Basic Structure of Inference

Statistical Inference makes educated guesses about the population based on information from the sample. All guesses are prone to error, and the quantification of imprecision is an important part of statistical inference.

1. **Estimation** estimates the state of population, which is typically characterized by some parameter, say θ.
2. **Hypothesis testing** chooses from among postulated states of population, such as $H_0: \ \theta = \theta_0$ versus $H_a: \ \theta \neq \theta_0$, where θ_0 is a known number.

Examples of Estimation and Testing

A plant physiologist grew 13 soybean seedlings of the type "Wells II". She measured the total stem length (cm) for each plant after 16 days of growth, and got $\bar{y} = 21.34$ and $s = 1.22$.

She may estimate the "average" stem length by a **point estimate**,

$$\mu \approx 21.34,$$

or by an **interval estimate**,

$$18.68 < \mu < 24.00.$$

As reported by AMA, 16 out of every 100 doctors in any given year are subject to malpractice claims. A hospital of 300 physicians received claims against 58 of their doctors in one year. Was the hospital simply "unlucky"? Or does the number possibly indicate some systematic "wrongdoings" at the hospital?

The number 58/300 is "within chance variation" of $\theta_0 = .16$.

Estimating Population Mean

Observing $X_1, \ldots, X_n$ from a population with mean μ and variance σ^2, one is to estimate μ. The procedure (or formula) one uses is called an **estimator**, which yields an **estimate** after the data are plugged in.

Observing $X_1, \ldots, X_5$, one may use one of the following **point estimators** for μ:

$$\hat{\mu}_1 = \bar{X}$$

$$\hat{\mu}_2 = X_1$$

$$\hat{\mu}_3 = (X_1 + X_3)/2$$

$$\hat{\mu}_4 = \text{Median}$$

$$\hat{\mu}_5 = \mu_0$$

Observing 5.1, 5.1, 5.3, 5.2, 5.2, one may use one of the following **point estimates** for μ:

$$\hat{\mu}_1 = \bar{x} = 5.18$$

$$\hat{\mu}_2 = x_1 = 5.1$$

$$\hat{\mu}_3 = (x_1 + x_3)/2 = 5.2$$

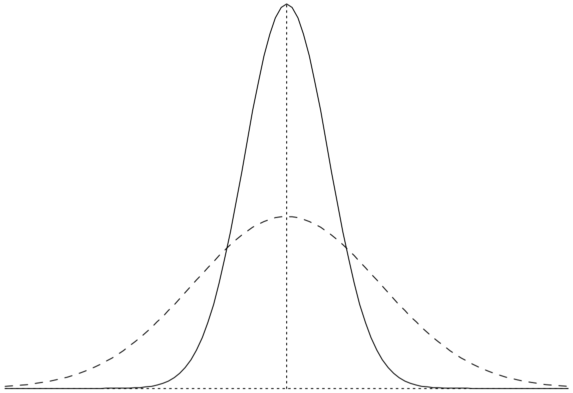
$$\hat{\mu}_4 = \text{Median} = 5.2$$

$$\hat{\mu}_5 = 5$$

Properties of Point Estimators

To choose among all possible estimators, one compares properties of the estimators.

- Unbiasedness: $\mu_{\hat{\theta}} = \theta$.
- Small SD: $\sigma_{\hat{\theta}}$.



$\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\mu}_3$ are all unbiased.

$$\begin{aligned} \mu_{\bar{X}} &= \mu_{X_1} = \mu_{(X_1+X_3)/2} = \mu \\ \sigma^2_{\bar{X}} &= \sigma^2/5 \\ \sigma^2_{X_1} &= \sigma^2 \\ \sigma^2_{(X_1+X_3)/2} &= \sigma^2/2 \end{aligned}$$

- A better estimator yields better estimates on average.
- A better estimator may *not* always yield a better estimate.

Mean Versus Median: A Simulation Study

Consider the estimation of the population mean/median μ of a symmetric distribution by the sample mean and the sample median.

```
x <- matrix(rnorm(100000),ncol=10000)
mn <- apply(x,2,mean) ## sample mean of 10
md <- apply(x,2,median)  ## sample median of 10
mean(mn); mean(md)  ## unbiasedness
mean(mn^2); mean(md^2)  ## mean is a better estimator
plot(density(mn)); xx <- seq(-1,1,len=101)
lines(xx,dnorm(xx,0,1/sqrt(10)),col=3)
lines(density(md),col=5); abline(v=0,lty=2)
sum(abs(mn)<abs(md))  ## mean can be the worse estimate
```

Sample Mean As Estimator of Population Mean

One usually uses the sample mean $\bar{x}$ to estimate the population mean μ, as $\bar{X}$ has the smallest standard deviation among all unbiased estimators of μ.

