Univariate Statistics: inference and testing

Denis Allard

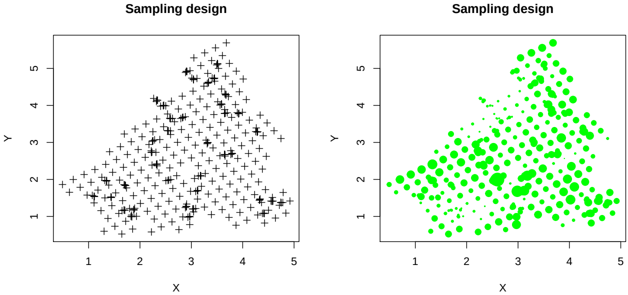
Biostatistique et Processus Spatiaux (BioSP), INRA, Avignon

Ecole Doctorale UAPV "Sciences et Agrosciences", Statistiques paramétriques,
7 juin 2016

Program

1. The statistical thinking; basic definitions
2. Univariate statistics: inference and testing
3. Simple regression, linear model and ANOVA
4. Applications to environmental statistics: time series and geostatistics

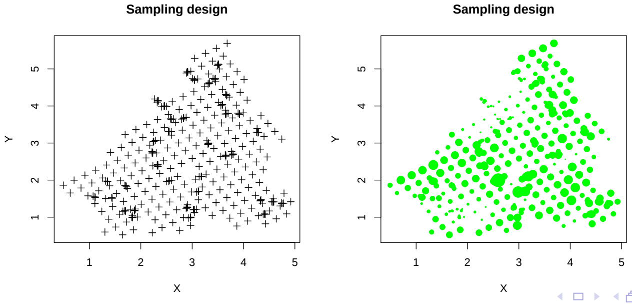
The data set



- 359 data, sampled in Swiss Jura, approx. 25 km^2 study area
- Sampling design: regular grid + random local densification
- Content of 7 heavy metals: Cd, Co, Cr, Cu, Ni, Pb, Zn
- 5 rock types: Argovian,Kimmeridgian, Sequenian, Portlandian, Quaternary
- 4 Land Use: forests, pastures, grasslands, tillage

The data set

```
> jura = read.table("jura_tout.txt",header=TRUE)
> jura[1,]
      x     y lu rt   Cd   Co    Cr    Cu    Ni    Pb    Zn
1 2.386 3.077  3  3 1.74 9.32 38.32 25.72 21.32 77.36 92.56
> Ni  = jura$Ni
> par(mfrow=c(1,2))
> plot(jura$x,jura$y,main="Sampling design",xlab="X",ylab="Y",pch=3)
> plot(jura$x,jura$y,main="Sampling design",xlab="X",ylab="Y",
> pch=19,cex=Ni/20,col="green")
```



Some probability facts

Gaussian (Normal) Random Variable

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}, \quad x \in \mathbf{R}.$$

Then,

$$E[X] = \mu; \quad \mathrm{Var}(X) = \sigma^2.$$

For short:

$$X \sim \mu + \sigma\mathcal{N}(0,1)$$

Quantiles

Let us denote u_p the value such that

$$\mathbb{P}(\mathcal{N}(0,1) \leq u_p) = p$$

Example:

$$p = 0.975 \Leftrightarrow u_p = 1.96$$

Estimating the mean

- Let $X_1, \ldots, X_n$ be an i.i.d. n-sample, arising from a population with mean μ and variance σ^2.
- The arithmetic average

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the natural estimator of μ.
- It is a random variable with

$$E[\bar{X}] = \mu; \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Central Limit Theorem

