TECHNISCHE UNIVERSITÄT WIEN

Introduction to Statistics
Confidence Intervals
Tests of Hypotheses

LV Nr. 105.692

Summer Semester 2021

CAN Estimators and Standard Errors

- Suppose $\hat{\theta}_n$ is a CAN estimator; i.e.,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} Y$$

  where

$$Y \sim \mathcal{N}(0, \tau^2)$$

- Suppose we also have a consistent estimator of the asymptotic variance

$$\hat{\tau}_n^2 \xrightarrow{P} \tau^2$$

CAN Estimators and Standard Errors

- We combined these using Slutsky's theorem to obtain

$$\frac{\hat{\theta}_n - \theta}{\hat{\tau}_n / \sqrt{n}} \xrightarrow{d} \frac{Y}{\tau}$$

- A linear function of a normal is normal, Y/τ is normal with parameters

$$\frac{\mathbb{E}(Y)}{\tau} = 0$$

$$\frac{\mathbb{V}\mathrm{ar}(Y)}{\tau^2} = 1$$

- Thus,

$$\frac{\hat{\theta}_n - \theta}{\hat{\tau}_n / \sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

CAN Estimators and Standard Errors

- Now only the parameter of interest is unknown, and we can make error bars based on

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| \leqslant \frac{2\hat{\tau}_n}{\sqrt{n}}\right) \approx 0.9545$$

  or,

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| \leqslant \frac{1.96 \times \hat{\tau}_n}{\sqrt{n}}\right) \approx 0.95$$

- We call the intervals with endpoints

$$\hat{\theta}_n \pm 2\frac{\hat{\tau}_n}{\sqrt{n}}$$

  or,

$$\hat{\theta}_n \pm 1.96\frac{\hat{\tau}_n}{\sqrt{n}}$$

  approximate (large n, asymptotic) 95% confidence intervals for θ.

Confidence intervals

- We can allow any probability: Let z_α denote the $1-\alpha$ quantile of $\mathcal{N}(0,1)$. Then,
$$\mathbb{P}\left(\left|\frac{\hat{\theta}_n - \theta}{\frac{\hat{\tau}_n}{\sqrt{n}}}\right| \leqslant z_{\alpha/2}\right) \approx 1-\alpha$$
- The interval with endpoints
$$\hat{\theta}_n \pm z_{\alpha/2}\frac{\hat{\tau}_n}{\sqrt{n}}$$
is called the asymptotic confidence interval for θ with coverage probability $1-\alpha$, or with confidence level $100(1-\alpha)\%$.

MLE CAN construction

Let $\{f(x|\theta) : \theta \in \Theta\}$ be any parametric model satisfying the regularity conditions of Theorem 1 of Fisher information and the Crameér-Rao lower bound section, where θ is a single parameter. To obtain a confidence interval for θ, consider the MLE $\hat{\theta}$ which satisfies

$$\frac{\hat{\theta} - \theta}{\frac{1}{\sqrt{n}}} \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{I(\theta)}\Big)$$

as $n \to \infty$, where $I(\theta)$ is the Fisher information. As θ is unknown, we can estimate $I(\theta)$ by the plugin estimator $I(\hat{\theta})$. If $I(\theta)$ is continuous in θ and $\hat{\theta}$ is consistent estimator of θ, then

$$I(\hat{\theta}) \xrightarrow{p} I(\theta).$$

Wald CI for MLE

Thus,

$$\frac{\hat{\theta}-\theta}{\sqrt{\frac{1}{nI(\hat{\theta})}}} = \sqrt{\frac{I(\hat{\theta})}{I(\theta)}} \cdot \frac{\hat{\theta}-\theta}{\sqrt{\frac{1}{nI(\theta)}}} \xrightarrow{d} \mathcal{N}(0,1)$$

so

$$P_\theta\Big(-z_{\alpha/2} \leqslant \frac{\hat{\theta}-\theta}{\sqrt{\frac{1}{nI(\hat{\theta})}}} \leqslant z_{\alpha/2}\Big) \to 1-\alpha.$$

After rearranging we obtain the asymptotic $100(1-\alpha)\%$ confidence interval

$$\Big[\hat{\theta} - z_{\alpha/2} \cdot \frac{1}{\sqrt{nI(\hat{\theta})}},\ \hat{\theta} + z_{\alpha/2} \cdot \frac{1}{\sqrt{nI(\hat{\theta})}}\Big], \tag{1}$$

also called the *Wald interval* for θ.