To quantify the imprecision of the estimation of μ by $\bar{x}$, one estimates $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ by $\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}}$, the **standard error of the sample mean $\bar{X}$**.

Soybean stem length: $n = 13$, $\bar{x} = 21.34$, and $s = 1.22$.

$$\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{1.22}{\sqrt{13}} = .338$$

- For $\bar{X}$ nearly normal, $\bar{x}$ lies within $\pm 2\frac{\sigma}{\sqrt{n}}$ of μ about 95% of the time.
- Do not confuse $\sigma_{\bar{X}}$, $\hat{\sigma}_{\bar{X}}$ with σ.

Confidence Intervals

A point estimate will almost surely miss the target, although its standard error indicates by how far the miss is likely to be. An **interval estimate** provides a range for the parameter estimate.

Soybean stem length: *Assume* normality with $\sigma = 1.2$ *known*. One has $\bar{X} \sim N(\mu, (1.2)^2/13)$, so

$$P(\frac{|\bar{X}-\mu|}{1.2/\sqrt{13}} \le 1.96) = .95.$$

Solving for μ, one obtains

$$\bar{X} - 1.96\frac{1.2}{\sqrt{13}} \le \mu \le \bar{X} + 1.96\frac{1.2}{\sqrt{13}}$$

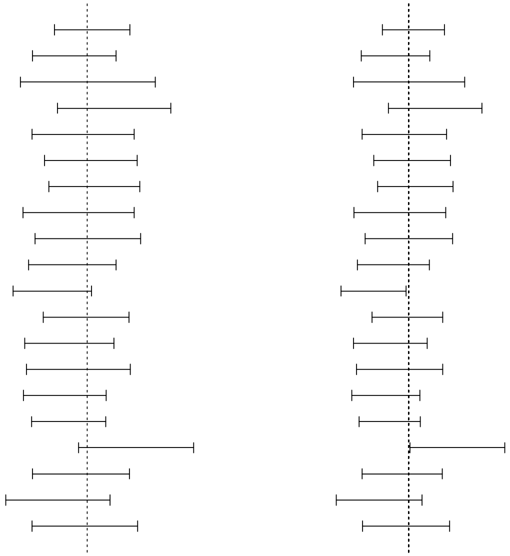
For $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$ with σ^2 known,

$$\boxed{\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

provides an interval estimator that covers μ with probability $(1-\alpha)$. It yields a $(1-\alpha)100\%$ **confidence interval** for μ, with a **confidence coefficient** $(1-\alpha)100\%$.

Coverage, Large Sample CIs

As an *estimator*, a CI is a moving bracket "chasing" a fixed target. As an *estimate*, a CI may or may not cover the "truth".



With a **large sample** from an "arbitrary" distribution for σ unknown, an confidence interval for μ with an approximate conf. coef. $(1-\alpha)100\%$ is given by

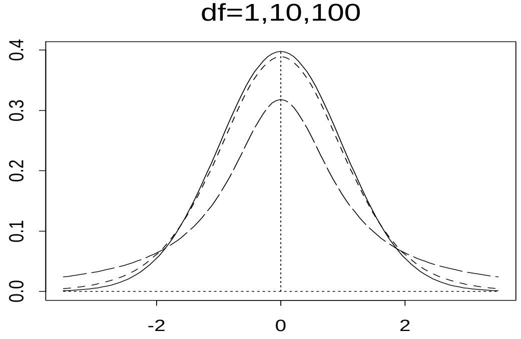
$$\boxed{\boxed{\bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}}}.$$

- Normality comes from CLT.
- Unknown σ estimated by s.
- Replace s by σ if known.

Small Sample CIs based on t-Distribution

For a **small sample** with σ unknown, one "*has to*" assume normality.

Consider $Z_i \sim N(0,1)$, $i = 1, \dots, n$. The distribution of $\frac{\bar{Z}}{s/\sqrt{n}}$ is called a ***t*-distribution** with a **degree of freedom** (df) $\nu = n - 1$. A t-distribution with $\nu = \infty$ reduces to $N(0,1)$.