Provided $\sigma^2 < \infty$, as $n \to \infty$, we have

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0, 1)$$

Estimating the mean

Sample of size 30, without repetitions

1. Mean # 1: 20.95
2. Mean # 2: 20.35
3. Mean # 3: 20.47

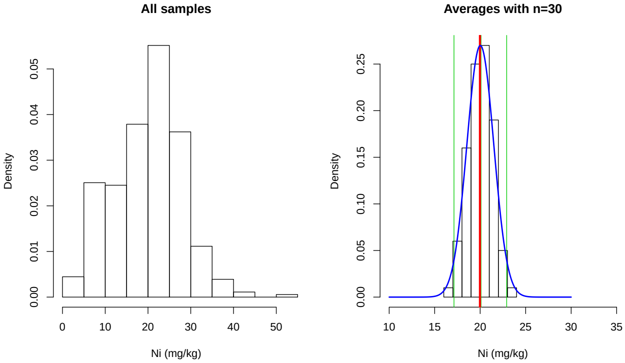
   ⋮

4. Mean #100: 18.12

Average of means: 19.91

Average of Ni: 20.02

Estimating the mean



Estimating the mean

```
> ndata = 30;  nrep  = 100
> xbar  = rep(NULL,nrep)
> for (i in 1:nrep) xbar[i] = mean(sample(Ni)[1:ndata])
> par(mfrow=c(1,2))
> hist(Ni,xlab="Ni (mg/kg)",main="All samples",
>      xlim=c(0,max(Ni)),probability=TRUE)
> hist(xbar,xlab="Ni (mg/kg)",main="Averages with n=30",
>      xlim=c(10,35),probability=TRUE)
> abline(v=mean(Ni),col=3,lwd=3)
> abline(v=mean(xbar),col=2,lwd=3)
> abline(v=mean(Ni)-1.96*sqrt(var(Ni)/ndata),col=3,lwd=1)
> abline(v=mean(Ni)+1.96*sqrt(var(Ni)/ndata),col=3,lwd=1)
> z = seq(10,30,by=0.1)
> f = dnorm(z,mean=mean(Ni),sd=sqrt(var(Ni)/ndata))
> lines(z,f,type="l",col="blue",lwd=2)
```

Estimating the mean

Summary

- The estimate of mean, $\bar{X}$ is random, because the sampling is random
- It is unbiased: $E[\bar{X}] = \mu$
- $\text{Var}[\bar{X}] = \sigma^2/n$
- The esimate of the mean is within

$$[\mu - 2\sigma/\sqrt{n}; \mu + 2\sigma/\sqrt{n}]$$

with probability 95%.

Confidence Interval for the mean

1. $X_1, \ldots, X_n$ i.i.d samples with $E[X] = \mu$
2. Let us find an interval $[\hat{\mu}_{inf}, \hat{\mu}_{sup}]$ containing the true value $\mu = E[X]$ with probability $1 - \alpha$: we call it the level.
3. We set the error on both sides

$$\mathbb{P}(\mu < \hat{\mu}_{inf}) = \mathbb{P}(\mu \geq \hat{\mu}_{sup}) = \alpha/2.$$

Confidence Interval: σ^2 is known

Let us first assume that σ^2 is known. Then, as $n \to \infty$.

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = [\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}],$$

where u_p is the value such that

$$\mathbb{P}(\mathcal{N}(0,1) \leq p) = u_p$$

Confidence Interval for the mean

Confidence Interval: σ^2 is known

Let us first assume that σ^2 is known. Then,

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = [\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

Proof

Using CTL,

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(-u_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq u_{1-\alpha/2}\right) \\ &= \mathbb{P}\left(-u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} - \bar{X} \leq -\mu \leq u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} - \bar{X}\right) \\ &= \mathbb{P}\left(\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

Confidence Interval for the mean

Confidence Interval: σ^2 is known

Let us first assume that σ^2 is known. Then,

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = [\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

The width of the CI interval

- Increases with $1-\alpha$:

$$\begin{aligned} \alpha = 10\% \quad &\Leftrightarrow \quad u_{0.950} = 1.64 \\ \alpha = 5\% \quad &\Leftrightarrow \quad u_{0.975} = 1.96 \\ \alpha = 1\% \quad &\Leftrightarrow \quad u_{0.995} = 2.58 \end{aligned}$$

- Increases with the variance
- Decreases as $1/\sqrt{n}$
- 1000 repetitions – samples of size 30. With $\alpha = 5\%$, one finds

$$\#\{\mu < \hat{\mu}_{inf}\} = 24; \quad \#\{\mu > \hat{\mu}_{sup}\} = 19,$$

where 25 expected.

Experiment

Sample of size 30, without repetitions

1. Variance # 1: 61.04
2. Variance # 2: 62.13
3. Variance # 3: 53.01

   ⋮

4. Variance #100: 68.7

Variance of Ni: 65.51

More probability facts

Estimation of the variance

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is unbiased, i.e.

$$E[\hat{\sigma}^2] = \sigma^2.$$

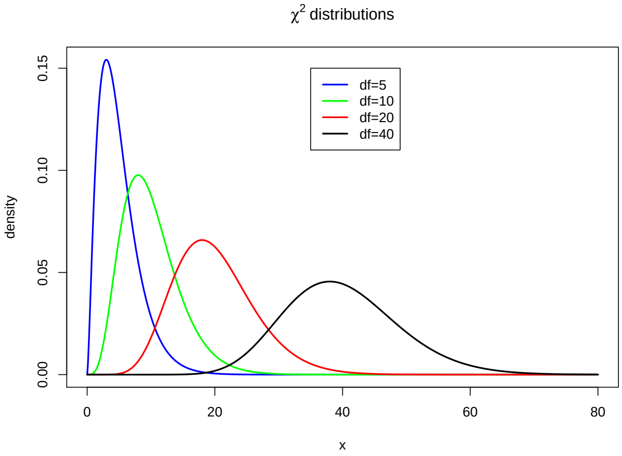
χ^2 distribution

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a $\mathcal{N}(\mu, \sigma^2)$ RV. Then,