Example: Poisson

Let $X_1, \dots, X_n$ be i.i.d. with $Poi(\lambda)$. To construct an asymptotic confidence interval for λ, we start with the estimator $\hat{\lambda} = \bar{X}$.
We showed that $\hat{\lambda} = \bar{X}$ is the MLE of λ. By the Central Limit Theorem,

$$\frac{\hat{\lambda} - \lambda}{\frac{1}{\sqrt{n}}} \xrightarrow{d} \mathcal{N}(0, \lambda)$$

As the variance λ of this limiting normal distribution is unknown, we can estimate it by $\hat{\lambda}$.
By the LLN

$$\hat{\lambda} \overset{p}{\to} \lambda \quad \text{as } n \to \infty$$

i.e. $\hat{\lambda}$ is consistent estimator of λ. Then by Slutsky's theorem we obtain

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\frac{\hat{\lambda}}{n}}} = \frac{\lambda}{\hat{\lambda}} \cdot \frac{\hat{\lambda} - \lambda}{\sqrt{\frac{\lambda}{n}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Example: Poisson

Hence,

$$P_\lambda\Big(-z_{\alpha/2} \leqslant \frac{\hat\lambda - \lambda}{\sqrt{\frac{\hat\lambda}{n}}} \leqslant z_{\alpha/2}\Big) \to 1-\alpha,$$

where $z_{\alpha/2}$ is the $(1-\frac{\alpha}{2})$-quantile of $\mathcal{N}(0,1)$. After rearranging these inequalities we obtain the asymptotic $100(1-\alpha)\%$ confidence interval for λ

$$\Big[\hat\lambda - z_{\alpha/2}\cdot\sqrt{\frac{\hat\lambda}{n}},\ \hat\lambda - z_{\alpha/2}\cdot\sqrt{\frac{\hat\lambda}{n}}\Big].$$

[HW] Use `R` to generate a random sample $X_1, \ldots, X_n$ from $Pois(1)$ distribution (for $n = 30$ and $n = 100$), Compute the 90% condidence interval for λ, check if it contains the true value of $\lambda = 1$, and repeat this 10000 times. What is the fraction of simulations for which the confidence interval covered λ?

Asymptotic confidence interval construction

1. We approximate the true distribution of $\sqrt{n}(\hat{\theta} - \theta)$ by a normal distribution.
2. We approximate the true variance of this normal distribution by a plugin estimator that is consistent.
3. If we have any two statistics T and V of the unknown parameter θ such that
$$\frac{T - \theta}{V} \xrightarrow{d} \mathcal{N}(0, 1)$$
as $n \to \infty$ then we can form the approximate confidence interval for θ
$$\left[T - z_{\frac{\alpha}{2}}\, V\,,\, T + z_{\frac{\alpha}{2}}\, V\right]$$
which is essentially a Wald type interval.
4. When we are interested in $g(\theta)$ and g is a nonlinear function, we approximate the value $g(\hat{\theta})$ by the Taylor expansion $g(\theta) + (\hat{\theta} - \theta)\, g(\theta)$ (the Delta method).

[HW] Let $X_1, \ldots, X_n$ be i.i.d. Bernulli (p), with unknown $p \in (0, 1)$.

1. Find the MLE $\hat{p}$ of p.
2. Show that $\hat{p}$ is a consistent estimator of p.
3. Use the Delta method to show that the log-odds $g(p) = \log \frac{p}{1-p}$ is the asympthotically normally distributed and

$$\frac{g(\hat{p}) - g(p)}{\frac{1}{\sqrt{n}}} \xrightarrow{\text{d}} \mathcal{N}\Big(0, \frac{1}{p(1-p)}\Big),$$

   where $\hat{p} = \bar{X}$.
4. Construct an asymptotic $100(1-\alpha)\%$ confidence interval for the log-odds $g(p) = \log \frac{p}{1-p}$.
5. Construct an asymptotic $100(1-\alpha)\%$ confidence interval for the odds $e^{g(x)} = \frac{p}{1-p}$.

Confidence intervals

- Based on the random sample $(X_1, \dots, X_n)$, we are interested in constructing a *random interval* that contains the true but unknown parameter value with a specified probability.
- This is the problem of *interval estimation*.
- In general, confidence region is a random subset of the parameter space Θ - random because it is a function of the data - that has a stated probability of covering the true unknown parameter value. If R is the region, then

$$\mathbb{P}_\theta(\theta \in R) = 1 - \alpha, \quad \text{for all } \theta \in \Theta$$

  where Θ is the parameter space and $1 - \alpha$ is the stated coverage probability.
- A confidence interval is the special case where the parameter is one-dimensional and the set R is always an interval.

