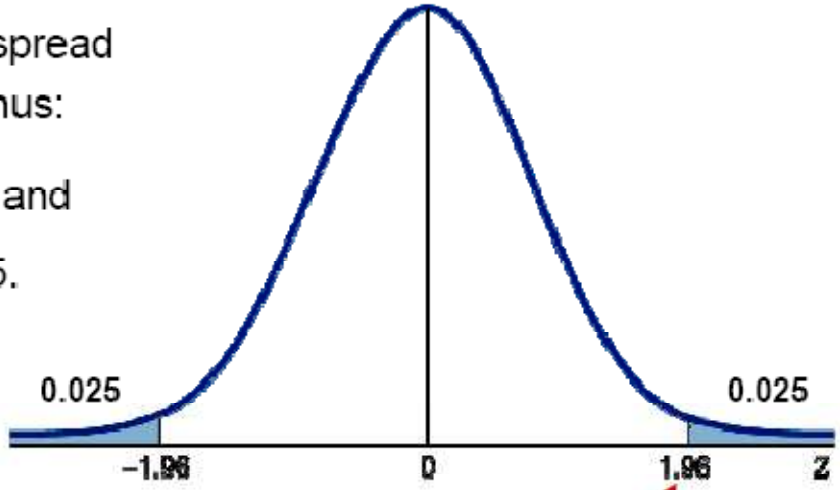
Two-sided test, α = 1%

Rejection region for a two-tail test of μ with $\alpha = 0.05$ (5%)

A two-sided test means that α is spread between both tails of the curve, thus:

-A middle area *C* of $1 - \alpha = 95\%$, and

-An upper tail area of $\alpha/2 = 0.025$.



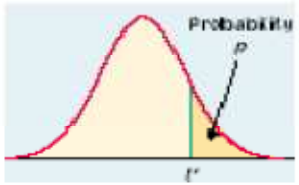


Table C

upper tail probability p	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
(…)												
z*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
Confidence interval C	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%

Confidence intervals to test hypotheses

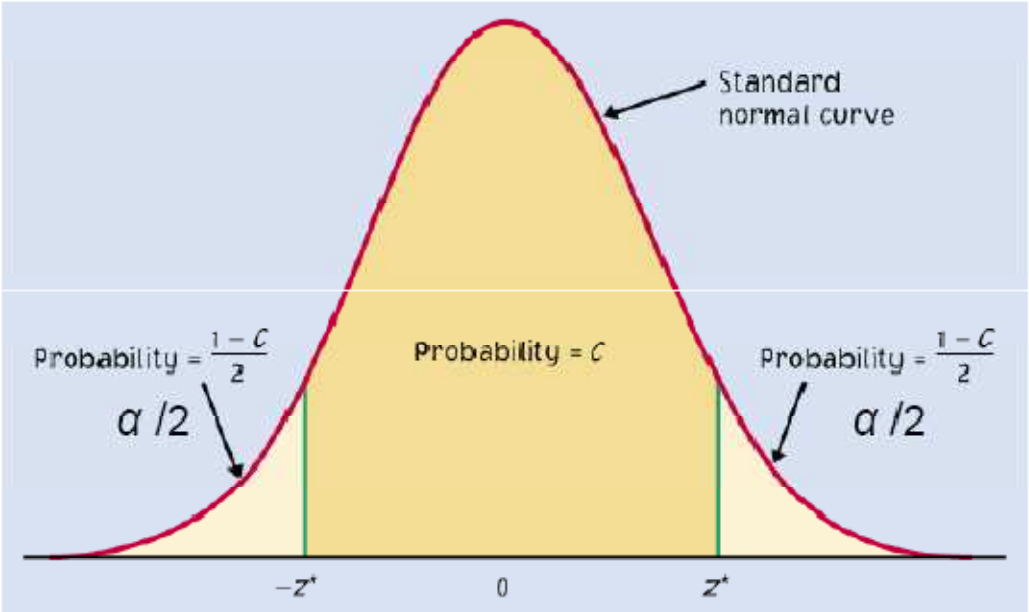
Because a two-sided test is symmetrical, you can also use a confidence interval to test a two-sided hypothesis.

In a two-sided test,

$C = 1 - \alpha$.

