Chapter 2: Statistical inference

Dr.Ting Wei

Section 2.1: Estimation

General setup

- We saw that a dataset can be modelled as a realization of a random sample from a probability distribution.
- Quantities of interest correspond to features of the model distribution, more precisely to one of the parameters of the model distribution or to a function of the parameters.
- Our task is to use the dataset to estimate the quantity of interest.

Model parameters: an example

- Consider arrivals of packages at a network server. We want to know the number of arrivals during one minute.
- Model this by a random variable with Poisson distribution with (unknown) parameter λ.
- Observe the actual process and get a dataset $x_1, \ldots, x_n$ with $x_i =$ number of arrivals in one minute.
- We want to use this dataset to estimate λ.

Model parameters

- The parameters that determine the model distribution are called model parameters.
- We are interested in estimating model parameters or a function of them.
- For instance, in the above example $e^{-\lambda}$ is the probability that no package arrives in one minute. We may want to estimate this as well.

Estimates

- Let $X_1, \dots, X_n$ be a random sample where X_i follows a distribution with parameter θ. Let $x_1, \dots, x_n$ be an observation of the random sample.
- An estimate of parameter θ is a value that only depends on the dataset, that is a function $\hat{\theta}(x_1, \dots, x_n)$.
- Our computation should give us an indication of the true value of the parameter of interest.

Estimators

- Since our dataset $x_1, \ldots, x_n$ is modelled as a realization of a random sample $X_1, \ldots, X_n$ the estimate $\hat{\theta}(x_1, \ldots, x_n)$ is the realization of the random variable $\hat{\theta}(X_1, \ldots, X_n)$.
- The random variable $\hat{\theta}(X_1, \ldots, X_n)$ is called an estimator of parameter θ.
- Do not confuse the estimate with the estimator! The estimate is a number, the estimator is a random variable.

Sampling distributions

- Let $\hat{\theta}(X_1, \ldots, X_n)$ be an estimator of parameter θ based on a random sample $X_1, \ldots, X_n$.
- The probability distribution of $\hat{\theta}(X_1, \ldots, X_n)$ is called the sampling distribution.

Example

- In the previous example, we can use the sample mean $\overline{x}_n = \frac{1}{n}(x_1 + \cdots + x_n)$ as a natural estimate for λ.
- Thus an estimate for $e^{-\lambda}$ is $e^{-\overline{x}_n}$.
- The corresponding estimator of λ is the random variable $\hat{\lambda} = \frac{1}{n}(X_1 + \cdots + X_n)$.
- The corresponding estimator of $e^{-\lambda}$ is the random variable $e^{-\hat{\lambda}} = e^{-\frac{1}{n}(X_1 + \cdots + X_n)}$.

Example, cont.

- Fact: the sum of n independent $Pois(\lambda)$ random variables has a $Pois(n\lambda)$ distribution.

- Thus the sampling distribution of $\hat{\lambda}$ in the previous example is

$$P\left(\hat{\lambda} = -\frac{k}{n}\right) = P\left(e^{-\hat{\lambda}} = e^{-\frac{k}{n}}\right)$$

$$= P(X_1 + \cdots + X_n = k) = \frac{(n\lambda)^k e^{-n\lambda}}{k!}$$

Unbiased estimators

- An estimator $\hat{\theta}$ is called an unbiased estimator for the parameter θ if $E(\hat{\theta}) = \theta$.
- The difference $E(\hat{\theta}) - \theta$ is called the bias of $\hat{\theta}$.
- If this difference is non-zero, then $\hat{\theta}$ is called a biased estimator.

Estimators of mean and variance

- Let $X_1, \ldots, X_n$ be a random sample with $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$, $i = 1, \ldots, n$.
- We saw in Section 1.5 that $E(\bar{X}) = \mu$ where $\bar{X}$ is the sample mean. Thus the sample mean is an unbiased estimator of μ.
- Recall from Section 1.5 the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
- It can be shown that the sample variance S^2 is an unbiased estimator of σ^2.

Example

- Let X and Y be two random variables with $E[X] = \mu$, $E[Y] = 2\mu$. Show that for any value of a constant c, the variable

$$Z = (1 - 2c)X + cY$$

is an unbiased estimator of μ.

Example

- Let X be the production at a department randomly selected in a day. We observe
  $X_1 = 210,\ X_2 = 220,\ X_3 = 210,\ X_4 = 225,\ X_5 = 220,\ X_6 = 217.$
  Find unbiased estimates for $\mu = E[X]$ and $\sigma^2 = \text{var}[X]$.