Exact CIs

- Let $X_1, \ldots, X_n$ be a random sample.
- By *random interval*, we mean an interval whose lower and upper endpoints $L(X_1, \ldots, X_n)$ and $U(X_1, \ldots, X_n)$ are functions of $X_1, \ldots, X_n$.
- The interval is random because a different realization of the data (random sample) leads to a different interval.
- The interval

  $$[L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$$

  is a $100(1-\alpha)\%$ confidence interval for $g(\theta)$ if for all $\theta \in \Theta$

  $$P_\theta\Big(L(X_1, \ldots, X_n) \leqslant g(\theta) \leqslant U(X_1, \ldots, X_n)\Big) = 1 - \alpha, \tag{2}$$

  where P_θ denotes the probability under $X_1, \ldots, X_n$ i.i.d. $f(x|\theta)$.
- A confidence interval for $g(\theta)$ is usually constructed from an estimate of $g(\theta)$ and an estimate of the associated standard error.

Example: Normal

- Let $X_1, \ldots, X_n$ be a random sample from a population with $\mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknnwn.
- To construct a confidence interval for μ, we consider the sample mean.
- We know that
$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}).$$
- The variance of $\bar{X}$ is $\frac{\sigma^2}{n}$ and its standard error is $\frac{\sigma}{\sqrt{n}}$.
- Since σ^2 is also unknown we estimate the variance by $\frac{S^2}{n}$, where
$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{X} - X_i)^2$$
is the sample variance (recall S^2 is unbiased estimator of σ^2).

Example: Normal (ctd)

- Recall that when $X_1, \ldots, X_n$ are i.i.d. random variables from $\mathcal{N}(\mu, \sigma^2)$ the quantity
$$\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} = \frac{\sqrt{n}\,(\bar{X}-\mu)}{S}$$
has a t-distribution with $(n-1)$ degrees of freedom.
- Then for the bounds we use $t_{\alpha/2}(n-1)$, the $\frac{\alpha}{2}$-quantile of the t-distribution with $(n-1)$ degrees of freedom, i.e.
$$P_{\mu,\sigma^2}\left(t_{1-\alpha/2}(n-1) \leqslant \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \leqslant t_{\alpha/2}(n-1)\right) = 1-\alpha$$
- Since the t-distribution is symmetric, $t_{1-\alpha/2}(n-1) = -t_{\alpha/2}(n-1)$. Therefore,
$$P_{\mu,\sigma^2}\left(-t_{\alpha/2}(n-1) \leqslant \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \leqslant t_{\alpha/2}(n-1)\right) = 1-\alpha,$$

Example: Normal (ctd)

which leads to

$$P_{\mu,\sigma^2}\left(\bar{X} - t_{\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + t_{\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}}\right) = 1-\alpha.$$

Hence,

$$\left[\bar{X} - t_{\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}}\right] \tag{3}$$

is a $100(1-\alpha)\%$ confidence interval for μ (when σ^2 is unknown).

The notation

$$\bar{X} \pm t_{\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}}$$

as shorthand for this interval is also used. Quantiles of t-districution can be read from Tables of t-ditribution or in R by using the command `qt`.

As a particular case of the previous example, we consider the following problem. A psychologist is interested in the reaction time in road traffic of ten-year-old students. An average response time of $\mu = 0.8$ seconds with empirical variance $s^2 = 0.04$ was measured for n = 51 students. Use `R` to determine the 95%-confidence interval for the mean response time.

The required confidence interval is of the form (3). We use the command `qt` to write

```
x.mean=0.8
s=sqrt(0.04)
n=51
alpha=0.05
dist = qt(1-alpha/2,n-1)*s/sqrt(n)
lower = x.mean - dist
upper = x.mean + dist
```

Thus, the confidence interval for the mean reaction time is approximately $(0.743, 0.856)$.

Exact CI for the normal variance

Assuming the data are IID normal, we can also construct confidence intervals for the unknown true population variance σ^2 based on

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

Let $\chi^2_\alpha(n-1)$ denote the $1-\alpha$ quantile of the $\chi^2(n-1)$ distrn. Then,

$$\mathbb{P}\left(\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S_n^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1)\right) = 1-\alpha$$

An exact $100(1-\alpha)\%$ CI for σ^2 is

$$\frac{(n-1)S_n^2}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S_n^2}{\chi^2_{1-\alpha/2}(n-1)}$$

CI based on the χ^2 distribution

- A criticism of intervals based on the chi-square distribution is that there is no particular reason to use equal tailed intervals.
- If the reference distribution is symmetric, then equal-tailed CIs make sense. But not otherwise.

Discrete Data

For discrete data, exact confidence intervals are impossible.

- Consider the binomial distribution with sample size n.
- There are only $n+1$ possible data values: $0, 1, \ldots, n$. Hence there are only $n+1$ possible intervals one can make that are functions of the data.
- Let R_x denote the confidence interval for data x. The coverage probability is

$$\mathbb{P}_p(p \in R_x) = \mathbb{E}_p(I_{R_x}(p)) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} I_{R_x}(p)$$

- This piecewise polynomial function of p cannot be constant, i.e., equal to $1-\alpha$ for all p such that $0 < p < 1$.
- In this case, we use *approximate coverage* CIs.