For $X_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$, $P(\frac{|\bar{X} - \mu|}{s/\sqrt{n}} \le t_{\alpha/2,n-1}) = 1 - \alpha$, so

$$\boxed{\bar{X} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}}$$

provides a $(1 - \alpha)100\%$ CI for μ.

- $t_{\alpha,\nu} \downarrow$ as $\nu \uparrow$: more data allow more accurate $\hat{\sigma}$, hence tighter bounds.
- For σ known, use $z_{\alpha/2}$ and σ.

Simulations of Coverage

```
x <- matrix(rnorm(100000),ncol=10000)
mn <- apply(x,2,mean); v <- apply(x,2,var)
hwd <- 1.96/sqrt(10); lcl <- mn-hwd; ucl <- mn+hwd
mean((lcl<0)&(ucl>0))  ## z-interval
hwd<-qt(.975,9)*sqrt(v/10); lcl<-mn-hwd; ucl<-mn+hwd
mean((lcl<0)&(ucl>0))  ## t-interval

x <- matrix(runif(100000),ncol=10000)
mn <- apply(x,2,mean); v <- apply(x,2,var)
hwd<-1.96*sqrt(1/120); lcl<-mn-hwd; ucl<-mn+hwd
mean((lcl<.5)&(ucl>.5))  ## z-interval
hwd<-qt(.975,9)*sqrt(v/10); lcl<-mn-hwd; ucl<-mn+hwd
mean((lcl<.5)&(ucl>.5))  ## t-interval
```

Confidence Intervals for μ: Summary

An agronomist measured stem diameter (mm) in 8 plants of a variety of wheat, and calculated $\bar{x} = 2.275$ and $s = .2375$.

Assuming normality, a 95% CI for μ is given by

$$2.275 \pm 2.365(.2375)/\sqrt{8},$$

or $(2.076, 2.474)$, where $t_{.025,7} = 2.365$. If one further knows that $\sigma = .25$, then he can use

$$2.275 \pm 1.96(.25)/\sqrt{8},$$

or $(2.102, 2.448)$.

- In the "ideal" situation with normality and known σ, always use

$$\boxed{\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

- With a small normal sample but unknown σ, estimate σ by s and replace $z_{\alpha/2}$ by $t_{\alpha/2,n-1}$ to allow for the uncertainty.
- When n is large, CLT grants normality, s estimates σ reliably, and $z_{\alpha/2} \approx t_{\alpha/2,n-1}$.
- The procedures may work under violated assumptions.

Coverage versus Precision

To cover the truth more often, one needs a higher confidence coefficient, but at the expense of wider intervals.

- The interval $(-\infty, \infty)$ has 100% coverage but is useless.
- A point estimate is the most precise but always misses.
- Given sample size n, $\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$ is the shortest interval estimate for μ among all that have a confidence coefficient $(1-\alpha)100\%$.
- To achieve both coverage and precision, one has to take a large enough sample.

Planning Sample Size

The agronomist is planning a new study of wheat stem diameter, and wants a 95% CI of μ no wider than .2 mm. From experience and pilot study, he believes that $\sigma = .25$ is about right.

The half-width of CI is

$$z_{.025}\frac{\sigma}{\sqrt{n}} = 1.96\frac{.25}{\sqrt{n}}.$$

Solving for n from

$$1.96(.25)/\sqrt{n} \leq .1,$$

one gets $\boxed{n \geq 24}$.

Let h be the desired half-width for a $(1-\alpha)100\%$ CI. Solving for n from $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq h$, one has

$$\boxed{\boxed{n \geq \left(\frac{z_{\alpha/2}\sigma}{h}\right)^2}}$$

- For n large, $z_{\alpha/2} \approx t_{\alpha/2,n-1}$.
- Need a conservative estimate of σ.
- To cut the width by half, one needs to quadruple the sample size n.

CI for Population Proportion

123 adult female deer were captured and 97 found to be pregnant. Construct a 95% CI for pregnant proportion in the population.

Since $\hat{p} = \frac{97}{123} = .7886$, $\hat{\sigma}_{\hat{p}} = \sqrt{\frac{.7886(1-.7886)}{123}} = .0368$, the 95% CI is given by

$$.7886 \pm 1.96(.0368),$$

or $(.7165, .8607)$.