C confidence level

α significance level



Packs of cherry tomatoes (σ = 5 g): H_0 : μ = 227 g versus H_a : $\mu \neq$ 227 g

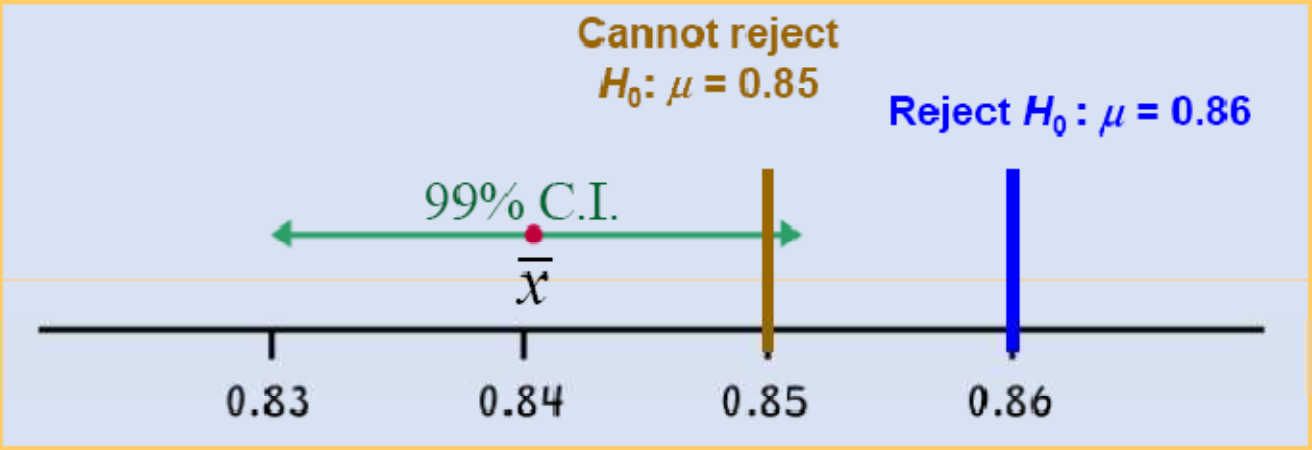
Sample average 222 g. 95% CI for $\mu = 222 \pm 1.96*5/\sqrt{4} = 222 \text{ g} \pm 4.9$ g

227 g does not belong to the 95% CI (217.1 to 226.9 g). Thus, we reject H_0.

Logic of confidence interval test

Ex: Your sample gives a 99% confidence interval of $\bar{x} \pm m = 0.84 \pm 0.0101$.

With 99% confidence, could samples be from populations with μ = 0.86? μ = 0.85?



A confidence interval gives a black and white answer: Reject or don't reject H_0.

But it also estimates a range of likely values for the true population mean μ.

A P-value quantifies how strong the evidence is against the H_0. But if you reject H_0, it doesn't provide any information about the true population mean μ.

More about the relationship to confidence intervals

Our engineer wants to know if the modification introduces an extra variance in the motors

Unmodified motor (Watts): 755 750 730 731 743 $s_X = 11.17$

Modified motor (Watts): 742 738 723 721 730 $s_Y = 9.28$

$$\frac{\sigma_X^2}{\sigma_Y^2} \in \left[\frac{\frac{11.17^2}{9.28^2}}{9.6}, \frac{\frac{11.17^2}{9.28^2}}{0.1} \right] = \left[0.39^2, 3.8^2 \right]_{\alpha = 0.05}$$

$$H_0 : \sigma_X^2 = \sigma_Y^2 \qquad H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1$$

$$H_1 : \sigma_X^2 \neq \sigma_Y^2 \qquad H_1 : \frac{\sigma_X^2}{\sigma_Y^2} \neq 1$$

If the confidence interval includes the null hypothesis, then we cannot reject H_0.
If the confidence interval does not include the null hypothesis, then we reject H_0.

Do not misinterpret confidence intervals, overlapping confidence intervals do not mean that we cannot reject the null hypothesis.