Hypothesis Tests

- Often confidence intervals do not do exactly what is wanted.
- Two related reasons for this.

1. Sometimes the size of an effect is not interesting, only the existence of the effect.
    - In the EU, a drug may be approved for marketing if it is safe and effective. The size of the treatment effect is irrelevant.
2. Sometimes the size of the effect depends on the details of the particular experiment and would not generalize to other situations. A phenomenon is hypothesized. An experiment is designed to study it. If the experiment shows that the phenomenon exists, then that generalizes to other situations, but the effect size seen does not generalize.

Hypothesis Tests (ctd)

- To relate to statistics, we need to turn these statements about existence of effects and phenomena into statements about a statistical model.
- A *statement about a statistical model* is called a statistical hypothesis, and formally statistical tests are called tests of statistical hypotheses
  - A statistical hypothesis asserts that the true unknown distribution lies in a sub-model of the statistical model under consideration.
  - If the model under consideration has parameter space Θ, then a statistical hypothesis asserts that the true unknown parameter value lies in a subset Θ_i of Θ.

Hypothesis Tests (ctd)

A hypothesis test considers two hypotheses: the null hypothesis H_0, and the alternative hypothesis H_1.

In terms of the parameter space, they are

$$H_0 : \Theta_0 \quad \text{vs.} \quad H_1 : \Theta_1$$

Example

- We have a random sample, $X_1, \ldots, X_n$ that denotes the observations of the subjects in a treatment in a medical experiment.
- Suppose we are willing to assume that the sample sample is iid normal with known variance.
- The question of scientific interest is whether the treatment has an effect, defined to be difference from not applying any treatment.
- We turn this into a question about statistical models: whether $\mathbb{E}(X) = \mu_X$ satisfies

$$H_0 : \mu_X \leqslant \mu_0 \quad \text{vs.} \quad H_1 : \mu_X > \mu_0$$

  where μ_0 stands for the average value of X in the population who is not submitted to the treatment.

Example

- Even though not obvious, this hypothesis test is equivalent to the simpler

$$H_0 : \mu_X = \mu_0 \quad \text{vs.} \quad H_1 : \mu_X > \mu_0$$

- How to decide in favor of H_0 or H_1?
- Let's recall that since X_i are iid $\mathcal{N}(\mu_X, \sigma^2)$,

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

or,

$$Z = \frac{\bar{X}_n - \mu_X}{\sigma/n} \sim \mathcal{N}(0, 1)$$

Example (ctd)

$$Z = \frac{\bar{X}_n - \mu_X}{\sigma/n} \sim \mathcal{N}(0, 1)$$

- Under H_0 the test statistic has sampling distribution centered at μ_0. Under H_1 the test statistic has sampling distribution centered at some number larger than μ_0. Thus larger than μ_0 values of Z are evidence in favor of H_1.
- Under H_0 we know the distribution of Z to be $\mathcal{N}(0, 1)$. Under H_1 we do not, because it depends on the unknown μ_X. Therefore, we base the probability calculation on H_0.

Example (ctd)

$$H_0 : \mu_X = \mu_0 \quad \text{vs.} \quad H_1 : \mu_X > \mu_0$$

- The *p*-value of the test is

$$p - val = \mathbb{P}(Z \geqslant z)$$

where $Z \sim \mathcal{N}(0, 1)$ ad $z = \frac{\bar{x} - \mu_0}{\sigma/n}$: the observed test statistic value. The probability is calculated under H_0.

- Under H_1, the observed value z is likely to be large and positive, hence far out in the tail of the distribution of Z under H_0. Thus, the p-value should be small when H_1 is true, but should be large (near 1/2) when H_0 is true.

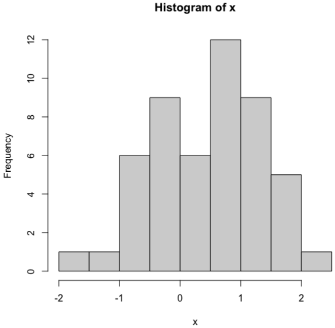
Example (ctd)

$$H_0 : \mu_X = \mu_0 \quad \text{vs.} \quad H_1 : \mu_X > \mu_0$$

- Small p-values are evidence in favor of H_1 and large p-values are evidence in favor of H_0.
- If we compute a small p-value we say that either H_1 is true or an unusual event has occurred since $Z \geqslant z$ occurs with small probability (the p-value).
- What is considered small is a topic of debate.