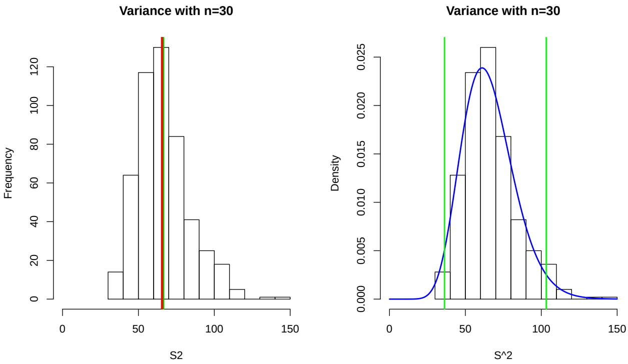
$$(n-1)S^2/\sigma^2 \sim \chi^2_{(n-1)}.$$

There are $(n-1)$ independent RV (degrees of freedom) when computing S^2.

The χ^2 distributions



Illustration



Confidence Interval for the variance

Confidence Interval at level α

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a $\mathcal{N}(\mu, \sigma^2)$ RV. Then,

$$[\hat{\sigma}^2_{inf}, \hat{\sigma}^2_{sup}] = [S^2(n-1)/x^{(n-1)}_{\alpha/2}, S^2(n-1)/x^{(n-1)}_{1-\alpha/2}],$$

where $x^{(n-1)}_p$ is such that $\mathbb{P}(\chi^2_{n-1} \leq p) = x^{(n-1)}_p$.

Proof

Using convergence towards χ^2,

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left(x^{(n-1)}_{\alpha/2} \leq (n-1)S^2/\sigma^2 \leq x^{(n-1)}_{1-\alpha/2}\right) \\
&= \mathbb{P}\left(1/x^{(n-1)}_{1-\alpha/2} \leq \sigma^2/(S^2(n-1)) \leq 1/x^{(n-1)}_{\alpha/2}\right) \\
&= \mathbb{P}\left(S^2(n-1)/x^{(n-1)}_{1-\alpha/2} \leq \sigma^2 \leq S^2(n-1)/x^{(n-1)}_{\alpha/2}\right)
\end{aligned}$$

Even more probability facts

Student t distribution

Let $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_{(n)}$ be independent. Then,

$$\frac{X}{Y/\sqrt{n}} \sim t_n$$

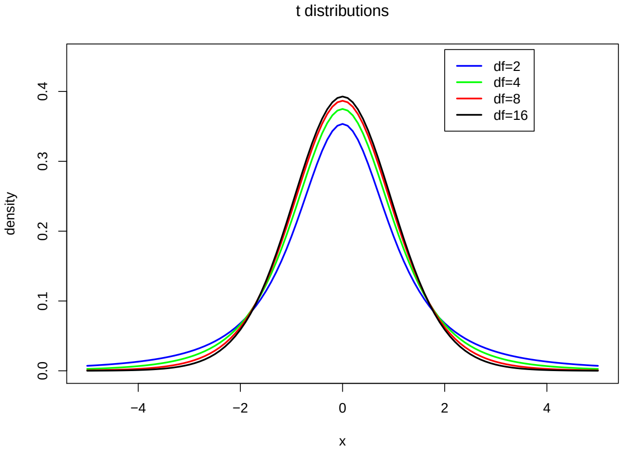
t distribution with n d.o.f.

Fisher F distribution

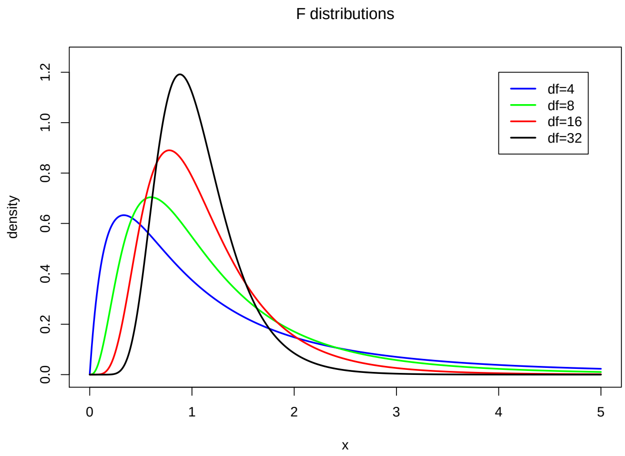
Let $X \sim \chi^2_{(n_X)}$ and $Y \sim \chi^2_{(n_Y)}$ be independent. Then,

$$\frac{X/\sqrt{n_X}}{Y/\sqrt{n_Y}} \sim F^{n_X}_{n_Y}.$$

The Student *t* distributions



The Fisher *F* distributions



Confidence interval Revisited

Confidence Interval: σ^2 is estimated

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a $\mathcal{N}(\mu, \sigma^2)$ RV. Using the definition of the t Student distribution:

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = [\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}]$$

where

$$\mathbb{P}(t_{(n-1)} \leq t_{1-\alpha/2}) = 1 - \alpha/2.$$

If $X_1, \ldots, X_n$ is not Gaussian, requires $n > 30$.