Our engineer wants to know if the modification introduces an extra variance in the motors

Unmodified motor (Watts): 755 750 740 741 743

Modified motor (Watts): 732 738 723 721 730

$\bar{x} = 745.8 \quad s_X = 6.45$

$\bar{y} = 728.8 \quad s_Y = 6.91$

$\bar{d} = -17$

$$\mu_X \in \bar{x} \pm t_{N_X - 1, \frac{\alpha}{2}} \frac{s_X}{\sqrt{N_X}} \longrightarrow \mu_X \in [727.9, 763.7]$$

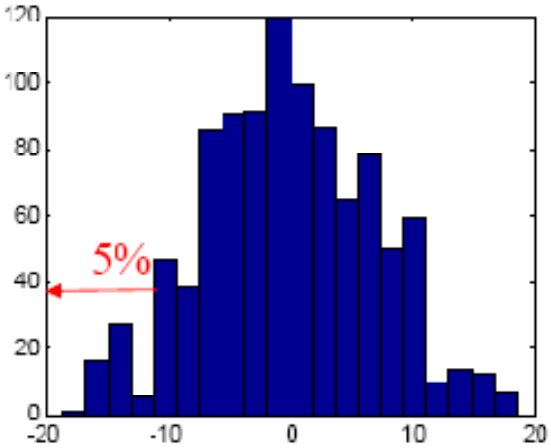
$$\mu_Y \in [709.6, 748.0]$$

$$\frac{\bar{d} - \mu_d}{\sqrt{\frac{s_X^2}{N} + \frac{s_Y^2}{N}}} \sim t_{2N-2} \longrightarrow \mu_d \in [-26.8, -7.2] \longrightarrow \text{Reject } H_0 \text{ because } \mu_d = 0 \text{ does not belong to the confidence interval}$$

Permutation / randomization tests

Our engineer wants to know if the modification reduces the consumption

	Unmodified					Modified					Diff
Actual measurement	755	750	730	731	743 (741.8)	742	738	723	721	730 (730.8)	-11
Permutation 1	742	731	721	755	730 (735.8)	750	723	730	743	738 (736.8)	1
Permutation 2	742	721	750	730	755 (739.6)	743	730	731	723	738 (733.0)	-6.6
Permutation 3	750	755	742	730	743 (744.0)	721	723	731	730	738 (728.6)	-15.4
…	…										



$$\Pr\{\hat{\mu} < -11\} = 0.05$$

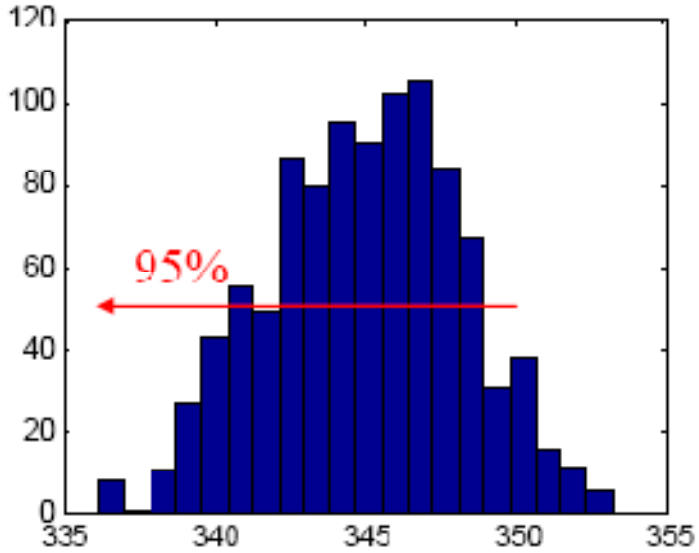
Bootstrap tests

Is it possible that the machine is filling less than programmed?

Data:

341 335 354 345 350 ⟶

354 350 350 354 350 → 351.6
341 335 345 341 335 → 339.4
335 335 354 335 335 → 338.8
341 345 354 350 354 → 348.8
...



$H_0 : \mu \geq 348$ ⟶ $\Pr\{\hat{\mu} \geq 348\} = 0.166$

$H_1 : \mu < 348$

Reject H_0 if the confidence interval at confidence level 1-α does not intersect the null region.

4.2 Hypothesis types

Simple hypothesis and Composite hypothesis

- A simple hypothesis is a hypothesis which specifies the population distribution completely.
- Examples
  1. H0: X ~ Bi(100,1/2), i.e. p is specified
  2. H0: X ~ N(5,20), i.e. μ and σ^2 are specified
- A composite hypothesis is a hypothesis which does not specify the population distribution completely.
- Examples
  1. X ~ Bi(100,p) and H1: p > 0.5
  2. X ~ N(0, σ^2) and H1: σ^2 unspecified