Example in R

1. We observe 50 data points from a population with variance 1. We draw a histogram:



2. One can claim the mean is zero; that is, $\mu_0 = 0$, and we want to test it.

3. The hypotheses are

$$H_0 : \mu_X = 0 \quad \text{vs.} \quad H_1 : \mu_X > 0$$

Example in R

1. The test statistic is

$$Z = \frac{\bar{X}_n - 0}{1/\sqrt{n}} = \frac{\bar{X}_{50}}{1/\sqrt{50}}$$

2. in R

```
> bar.x=mean(x)
> z=bar.x/(1/sqrt(length(x)))
> pnorm(z,lower.tail=FALSE)
[1] 0.0003458069
```

3. The *p*-value is 0.00034, pretty small. The evidence is against the null and in support of the alternative. We decide that the mean is larger than 0.

4. This is correct: the data were generated from $\mathcal{N}(0.5, 1)$.

Example in R

1. Now, we generate 50 values from $\mathcal{N}(0, 1)$ and recompute the test statistic and the p-value:

```
x=rnorm(50,0,1)
> hist(x)
> bar.x=mean(x)
> z=bar.x/(1/sqrt(length(x)))
> pnorm(z,lower.tail=FALSE)
[1] 0.001111735
```

2. The p-value is 0.001, pretty small again. The evidence is against the null and in support of the alternative. We decide that the mean is larger than 0.

Example in R

1. This is wrong: the data were generated from $\mathcal{N}(0, 1)$ and the true mean is 0!
2. Repeating the same for a new random sample from $\mathcal{N}(0, 1)$ I get a p-value of 0.415: pretty large indicating no evidence against the null (which is right!).
3. Morale: p-values are only probabilities associated with the specific sample!

Theory of Hypothesis Testing

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1$$

1. The parameter spaces Θ_0 and Θ_1 are any two disjoint subsets of the parameter space.
2. When Θ_0 is a singleton set (contains exactly one point), the null hypothesis is said to be *simple*.
    - For example, $\mu = 0$
3. The alternative hypothesis only motivates the choice of test statistic: e.g.,

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma / \sqrt{n}}$$

One-Tailed and Two-Tailed Tests

In tests where the hypotheses

- involve a single parameter θ,
- the distribution of the test statistic, say, T is symmetric about zero under H_0, and
- the distribution of the test statistic shifts in the same direction as θ does under H_1,

we distinguish three kinds of tests:

Upper-Tailed Tests

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0$$

with

$$p - value = \mathbb{P}_{\theta_0}(T \geqslant t)$$

One-Tailed and Two-Tailed Tests

Lower-Tailed Tests

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta < \theta_0$$

with

$$p - value = \mathbb{P}_{\theta_0}(T \leqslant t)$$

Two-Tailed Tests

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$$

with

$$\begin{aligned} p - value &= \mathbb{P}_{\theta_0}(|T| \geqslant |t|) \\ &= 2\,\mathbb{P}_{\theta_0}(T \leqslant -|t|) \\ &= 2\,\mathbb{P}_{\theta_0}(T \geqslant |t|) \end{aligned}$$

Principle of Hypothesis Testing

- Do only one test per data set.
- Not just report only one test, **do only one test**.
- Moreover the test to be done is chosen before the data are observed.
- This principle is often violated, but unrestricted multiple testing without correction makes HT a meaningless exercise. More on this later.

$H_0 : \theta = 0$ is equivalent to $H_0 : \theta \leqslant \theta_0$

Suppose $X_1, \ldots, X_n$ iid $f_\theta(x)$ and $\hat{\theta}_n$ is a CAN for θ; i.e., $\sqrt{n}(\hat{\theta}_n - \theta) \approx \mathcal{N}(0, \tau^2)$.

- Consider a one-tailed test based on the exact or asymptotic pivotal quantity

$$\frac{\hat{\theta}_n - \theta}{\hat{\tau}_n / \sqrt{n}}$$

  to test the hypotheses

$$H_0 : \theta \leqslant \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0$$

- Claim: the test having the test statistic

$$T = \frac{\hat{\theta}_n - \theta_0}{\hat{\tau}_n / \sqrt{n}}$$

  and

$$p - value = \mathbb{P}_{\theta_0}(T \geqslant t)$$

  is valid.