Summary

- Estimate and estimators.
- Sampling distributions.
- Unbiased estimators.
- Estimators of means and variance.

Section 2.2: Maximum Likelihood Estimation

Recap: estimates and estimators

- Let the dataset $x_1, \ldots, x_n$ be modelled as a realization of a random sample $X_1, \ldots, X_n$.
- An estimate is a value that only depends on the dataset, that is a function $\hat{\theta}(x_1, \ldots, x_n)$.
- The estimate $\hat{\theta}(x_1, \ldots, x_n)$ is the realization of the random variable $\hat{\theta}(X_1, \ldots, X_n)$.
- The random variable $\hat{\theta}(X_1, \ldots, X_n)$ is called an estimator.
- Do not confuse the estimate with the estimator! The estimate is a ...., the estimator is ....

Fill in the gaps

The need for a general principle

- Sometimes it is easy construct estimators for parameters of interest because these parameters had a natural analogue, such as expectation versus sample mean.
- However, in many situations such an analogue does not exist.
- We therefore need a general principle to construct estimators.

The maximum likelihood principle: the idea

- The idea of the maximum likelihood principle is that, given a dataset, we choose the parameters of interest in such a way that the data are most likely.
- We need a precise mathematical method to formalize this idea.

The likelihood function

- Let $x_1, \cdots x_n$ be a dataset which is the realization from a random sample $X_1, \cdots X_n$.
- The likelihood function $\mathcal{L}(\theta)$ is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} P(X_i = x_i|\theta)$$

if the X_i are discrete, and

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta)$$

if the X_i are continuous.

- Here $P(X_i = x_i|\theta)$ denotes the probability X_i takes the value x_i if the parameter is θ, and $f_{X_i}(x_i|\theta)$ is the density function of X_i if the parameter is θ.

Maximum likelihood estimates (MLE)

- The maximum likelihood estimates (MLE) of the parameter θ is the value $\hat{\theta}(x_1, \ldots, x_n)$ that maximizes the likelihood function $\mathcal{L}(\theta)$.

- The corresponding random variable $\hat{\theta}(X_1, \ldots, X_n)$ is called the maximum likelihood estimator for θ.

Example

- Suppose we have a dataset $x_1, \ldots, x_n$ modelled as a realization of a random sample from an exponential distribution with probability density function

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

- The likelihood function is

$$\mathcal{L}(\lambda) = \ldots\ldots\ldots\ldots$$

Fill in the gaps

Example, cont.

- $\mathcal{L}(\lambda) = \lambda^n e^{-\lambda(x_1+\cdots+x_n)}$.
- To find the MLE of λ we compute the derivative

$$\frac{d\mathcal{L}}{d\lambda} = n\lambda^{n-1}e^{-\lambda\sum_{i=1}^n x_i} - \lambda^n\left(\sum_{i=1}^n x_i\right)e^{-\lambda\sum_{i=1}^n x_i} =$$

$$= n\left(\lambda^{n-1}e^{-\lambda\sum_{i=1}^n x_i}\right)\left(1 - \frac{\lambda}{n}\sum_{i=1}^n x_i\right)$$

- Thus $\dfrac{d\mathcal{L}}{d\lambda} = 0$ if and only if ....

Fill in the gaps

Example, cont.

- Thus $\dfrac{d\mathcal{L}}{d\lambda} = 0$ iff $1 - \lambda\overline{x}_n = 0$
- The MLE of λ is .... while the maximum likelihood estimator is ....
- Checking: As $\dfrac{d^2\mathcal{L}}{d\lambda^2} = -\overline{x}_n < 0$, the MLE of λ is the maximal of $\mathcal{L}(\lambda)$.

Fill in the gaps

Finding the maximum of $\mathcal{L}(\theta)$

- In the previous example it was easy to find the value of the parameter for which the likelihood is maximal.
- However, in general computing the derivative $\dfrac{d\mathcal{L}}{d\theta}$ can be tedious because $\mathcal{L}(\theta)$ is a product of terms, so one needs to use the product rule for differentiation.

The log likelihood

- We thus introduce the log likelihood

$$\ell(\theta) = \log \mathcal{L}(\theta)$$

in which products are turned into sums, which are easier to differentiate.