For a 95% CI with half-width $h \leq 3\%$, it is safe to have

$$n \geq (1.96(0.5)/0.03)^2 = 1067.1.$$

Consider $X_i \sim \text{Bin}(1, p)$, $i = 1, \ldots, n$, independent. One has

$$X = \sum_i X_i \sim \text{Bin}(n, p).$$

For n large, by CLT,

$$P\left(\frac{X/n - p}{\sqrt{p(1-p)/n}} \leq z\right) \approx \Phi(z).$$

The sample proportion $\hat{p} = X/n$ is actually an $\bar{X}$. As an estimate of $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$ one may use $\sqrt{\hat{p}(1-\hat{p})/n}$. A $(1-\alpha)100\%$ CI for p is thus

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

- $\sigma = \sqrt{p(1-p)} \leq 0.5$.

Structure of Hypothesis Testing

A drug maker claimed that the mean potency of one of its antibiotics was 80%. A sample of 100 capsules were tested and produced $\bar{x} = 80.2\%$ and $s = .8\%$.

Decide between the alternatives $H_0 : \mu = 80\%$ and $H_a : \mu \neq 80\%$.

If H_0 is true, then by CLT,

$$Z = \frac{\bar{X} - 80}{s/\sqrt{100}} \sim N(0,1).$$

Large magnitude of Z indicates departure of μ from 80%.

One may decide to accept H_0 when $|Z| \leq 1.96$, and reject o.w.

Statistical tests test hypotheses concerning the state of population, which is often characterized by the value of some parameter. A test involves four elements.

1. Null hypothesis H_0: the postulated "default" state.
2. Alternative hypothesis H_a: the "abnormal" state.
3. Test statistic: the empirical information from data.
4. Rejection region: the decision rule.

Formulation of Hypotheses

As will be seen later, it takes “effort” to “prove” the alternative H_a whereas H_0 is assumed but never established, so one usually sets the “hoped-for” or “feared-of” hypothesis as the alternative.

♣ In fear of being cheated, we may set $H_0 : \mu \leq 80\%$ versus $H_a : \mu > 80\%$ for the antibiotic potency, and refuse the product when H_0 is accepted.

♣ To monitor industrial pollution to the environment, the regulating agency may set $H_0 : \mu \geq \mu_0$ versus $H_a : \mu < \mu_0$, and punish the polluter when H_0 is not rejected.

♣ When testing a new drug, one usually presumes (H_0) that the new drug has no effect.

Construction of Tests

Suppose $\sigma = .8$ is known and one is to take a sample of size 100.

For $H_0 : \mu = 80$ vs. $H_a : \mu \neq 80$, one may calculate

$$Z = \frac{\bar{X} - 80}{\sqrt{.8^2/100}} = \frac{\bar{X} - 80}{.08}$$

and reject H_0 when $|Z| > 1.96$, or $\bar{X} < 79.8432$, $\bar{X} > 80.1568$.

- When $\mu = 80$, $Z \sim N(0, 1)$, so $P(\text{reject } H_0 | \mu = 80) = .05$.

For $H_0 : \mu \leq 80$ vs. $H_a : \mu > 80$, one calculates $Z = (\bar{X} - 80)/.08$, and rejects H_0 when $Z > 1.645$, or $\bar{X} > 80.1316$.

- By convention, decision rules are constructed to control *the probability of rejecting a true null* (α-risk) to .05 or .01, and the resulting tests are said to be at the 5% or 1% **significance level**.
- The roles of H_0 and H_a are *asymmetric*. H_0 is protected.
- A test parallels a court room: H_0 corresponds to *innocence*, H_a to *guilt*, the test statistic to *evidence*, and the decision to *verdict*. One is assumed innocent until proven guilty.

Drawing Conclusions from Tests

Potency of antibiotic: $\sigma = .8$, $n = 100$, and $\bar{x} = 80.2$. Test

$$H_0 : \mu = 80 \text{ vs. } H_a : \mu \neq 80.$$

It is easy to calculate

$$z = \frac{80.2 - 80}{.8/\sqrt{100}} = 2.5.$$

Since $|z| = 2.5 > 1.96 = z_{.975}$, one rejects H_0 at the 5%-level.

Since $|z| = 2.5 < 2.576 = z_{.995}$, one accepts H_0 at the 1%-level.

- It takes more than data to reach conclusions.

- With the same empirical evidence, one may draw different conclusions depending on the amount of protection desired for H_0. Smaller α favors H_0.
- Accepting H_0 does *not* imply that H_0 is true, it merely says that one does not have strong enough evidence against H_0.
- With a small sample, one often has to accept H_0, simply because there isn't much information in the data.

p-Values of Tests

p-value is the probability that the test statistic *would* look more "weird" than the observed one if H_0 is true.