$H_0 : \theta = 0$ is equivalent to $H_0 : \theta \leqslant \theta_0$

We must show that

$$\mathbb{P}_\theta(T \geqslant t) \leqslant \mathbb{P}_{\theta_0}(T \geqslant t), \quad \theta < \theta_0$$

If the true unknown parameter value is θ, then

$$\frac{\hat{\theta}_n - \theta}{\hat{\tau}_n/\sqrt{n}} = \frac{\hat{\theta}_n - \theta_0}{\hat{\tau}_n/\sqrt{n}} + \frac{\theta_0 - \theta}{\hat{\tau}_n/\sqrt{n}}$$

has the same distribution as T does when the true unknown parameter value is θ_0. Hence,

$$\mathbb{P}_\theta\left(T + \frac{\theta_0 - \theta}{\hat{\tau}_n/\sqrt{n}} \geqslant s\right) = \mathbb{P}_\theta\left(T \geqslant s\right)$$

or,

$$\mathbb{P}_\theta(T \geqslant t) = \mathbb{P}_{\theta_0}\left(T \geqslant t + \frac{\theta_0 - \theta}{\hat{\tau}_n/\sqrt{n}}\right)$$

$H_0 : \theta = 0$ is equivalent to $H_0 : \theta \leqslant \theta_0$

We assume $\hat{\tau}_n > 0$, since it is an estimate of standard deviation.

If H_0 is true, i.e., $\theta \leqslant \theta_0$, then

$$\frac{\theta_0 - \theta}{\hat{\tau}_n / \sqrt{n}} \geqslant 0$$

and

$$\mathbb{P}_\theta(T \geqslant t) = \mathbb{P}_{\theta_0}\left(T \geqslant t + \frac{\theta_0 - \theta}{\hat{\tau}_n / \sqrt{n}}\right) \leqslant \mathbb{P}_{\theta_0}\left(T \geqslant t\right)$$

$H_0 : \theta = 0$ is equivalent to $H_0 : \theta \leqslant \theta_0$

In conclusion: the test with *p*-value

$$\mathbb{P}_{\theta_0}(T \geqslant t)$$

is valid for either

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0$$

or

$$H_0 : \theta \leqslant \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0$$

whether the null hypothesis is an equality or inequality is irrelevant. And similarly for the other one-tailed test.

Decision theory and α levels

- The theory of statistical decisions is a large subject. When applied to hypothesis tests, it gives a different view.
- The point of a hypothesis test is to decide in favor of H_0 or H_1. The result is one of two decisions:
  1. accept H_0 or reject H_1 (both mean the same)
  2. reject H_0 or accept H_1 (both mean the same)
- In the decision-theoretic mode, the result of a test is just reported in these terms. No p-value is reported, hence no indication of the strength of evidence.

Decision theory and α levels

If no p-value is reported, how is the test done?

A level of significance α is chosen

- If p-value $< \alpha$, then the test decides reject H_0.
- If p-value $\geqslant \alpha$, then the test decides accept H_0.

The decision theoretic view provides *less information*. Instead of giving the actual p-value, it is only reported whether the p-value is above or below α.

Decision theory and α levels

- Ideally, the significance level α should be chosen carefully and reflect the costs and probabilities of false positive and false negative decisions.
- Since the decision-theoretic mode provides less information and isn't usually done properly
  - In practice, $\alpha = 0.05$ is usually thoughtlessly chosen

  many recent textbooks say it should not be used: always report the p-value, never report only a decision.

Power

- Everything we have said about hypothesis tests so far is only about validity. Is the test defensible?
- A different issue is power, which is, roughly, how probable is it that the test will do what is wanted:

$$\text{power of a test} = \mathbb{P}(\text{it will accept } H_1 \mid H_1 \text{ is true})$$

- Since H_1 is always a composite hypothesis, the power is always a function of the true unknown parameter value θ. It also depends on the sample size.
- The power of a hypothesis is useful in planning an experiment or getting funding for an experiment. What is the point of funding an experiment that probably won't detect anything anyway because the planned sample size is too small?

Power

- Power calculations are simplest when the reference distribution (the distribution of the test statistic under H_0, either exact or approximate) is normal.
- **Example:** suppose the asymptotically pivotal quantity

$$\frac{\hat{\theta}_n - \theta}{\hat{\tau}_n/\sqrt{n}} \approx \mathcal{N}(0, 1)$$

  and the test statistic is

$$T = \frac{\hat{\theta}_n - \theta_0}{\hat{\tau}_n/\sqrt{n}}$$

Power

- Since we showed that

$$\mathbb{P}_\theta(T \geqslant t) = \mathbb{P}_{\theta_0}\left(T \geqslant t + \frac{\theta_0 - \theta}{\hat{\tau}_n/\sqrt{n}}\right)$$

  but we do not know the sampling distribution of $\hat{\tau}_n$ (we only know that it is a consistent estimator of the nuisance parameter τ), we write

$$\mathbb{P}_\theta(T \geqslant t) \approx \mathbb{P}_{\theta_0}\left(T \geqslant t + \frac{\theta_0 - \theta}{\tau/\sqrt{n}}\right)$$

- Using the asymptotic normality of the pivotal, we re-write as

$$\mathbb{P}_\theta(T \geqslant z_\alpha) \approx \mathbb{P}_{\theta_0}\left(T \geqslant z_\alpha - \frac{\theta - \theta_0}{\tau/\sqrt{n}}\right)$$