- Since 'log' is a positively increasing monotone function, $\mathcal{L}(\hat{\theta})$ is maximal if and only if $\ell(\hat{\theta})$ is maximal.

MLE for normal distribution

- Suppose that the dataset $x_1, \ldots, x_n$ is a realization from a $N(\mu, \sigma^2)$ distribution.
- In this case θ is the vector (μ, σ) and the likelihood function is a function of two variables

$$\mathcal{L}(\mu, \sigma) = f_{\mu,\sigma}(x_1) f_{\mu,\sigma}(x_2) \cdots f_{\mu,\sigma}(x_n)$$

where

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

MLE for normal distribution, cont.

$$
\begin{aligned}
\ell(\theta) &= \log \mathcal{L}(\theta) = \log \prod_{i=1}^{n} f_{\mu,\sigma}(x_i) = \\
&= \sum_{i=1}^{n} \log f_{\mu,\sigma}(x_i) = \sum_{i=1}^{n} \log \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} \right) = \\
&= \sum_{i=1}^{n} \left(\log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2}(x_i-\mu)^2 \right) = \\
&= -n\log\sigma - n\log\sqrt{2\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}
$$

Differentiating the log likelihood

- From previous slide

$$\ell(\mu, \sigma) = -n \log \sigma - n \log \sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 .$$

- We compute

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = \frac{n}{\sigma^2} (\overline{x}_n - \mu)$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$= -\frac{n}{\sigma^3} \left(\sigma^2 - \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$$

Conclusion

- Solving $\dfrac{\partial \ell}{\partial \mu} = 0$ and $\dfrac{\partial \ell}{\partial \sigma} = 0$ yields

$$\hat{\mu} = \overline{x}_n, \qquad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}$$

Invariance property

- The following is an important property of MLE. We do not prove it, but we will use it repeatedly in studying linear statistical models.
- Invariance property of MLE: if $\hat{\theta}$ is the maximum likelihood estimator of a parameter θ and $g(\theta)$ is an invertible function of θ, then $g(\hat{\theta})$ is the maximum likelihood estimator for $g(\theta)$.
- Notation: we will sometimes denote the MLE of θ by $\hat{\theta}$. In this case the above reads $\widehat{g(\theta)} = g(\hat{\theta})$. That is you can 'bring the hat inside the function'.

Summary

- Likelihood function.
- Maximum likelihood estimates and estimators.
- Log likelihood.
- Example: normal distribution.
- Invariance property of MLE.

Section 2.3: Confidence Intervals

The limitations of point estimates

- Suppose we have an estimator $\hat{\theta}(X_1, \ldots, X_n)$ of an unknown parameter θ.
- We use its realization $\hat{\theta}(x_1, \ldots, x_n)$, based on a dataset from an experiment, as our estimate for θ.
- Suppose we repeat the experiment many times: do you expect the estimates to remain the same?

The need for interval estimates

- We cannot say that the estimate $\hat{\theta}(x_1, \ldots, x_n)$ equals the true value of θ, but rather than it is only close to the true θ.
- We want to provide an interval of plausible values for θ and also add a specific statement about how confident we are that the true θ is among them.
- This will be based on the knowledge of the sampling distributions of corresponding estimators.

Confidence intervals

- Suppose a dataset $x_1, \ldots, x_n$ is given, modelled as realization of random variables $X_1, \ldots, X_n$. Let θ be the parameter of interest, and γ a number between 0 and 100.

- If there exist sample statistics $\mathcal{L}_n = g(X_1, \ldots, X_n)$ and $\mathcal{U}_n = h(X_1, \ldots, X_n)$ such that

  $$P(\mathcal{L}_n < \theta < \mathcal{U}_n) = \gamma\%$$

  for every value of θ, then (ℓ_n, u_n) where $\ell_n = g(x_1, \ldots, x_n)$ and $u_n = h(x_1, \ldots, x_n)$ is called a $\gamma\%$ confidence interval for θ.

- The number $\gamma\%$ is called the confidence level.

Confidence intervals, cont.

- There is no way of knowing whether an individual confidence interval is correct, in the sense it indeed does cover θ.

- The procedure guarantees that each time we make a confidence interval we have the probability γ% of covering θ.

Example: Normal data (variance known)

- Suppose the data can be seen as the realization of a sample $X_1, \ldots, X_n$ from $N(\mu, \sigma^2)$ distribution and μ is the unknown parameter of interest, while σ^2 is known.
- The mean $\overline{X}_n$ has an $N\left(\mu, \frac{\sigma^2}{n}\right)$ distribution.
- Therefore,

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \ .$$

by central limit theorem.