Potency of antibiotic: $\sigma = .8$, $n = 100$. Observing $\bar{x} = 80.2$, test

$$H_0: \mu = 80 \quad \text{vs.} \quad H_a: \mu \neq 80, \text{ or}$$
$$H_0: \mu = 80 \quad \text{vs.} \quad H_a: \mu > 80,$$

It was calculated that

$$z = \frac{\bar{x} - 80}{.8/\sqrt{100}} = \frac{80.2 - 80}{.08} = 2.5.$$

The p-values are

$$P(|Z| > 2.5) = .0124, \text{ or}$$
$$P(Z > 2.5) = .0062.$$

- The p-value summarizes the empirical evidence against H_0. The smaller the p-value is, the stronger the evidence is against H_0.
- If $p < \alpha$, one rejects H_0 at the α-level.
- With the same data, one-tailed and two-tailed tests have different p-values. The direction of H_a also matters.

Type-I/Type-II Errors, α/β-Risks

Tests are prone to errors. Consider $H_0 : \mu = 80$ vs. $H_a : \mu \neq 80$.

	$\mu = 80$	$\mu \neq 80$
$\lvert Z \rvert \leq 1.96$	correct decision	**type-II error**
$\lvert Z \rvert > 1.96$	**type-I error**	correct decision

The probabilities of type-I and type-II errors are called the **α-risk** and **β-risk**, respectively.

The α-risk and β-risk are properties of the decision rule, similar to an *estimator*. For a given sample, the decision made, similar to an *estimate*, is either correct or erroneous.

The α-risk and β-risk are in general functions of the true parameter, which is unknown in practice.

α/β-Risks of One-Sided Tests

Consider $H_0 : \mu \le 80$ vs. $H_a : \mu > 80$ with $\sigma = .8$ and $n = 100$. One rejects H_0 when $(\bar{X} - 80)/.08 > 1.645$, or $\bar{X} > 80.1316$.

When $\mu = 80, 79.9$, the α-risks are:

$$\alpha(80.0) = P(\bar{X} > 80.1316) = P(\tfrac{\bar{X}-80}{.08} > \tfrac{80.1316-80}{.08}) = .05,$$

$$\alpha(79.9) = P(\tfrac{\bar{X}-80}{.08} > 1.645) = P(\tfrac{\bar{X}-79.9}{.08} > 1.645 + \tfrac{.1}{.08}) = .0019.$$

All α-risks are no more than .05, the significance level.

When $\mu = 80.1, 80.5$, the β-risks are:

$$\beta(80.1) = P(\bar{X} \le 80.1316) = P(\tfrac{\bar{X}-80.1}{.08} \le \tfrac{80.1316-80.1}{.08}) = .6536,$$

$$\beta(80.5) = P(\tfrac{\bar{X}-80}{.08} \le 1.645) = P(\tfrac{\bar{X}-80.5}{.08} \le 1.645 - \tfrac{.5}{.08}) = 2 \times 10^{-6}.$$

All β-risks are no more than .95, $1 - \alpha$ on the border line.

α/β-Risks of Two-Sided Tests

Consider $H_0 : \mu = 80$ vs. $H_a : \mu \neq 80$ with $\sigma = .8$ and $n = 100$. One accepts H_0 when $|\bar{X} - 80|/.08 \leq 1.96$, or $79.8432 \leq \bar{X} \leq 80.1568$.

The α-risk only makes sense when $\mu = 80$, which is .05 by design.

When $\mu = 79.9$ (and $\mu = 80.1$, by symmetry), the β-risk is:

$$P(79.8432 \leq \bar{X} \leq 80.1568) = P(\tfrac{79.8432-79.9}{.08} \leq \tfrac{\bar{X}-79.9}{.08} \leq \tfrac{80.1568-79.9}{.08})$$
$$= P(-.71 \leq \tfrac{\bar{X}-79.9}{.08} \leq 3.21) = .7605$$

When $\mu = 80.5$, the β-risk is:

$$P(-1.96 \leq \tfrac{\bar{X}-80}{.08} \leq 1.96) = P(-1.96 - \tfrac{.5}{.08} \leq \tfrac{\bar{X}-80.5}{.08} \leq 1.96 - \tfrac{.5}{.08})$$
$$= P(-8.21 \leq \tfrac{\bar{X}-80.5}{.08} \leq -4.29) = 9 \times 10^{-6}.$$

Effects of Sample Size on α/β-Risks

Consider $H_0: \mu \leq 80$ vs. $H_a: \mu > 80$ with $\sigma = .8$ and $\mathbf{n} = \mathbf{25}$. One rejects H_0 when $(\bar{X} - 80)/.16 > 1.645$, or $\bar{X} > 80.2632$.