Power

- This, considered as a function of θ, is the power function of the upper-tailed test.
- It depends on the hypothetical value of the nuisance parameter τ and the sample size.
- Alternatively, we can consider it a function of the standardized treatment effect

$$\frac{\theta - \theta_0}{\tau}$$

  and the sample size n

Power in R

```r
## 1-tailed mean z-test

alpha <- 0.05
n <- c(200, 100, 50)
z <- qnorm(1 - alpha)

curve(pnorm(z - x * sqrt(n[1]), lower.tail = FALSE),
      from = 0, to = 1 / 2, ylab = "power",
      xlab = expression((theta - theta[0]) / tau), lwd = 2)
for (i in 2:length(n))
  curve(pnorm(z - x * sqrt(n[i]), lower.tail = FALSE),
        add = TRUE, lty = i, lwd = 2)
```

One-tail z-test power plot

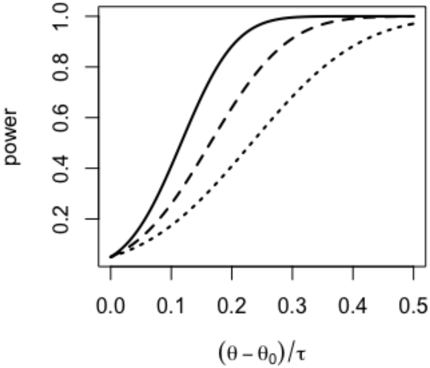


Figure: Power curves for upper-tailed z test and $\alpha = 0.05$: solid line is $n = 200$, dashed line is $n = 100$, dotted line is $n = 50$.

- Power increases from α to 1 as standardized effect $(\theta - \theta_0)/\tau$ increases from zero to infinity.
- Power increases from α to 1 as sample size increases from 0 to infinity.
- The first is not under control of the experimenters. The effect size is what it is, and although hypothetical in the power calculation, should be realistic. The second is under control of the experimenters. The sample size should be chosen so that the power will be reasonably large.

Power for 2-tailed Tests

$$\begin{aligned}\mathbb{P}_{\theta}(|T| \geqslant z_{\alpha/2}) \approx \mathbb{P}_{\theta_0}&\left(T \geqslant z_{\alpha/2} - \frac{\theta - \theta_0}{\tau/\sqrt{n}}\right) \\ &+ \mathbb{P}_{\theta_0}\left(T \leqslant -z_{\alpha/2} - \frac{\theta - \theta_0}{\tau/\sqrt{n}}\right)\end{aligned}$$

Power in R

```r
## 2 tailed mean z-test
alpha <- 0.05
n <- c(200, 100, 50)
z <- qnorm(1 - alpha / 2)

curve(pnorm(z - x * sqrt(n[1]), lower.tail = FALSE) +
        pnorm(- z - x * sqrt(n[1])), from = - 1 / 2,
      to = 1 / 2, ylab = "power",
      xlab = expression((theta - theta[0]) / tau), lwd = 2)
for (i in 2:length(n))
  curve(pnorm(z - x * sqrt(n[i]), lower.tail = FALSE) +
          pnorm(- z - x * sqrt(n[i])), add = TRUE, lty = i,
        lwd = 2)
```

2-tail z-test power plot

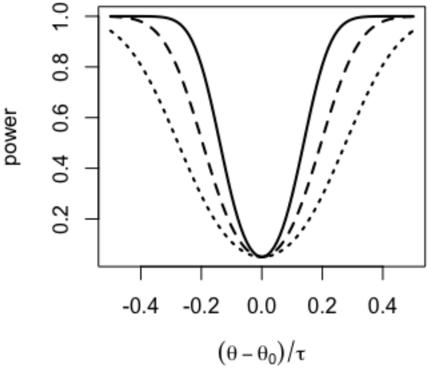


Figure: Power curves for 2-tailed z test and $\alpha = 0.05$: solid line is $n = 200$, dashed line is $n = 100$, dotted line is $n = 50$.

Power (ctd)

- Power calculations are more complicated when the reference distribution is not normal.
- When the reference distribution is t, F, or χ^2, then the distribution under the alternative hypothesis is so-called noncentral t, F, or χ^2, respectively.
- `R` can calculate probabilities for these distributions with `pt, pf`, and `pchisq` by setting the noncentrality parameter argument `ncp`.

Non-central t

- If $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2(\nu)$ and Z and Y are independent, then

$$T = \frac{Z}{\sqrt{Y/\nu}} \sim t(\nu)$$

- We define

$$T = \frac{Z+\delta}{\sqrt{Y/\nu}} \sim t(\nu)$$

  to have the noncentral t distribution with degrees of freedom ν and noncentrality parameter δ, denoted $t(\nu, \delta)$.