Example, cont.

- If c_ℓ and c_u are chosen such that $P(c_\ell < Z < c_u) = \gamma\%$ for Z a $N(0,1)$ distributed random variable, then

$$\begin{aligned}\gamma\% =& P\left(c_\ell < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < c_u\right) = P\left(c_\ell \frac{\sigma}{\sqrt{n}} < \overline{X}_n - \mu < c_u \frac{\sigma}{\sqrt{n}}\right) \\ =& P\left(\overline{X}_n - c_u \frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n - c_\ell \frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

- Thus, if $\ell_n = \overline{x}_n - c_u \frac{\sigma}{\sqrt{n}}$ and $u_n = \overline{x}_n - c_\ell \frac{\sigma}{\sqrt{n}}$ the interval (ℓ_n, u_n) covers μ with probability $\gamma\%$.

Example, cont.

- Thus $\left(\overline{x}_n - c_u \frac{\sigma}{\sqrt{n}}, \overline{x}_n - c_\ell \frac{\sigma}{\sqrt{n}}\right)$ is a γ% confidence interval for parameter μ.

- A common choice is to divide $\alpha = 1 - \gamma\%$ evenly between the two tails of the distribution, that is

$$P(Z \geq c_u) = \alpha/2 \text{ and } P(Z \leq c_\ell) = \alpha/2$$

so that $c_u = Z_{\alpha/2}, \quad c_\ell = -Z_{\alpha/2}$.

- In summary, the $100(1-\alpha)\%$ C.I. for μ is

$$(\overline{x}_n - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \overline{x}_n + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}})$$

- If $\alpha = 0.05$ then $Z_{\alpha/2} = 1.96$.

Example: Normal data (variance unknown)

- Suppose the data can be seen as a random sample $X_1, \dots, X_n$ from a $N(\mu, \sigma^2)$ distribution where both μ and σ^2 are unknown.
- The fact that $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ has the standard normal distribution is not useful, since the corresponding C.I. contains σ, which is unknown.

Example: cont.

- We use the following general fact (without proof):

  For a random sample $X_1, \ldots, X_n$ from a $N(\mu, \sigma^2)$ distribution,

  $$T_n = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$

  where S_n^2 is the sample variance.

- In other words, replacing σ with S_n changes $N(0,1)$ into t_{n-1}.

- Discussion question: what is the corresponding $100(1-\alpha)\%$ confidence interval for μ?

Example, cont.

- The $100(1-\alpha)\%$ C.I. for μ is

$$\left(\overline{x}_n - t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}},\ \overline{x}_n + t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}}\right)$$

  where $P(T_n \geq t_{\alpha/2,n-1}) = \alpha/2$ and $P(T_n \leq -t_{\alpha/2,n-1}) = \alpha/2$.

- Find the value of $t_{\alpha/2,n-1}$ from the t-distribution table.

- Or using R programming, $t_{\alpha/2,n-1} =$

```
-qt(alpha/2,df)
```

Approximate C.I.

- If we have a random sample whose distribution we approximately know, we can use our confidence interval procedure to derive approximate C.I. for the parameters.

- For large datasets the central limit theorem ensures that this method provides confidence intervals with approximately correct confidence levels.

Example

- Suppose we know that $B \sim \text{Bin}(n, p)$, but we don't know p and would like to produce a 95% confidence interval for it.

- In this case we can use the normal approximation to the binomial for large n, say

$$\frac{B - np}{\sqrt{np(1-p)}} \sim N(0, 1) \; .$$

Example

- Recall that if $Z \sim N(0,1)$ then $P(-1.96 < Z < 1.96) = 0.95$. Therefore

$$P\left(-1.96 < \frac{B - np}{\sqrt{np(1-p)}} < 1.96\right) \approx 0.95$$

- Suppose we observe a value b of B, so that the approximate 95% confidence interval for p is based on this observation, that is:

$$-1.96 < \frac{b - np}{\sqrt{np(1-p)}} < 1.96.$$

Example, cont.

- Squaring the last inequalities we obtain

$$\frac{(b-np)^2}{np(1-p)} < 1.96^2$$

- Rearranging as a function of p we have

$$p^2 - \frac{n(2b+1.96^2)}{1+1.96^2 n} p + b^2 < 0.$$

- The left hand side is a quadratic function of p, so the interval we want is the interval between the two roots of the quadratic.