When $\mu = 80, 79.9$, the α-risks are:

$$\alpha(80.0) = P(\bar{X} > 80.2632) = P(\tfrac{\bar{X}-80}{.16} > \tfrac{80.2632-80}{.16}) = .05,$$

$$\alpha(79.9) = P(\tfrac{\bar{X}-80}{.16} > 1.645) = P(\tfrac{\bar{X}-79.9}{.16} > 1.645 + \tfrac{.1}{.16}) = .0116.$$

Compared to $n = 100$, $.0116 > .0019$.

When $\mu = 80.1, 80.5$, the β-risks are:

$$\beta(80.1) = P(\bar{X} \leq 80.2632) = P(\tfrac{\bar{X}-80.1}{.16} \leq \tfrac{80.2632-80.1}{.16}) = .8461,$$

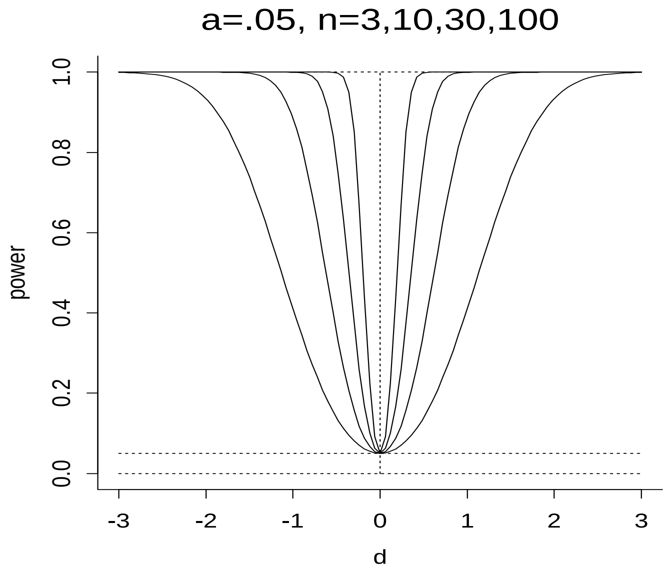
$$\beta(80.5) = P(\tfrac{\bar{X}-80}{.16} \leq 1.645) = P(\tfrac{\bar{X}-80.5}{.16} \leq 1.645 - \tfrac{.5}{.16}) = .0694.$$

Compared to $n = 100$, $.8461 > .6536$, $.0694 > 2 \times 10^{-6}$.

Power of Tests

Protection against the type-I error is built into the tests. To guard against the type-II error, one needs **power**, $P(\text{Reject } H_0)$.

The power of tests for $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$ depends on n and $d = \frac{\mu - \mu_0}{\sigma}$.



- The power is a property of the decision rule.
- One gains more power with larger n.
- To achieve desired power at a specific d, one needs a large enough n. This is similar to sample size planning for CI.
- Power for one-sided tests are different from two-sided.

t-Tests For Population Mean

Consider a sample of size 9 with $\bar{x} = 1.9$ and $s = .51$.

To test for $H_a : \mu \neq 1.5$, one has

$$t = \frac{1.9-1.5}{.51/\sqrt{9}} = 2.353$$

Since $|t| > 2.306 = t_{.025,8}$, one rejects $H_0 : \mu = 1.5$ at the 5%-level. The p-value is

$$p = P(|T_8| > 2.353) = .0465.$$

To test for $H_a : \mu > 1.5$, one rejects $H_0 : \mu \leq 1.5$ at the 5%-level as $t > 1.860 = t_{.05,8}$. The p-value is $p = P(T_8 > 2.353) = .0232$.

For $H_a : \mu < 1.5$, the p-value is $p = P(T_8 < 2.353) = .9768$.

In practice, σ is typically unknown and is to be estimated by s, and one may use the t-tests for hypotheses involving a population mean μ:

1. Calculate $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$.
2. Compare t with critical values from the t-table or calculate p-value using t-distributions.

• With small samples, one needs to assume normal population.

• With large samples, CLT grants normality of $\bar{X}$, s estimates σ reliably, and t_α approaches z_α.