Non-central t

- Let's derive the power for a t test:
- We want to test
$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \geqslant \mu_0$$
- The pivotal quantity is
$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S_n^2}{\sigma^2}/(n-1)}} = \frac{Z}{\sqrt{Y/\nu}} \sim t(\nu)$$

if X_i iid $\mathcal{N}(\mu, \sigma^2)$ with σ^2 unknown.

Non-central t

- The test statistic is
$$T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$
- When the true unknown parameter value is μ, the numerator of the T
$$Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$$
is normal but not standard normal, since
$$\mathbb{E}(Z) = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$
$$\mathbb{V}\mathrm{ar}(Z) = 1$$

Non-central t

- Therefore,

$$Z = Z^\star + \frac{\mu - \mu_0}{\sigma / \sqrt{n}}$$

  where $Z^\star \sim \mathcal{N}(0, 1)$.

- In consequence,

$$T = \frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}} \sim t(n - 1, \delta)$$

  with

$$\delta = \frac{\mu - \mu_0}{\sigma / \sqrt{n}}$$

Power in R

```r
## One-tail t-test
alpha <- 0.05
nn <- c(20, 10, 5)

n <- nn[1]
z <- qt(1 - alpha, n - 1)
curve(pt(z, df = n - 1, ncp = sqrt(n) * x,
         lower.tail = FALSE), from = 0, to = 1,
      xlab = expression((mu - mu[0]) / sigma),
      ylab = "power", lwd = 2)
for (i in 2:length(nn)) {
  n <- nn[i]
  z <- qt(1 - alpha, df = n - 1)
  curve(pt(z, df = n - 1, ncp = sqrt(n) * x,
           lower.tail = FALSE), add = TRUE, lty = i, lwd = 2)
}
```

One-tail t-test power plot

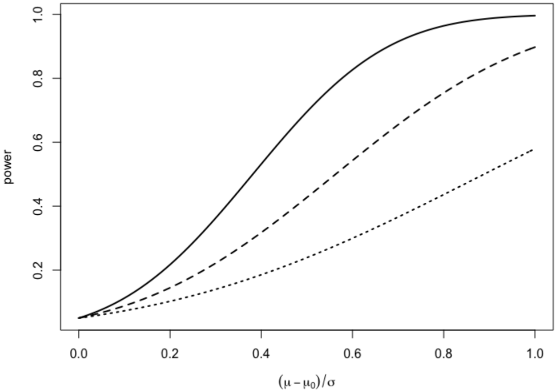


Figure: Power curves for upper-tailed t test and $\alpha = 0.05$: solid line is $n = 200$, dashed line is $n = 100$, dotted line is $n = 50$.

Power in R

```r
## 2-tail t-test
alpha <- 0.05
nn <- c(20, 10, 5)

n <- nn[1]
z <- qt(1 - alpha / 2, n - 1)
curve(pt(z, df = n - 1, ncp = sqrt(n) * x,
         lower.tail = FALSE) +
        pt(- z, df = n - 1, ncp = sqrt(n) * x),
      from = -1, to = 1, ylab = "power",
      xlab = expression((mu - mu[0]) / sigma), lwd = 2)
for (i in 2:length(nn)) {
  n <- nn[i]
  z <- qt(1 - alpha / 2, df = n - 1)
  curve(pt(z, df = n - 1, ncp = sqrt(n) * x,
           lower.tail = FALSE) +
          pt(- z, df = n - 1, ncp = sqrt(n) * x),
        add = TRUE, lty = 2, lwd = 2)
}
```

Two-tail t-test power plot

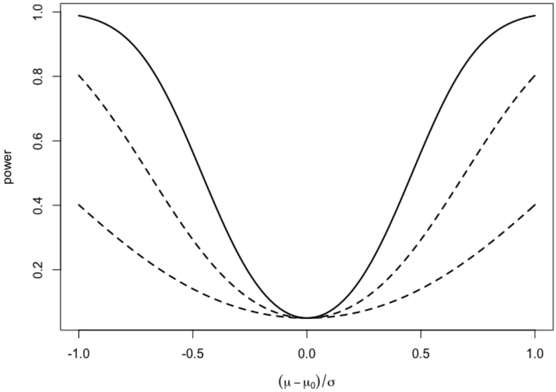


Figure: Power curves for two-tailed t test and $\alpha = 0.05$: solid line is $n = 200$, dashed line is $n = 100$, dotted line is $n = 50$.

- The same behavior as in z tests
- The power increases from α to 1 as either the standardized effect size increases or the sample size increases.
- The power curves look similar whether the normal distribution or the noncentral t distribution is used.