Summary

- The need for interval estimates.
- Confidence intervals.
- Normal data (variance known).
- Normal data (variance unknown).
- Approximate confidence intervals.

Section 2.4: Testing Hypotheses

Null and alternative hypotheses

- The first of the two competing propositions is called the null hypothesis, denoted H_0 and the second one is called the alternative hypothesis, denoted H_1.
- The null hypothesis is presumed to be true until the data provide convincing evidence against it.
- If we reject the null hypothesis we will accept H_1.
- The next step is a criterion that provides an indication about whether H_0 is false. This involves a test statistic.

Test statistic

- Suppose that the dataset is modelled as the realization of random variables $X_1, \ldots, X_n$.
- A test statistic is any sample statistic $\mathrm{T} = h(X_1, \ldots, X_n)$ whose numerical value is used to decide whether we reject H_0.
- The values of the test statistic can be viewed on a credibility scale for H_0, and we must determine which of these values provide evidence in favor of H_0, and which provide evidence in favor of H_1.

Test for a single mean (variance known)

- For a random sample $X_1, \ldots, X_n$, let $E(X_i) = \mu$, $\mathrm{var}(X_i) = \sigma^2$, $i = 1, \ldots, n$. Let the sample statistic T be the sample mean $\bar{X}$.
- We want to test if μ is equal to a constant μ_0 or not. When σ^2 is known, $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$,
  - $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$,

    If $P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right) \leqslant \alpha$, reject H_0.
  - $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$,

    If $P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right) \leqslant \alpha$, reject H_0.
  - $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$,

    If $P\left(\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| > \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right|\right) \leqslant \alpha$, reject H_0.

Test for a single mean (variance unknown)

- We want to test if μ is equal to a constant μ_0 or not. When σ^2 is unknown, and let S^2 be the sample variance, $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$,

  - $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$,

    If $P\left(\frac{\bar{X}-\mu_0}{S/\sqrt{n}} > \frac{\bar{x}-\mu_0}{S/\sqrt{n}}\right) \leqslant \alpha$, reject H_0.

  - $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$,

    If $P\left(\frac{\bar{X}-\mu_0}{S/\sqrt{n}} < \frac{\bar{x}-\mu_0}{S/\sqrt{n}}\right) \leqslant \alpha$, reject H_0.

  - $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$,

    If $P\left(\left|\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\right| > \left|\frac{\bar{x}-\mu_0}{S/\sqrt{n}}\right|\right) \leqslant \alpha$, reject H_0.

Example

- Suppose that we know that blood pressures in humans from Honolulu follow a normal distribution, but we don't know the mean μ.
- For the rest of the U.S. The mean is known to be 120 mm Hg, and the standard deviation is 10 mm Hg.
- Some people think that Honolulans have different blood pressure on average with other Americans, so we want to test the hypothesis H_0 that $\mu = 120$ against the alternative H_1 that $\mu \neq 120$.

Example, cont.

- We measure the blood pressure of 100 Honolulans selected independently at random and compute the mean $\bar{x}$, which turns out to be 130.1 mm Hg.
- We make the assumption that the standard deviation for blood pressure of Honolulans is also 10mm Hg.
- Since we have 100 independent observations of a $N(\mu, 100)$ random variable, we have

  $$\frac{\bar{X} - \mu}{10/\sqrt{100}} \sim N(0, 1)$$

  where $\bar{X}$ is the sample mean.

Example, cont.

- If the mean is really 120 (that is, H_0 is true) then

$$Z = \frac{\bar{X} - 120}{10/\sqrt{100}} \sim N(0,1)$$

- Hence the probability of observing a sample mean of not 130.1 mm Hg

$$P\left(\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| > \left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right|\right) = P\left(\left|\frac{\bar{X} - 120}{10/\sqrt{100}}\right| > \left|\frac{130.1 - 120}{10/\sqrt{100}}\right|\right)$$
$$= P\left(Z > \frac{130.1 - 120}{10/\sqrt{100}}\right) + P\left(Z < -\frac{130.1 - 120}{10/\sqrt{100}}\right)$$
$$= P(Z > 10.1) + P(Z < -10.1) < 0.001$$

- This probability is called the *p*-value of our test. Since it is very small, we can regard it as evidence to reject H_0 in favour of our alternative that $\mu \neq 120$.