Planning Sample Size For z/t-Tests

Consider $H_0 : \mu \leq 80$ vs. $H_a : \mu > 80$ with $\sigma = .8$. One needs a 5%-level test with power $1 - \beta \geq .9$ at $\mu = 80.2$.

The z-test rejects H_0 when $(\bar{X} - 80)/(.8/\sqrt{n}) > 1.645$.

The power at $\mu = 80.2$ is:

$$1 - \beta(80.2) = P(\tfrac{\bar{X}-80}{.8/\sqrt{n}} > 1.645) = P(\tfrac{\bar{X}-80.2}{.8/\sqrt{n}} > 1.645 - \tfrac{.2}{.8/\sqrt{n}}) \geq .9.$$

So one needs $1.645 - .2/(.8/\sqrt{n}) \leq -1.282$, or

$$n \geq \left(\frac{1.645 + 1.282}{.2/.8}\right)^2 = 137.07.$$

- The sample size depends on α, β, as well as $d = |\mu - \mu_0|/\sigma$.
- Check the box on page 314 for formulas.

Relation Between Test and CI

If the two-sided test for $H_0 : \mu = \mu_0$ accepts H_0 at the α-level, then the $(1-\alpha)100\%$ CI for μ contains μ_0. The converse is also true.

Potency of antibiotic: $n = 100$, $\bar{x} = 80.2$ and $s = .8$.

It was calculated that

$$Z = \frac{80.2 - 80}{.8/\sqrt{100}} = 2.5.$$

Since $|Z| \leq 2.576 = z_{.005}$, $H_0 : \mu = 80$ is accepted at the 1%-level.

The 99% CI for μ is $80.2 \pm 2.576(.08)$, or $(79.994, 80.406)$, which contains 80.

The test's acceptance region is

$$\frac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}} \leq z_{\alpha/2},$$

which is equivalent to

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Ditto for t-test and t-interval.

- The CI simply provides a range of values for μ that are compatible with the data.

Comparing Two Means

Hematocrit level measures the red cell concentration in blood. Large samples of hematocrit values were obtained for males and females of 17-year-old.

	M	F
n	489	469
$\bar{x}$	45.8	40.6
s	2.8	2.9

Is there a gender difference?

The SDs appear to be similar, but the means look different. Need inferences for $\mu_M - \mu_F$.

By assumption or by CLT,

$$\bar{X}_1 \sim N(\mu_1, \sigma_1^2/n_1),$$
$$\bar{X}_2 \sim N(\mu_2, \sigma_2^2/n_2)$$

independent, so

$(\bar{X}_1 - \bar{X}_2) \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$,

or, equivalently,

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

A $(1-\alpha)100\%$ CI for $(\mu_1 - \mu_2)$ is given by

$$\boxed{(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.}$$

CI with Pooled Variance

For the hematocrit data

$$\bar{x}_1 - \bar{x}_2 = 45.8 - 40.6 = 5.2,$$

$$\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} = \frac{2.8^2}{489} + \frac{2.9^2}{469} = .184^2.$$

So a 95% CI for $(\mu_1 - \mu_2)$ is given by $5.2 \pm 1.96(.184),\ \text{or}\ (4.839, 5.561)$. Now as

$$s_p^2 = \frac{488(2.8)^2 + 468(2.9)^2}{488 + 468} = 2.85^2,$$

$$\frac{1}{n_1} + \frac{1}{n_2} = \frac{1}{489} + \frac{1}{469} = .0646^2,$$

$2.85(.0646) = .184$, so the “pooled” version yields the same result.

With normality and $\sigma_1^2 = \sigma_2^2 = \sigma^2$,

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0, 1),$$

and σ^2 can be estimated by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)},$$

the **pooled variance**. Since

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{1/n_1 + 1/n_2}} \sim t_\nu,$$

where $\nu = n_1 + n_2 - 2$, so a CI for $(\mu_1 - \mu_2)$ is given by

$$\boxed{(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,\nu} s_p \sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}}.}$$

Tests Concerning Two Means

Amine serotonin levels were measured from heart disease patients and a control group.

	D	C
n	8	12
$\bar{x}$	3840	5310
s	850	640

$$s_p^2 = \frac{7(850)^2 + 11(640)^2}{7+11} = 729^2,$$

$$t = \frac{3840 - 5310}{729\sqrt{\frac{1}{8} + \frac{1}{12}}} = -4.42.$$

As $|t| = 4.42 > 2.878 = t_{.005,18}$, reject H_0 at the 1%-level. The p-value is $P(|T_{18}| > 4.42) = .00033$.