Error types

- There are two situations in which the decision made on the basis of data is wrong:
  - The null hypothesis H_0 may be true, whereas data lead to rejection of H_0.
  - The alternative hypothesis H_1 may be true, whereas we do not reject H_0 on the basis of the data,
- A type I error occurs if we falsely reject H_0. A type II error occurs if we falsely do not reject H_0.
- The question is: what should be the probability of committing a type I error, i.e. for which values of T should we reject H_0?

Significance level

- The significance level is the largest acceptable probability of committing a type I error and is denoted by α, where $0 < \alpha < 1$.
- We speak of 'performing the test at level α' as well as 'rejecting H_0 in favor of H_1 at level α'.
- We usually take $\alpha = 0.05$.

Critical region and critical values

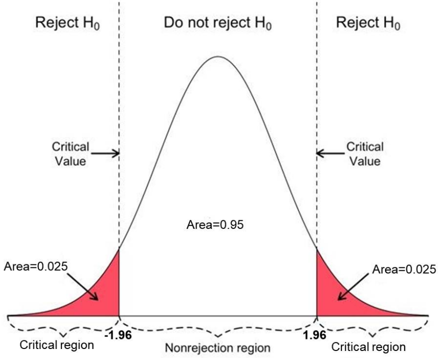
- Suppose we test H_0 against H_1 at significance level α by means of a test statistic T.
- The set $K \subset \mathbb{R}$ that corresponds to all values of T for which we reject H_0 in favour of H_1 is called the critical region. Values on the boundary of the critical region are called critical values.
- The precise shape of the critical region depends on both the chosen significance level α and the test statistic T that is used. But it will always be such that the probability that $\mathrm{T} \in K$ satisfies

$$P(\mathrm{T} \in K) \leq \alpha \qquad \text{in the case that } H_0 \text{ is false.}$$

Example

- Suppose the test statistic is $N(0,1)$ and the significance level is $\alpha = 0.05$. Then the critical values are ± 1.96, while the critical region is $(-\infty, -1.96) \cup (1.96, +\infty)$.

- In picture: the sum of the areas of the two tails (red) is 0.05.



p-values

- If the observed value of the statistic falls in the critical region, we reject the null hypothesis H_0.

- The 2-sided *p*-value is the sum of the areas of the two tail probabilities $P(\mathrm{T} \leq -\mathrm{t}) + P(\mathrm{T} \geq \mathrm{t})$. The left-sided *p*-value is $P(\mathrm{T} \leq -\mathrm{t})$ and the right-sided *p*-value is $P(\mathrm{T} \geq \mathrm{t})$.

- The *p*-value expresses how likely is to obtain a value of the test statistic T at least as extreme as the value t obtained for the data.

- The smaller the *p*-value, the stronger evidence the observed value t bears against H_0.

Test for a single mean, unknown variance

- Suppose we can take independent samples from a normal distribution $N(\mu, \sigma^2)$ in which both μ and σ^2 are unknown. Then

$$\mathrm{T} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

- Suppose that we observe $\bar{x} = 102$, $s = 4.7$ with a sample size $n = 25$. We want to test the null hypothesis H_0 that $\mu = 100$ against the alternative H_1 that $\mu > 100$ at significance level $\alpha = 0.05$.

Test for a single mean, unknown variance, cont.

- We have $\frac{102-100}{4.7/5} = 2.13 \sim t_{24}$.
- The p-value is $P\left(\mathrm{T} > \frac{102-100}{4.7/5}\right) = 0.0218 < 0.05$.
- Alternatively, read from the t distribution table, the critical value is 1.711, the critical region is $(1.711, +\infty)$.
- Therefore we reject the null hypothesis.

Hypothesis tests and confidence intervals

- Hypothesis tests and confidence intervals are equivalent ways to do interval estimation.
- Suppose that for some parameter θ we test $H_0 : \theta = \theta_0$. Hence we reject $H_0 : \theta = \theta_0$ in favour of H_1 at level α if and only if θ_0 is not in the $100(1-\alpha)\%$ C.I. for θ.
- Note: If the hypothesis test and the C.I. give contradictory results, it means you have made a calculation mistake!

Summary

- Null and alternative hypotheses.
- Test statistic for hypothesis testing.
- Type I and type II errors.
- Significance level.
- Critical values and critical regions.
- p-values.
- Example: test for single mean, unknown variance.
- Relation between hypothesis tests and confidence intervals.