To test the hypotheses

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_a: \mu_1 - \mu_2 \neq 0,$$

calculate

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

or

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

and reject H_0 when $|Z| > z_{\alpha/2}$ or $|T| > t_{\alpha/2,\nu}$, where $\nu = n_1 + n_2 - 2$.

- Use table to bracket p-value for t.

CIs and Tests With Unequal Variances

For the serotonin data

$$\bar{x}_1 - \bar{x}_2 = 3840 - 5310 = -1470,$$

$$\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} = \frac{850^2}{8} + \frac{640^2}{12} = 352.8^2,$$

with the df given by

$$\frac{(\frac{850^2}{8})^2(\frac{1}{7}) + (\frac{640^2}{12})^2(\frac{1}{11})}{352.8^4} = \frac{1}{12.18}.$$

Thus a 95% CI for $\mu_1 - \mu_2$ is given by $-1470 \pm 2.175(352.8) = (-2237, -703)$, where $t_{.025,12.18} = 2.175$.

- `qt(...)` does take fractional df.

For small sample normal data with unequal variances, one has

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \overset{\text{approx.}}{\sim} t_\nu,$$

where the df ν is given by

$$\frac{1}{\nu} = \frac{\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{(n_2-1)}}{(s_1^2/n_1 + s_2^2/n_2)^2}.$$

Note that when $n_1 = n_2 = n$ and $s_1^2 = s_2^2$, $\nu = 2(n-1)$.

CI's and tests concerning $\mu_1 - \mu_2$ can then be constructed accordingly.

Paired Data

Blocks of land were divided into two plots and the plots were planted with two varieties of wheat. The yields follow.

	Variety		
Block	1	2	d
1	32.1	34.5	-2.4
2	30.6	32.6	-2.0
3	33.7	34.6	-0.9
4	29.7	31.0	-1.3
mean	31.53	33.18	-1.65
SD	1.76	1.72	.676

Compare the mean yields of the two varieties.

- Typical pairing: blocking designs, before-after studies, left-right organs, repeated measurements, etc.
- Pairing effectively reduces “background” noise.
- If pairing is ignored, features may be swamped by background noise.
- When done on irrelevant factors, pairing may yield loss of power.

Inference for Paired Data

Working on d, a 95% CI for wheat yield difference is given by $-1.65 \pm 3.182(.676)/\sqrt{4}$, or $(-2.73, -0.57)$, where $t_{.025,3} = 3.182$. Variety 2 appears to yield significantly more.

Ignoring pairing,

$$s_p^2 = \frac{3(1.76)^2 + 3(1.72)^2}{3+3} = 1.74^2,$$

thus $s_p\sqrt{\frac{1}{4}+\frac{1}{4}} = 1.23$, so a 95% CI is $-1.65 \pm 2.447(1.23)$, or $(-4.66, 1.36)$, where $t_{.025,6} = 2.447$. The result is inconclusive.

- Note that $\hat{\sigma}_{\bar{d}} = .338$, $\hat{\sigma}_{\bar{x}_1-\bar{x}_2} = 1.23$, almost a 4-fold difference. $\hat{\sigma}_{\bar{x}_1-\bar{x}_2}$ includes block to block variability.
- Pairing results in a loss of df: compare 3.182 with 2.447. When pairing is ineffective, say $\hat{\sigma}_{\bar{d}} \approx \hat{\sigma}_{\bar{x}_1-\bar{x}_2}$, this would yield loss of power. This is negligible for larger sample sizes, though.

CIs and t-Tests in R

In R, one may use `t.test(...)` to calculate one or two sample t-tests and the associated CI's.

```
x <- rnorm(30); y <- rnorm(20,mean=2)
t.test(x); t.test(y,mu=2,alt="less"); t.test(x,alt='gre')
t.test(x,y) ## unequal var with approximate df
t.test(x,y,var.equal=TRUE) ## with pooled var est
t.test(x,y,conf.level=.9) ## 90% CI
t.test(x[1:20],y,paired=T); t.test(x[1:20]-y)
x <- matrix(rnorm(100000),10,10000) ## coverage simulation
y <- matrix(rnorm(150000,sd=2),15,10000)
fun <- function(x) {t.test(x[1:10],x[11:25])$conf.int}
jk <- apply(rbind(x,y),2,fun); sum(jk[1,]*jk[2,]>0)
```