Confidence interval Revisited

Proof

Using $(n-1)S^2/\sigma^2 \sim \chi^2_{(n-1)}$ and the definition of a t distribution,

$$
\begin{aligned}
1-\alpha &= \mathbb{P}\left(-t_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{S/\sqrt{n-1}} \leq t_{1-\alpha/2}\right) \\
&= \mathbb{P}\left(-t_{1-\alpha/2}\frac{S}{\sqrt{n-1}} - \bar{X} \leq -\mu \leq t_{1-\alpha/2}\frac{S}{\sqrt{n-1}} - \bar{X}\right) \\
&= \mathbb{P}\left(\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n-1}} \leq \mu \leq \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n-1}}\right)
\end{aligned}
$$

Confidence interval Revisited

Confidence Interval: σ^2 is estimated

Using the definition of the t Student distribution:

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = [\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}]$$

Since $t_{1-\alpha/2} \geq u_{1-\alpha/2}$, the interval is wider as compared to the case with known σ^2.
With $n = 30$:

$$\begin{aligned} \alpha = 10\% \quad &\Leftrightarrow \quad t_{0.950} = 1.70 \quad [u_{0.950} = 1.64] \\ \alpha = 5\% \quad &\Leftrightarrow \quad t_{0.975} = 2.04 \quad [u_{0.975} = 1.96] \\ \alpha = 1\% \quad &\Leftrightarrow \quad t_{0.995} = 2.76 \quad [u_{0.995} = 2.58] \end{aligned}$$

Same 1000 samples of size 30. One finds

$$\#(\mu < \hat{\mu}_{inf}) = 29 \quad [24]; \quad \#(\mu > \hat{\mu}_{sup}) = 24 \quad [19].$$

Expected value: 25.

Statistical tests

- According to former studies and/or expertise, one should have $\mu = 20$.
- A sample of size 30 provides $\bar{X} = 22.2$ and $S^2 = 52$.
- Is this a significant difference?

$\rightsquigarrow$ Need for formal statistical tests

Definition

Statistical test = Mathematical decision tool to check an hypothesis.

- Neutral, or "null" hypothesis, H_0
- Alternative hypothesis, H_1

H_0 is not guilty unless proven otherwise.

Statistical tests

Test

$$H_0 \text{ vs. } H_1$$

We always test H_0 against an alternative. Both have to be clearly defined.

Two types of errors

	Decision	
	Do not reject H_0 Keep H_0	Reject H_0 Prefer H_1
H_0 true	Correct	Type I Error
H_1 true	Type II Error	Correct

- Level

$$\alpha = \mathbb{P}(\text{Type I Error})$$

(to be computed conditional on H_0 being true)

- Power

$$1 - \beta = \mathbb{P}(\text{No Type II Error})$$

(to be computed conditional on H_1 being true)

Statistical test: the very, very general procedure

H_0 is supposed to be true unless proven to be false.
$\Rightarrow$ computations are done conditional on H_0.

1. Define clearly the hypotheses H_0 and H_1
2. Set the level α
3. Use the relevant statistics (this is where the mathematical theory comes in), say T
4. Find the critical value of T, denoted t_c, as a function of n, α
5. Compute the value of T for the given sample, and compare to t_c
6. Conclude whether H_0 should be rejected or not

Power of a test

- The level α is set by the user.

$$1 - \alpha = \mathbb{P}(H_0 \mid H_0) = \mathbb{P}(\text{Not rejecting } H_0 | H_0 \text{ is true})$$

- Power

$$1 - \beta = \mathbb{P}(H_1 \mid H_1) = \mathbb{P}(\text{Rejecting } H_0 | H_1 \text{ is true})$$

Necessitates a complete specification H_1.

Example: testing the mean

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu = \mu_1 > \mu_0$$

The power $1 - \beta$ is a function of μ_1.

Testing the mean

The average of Ni should be $\mu = 20$. In a sample of size 30, it is found that $\bar{X} = 22.2$ and $S^2 = 52$.

Should H_0 be rejected ?

1. Define the hypotheses $H_0 : \mu = 20$; $H_1 : \mu > 20$
2. Set a level: $1 - \alpha = 0.05$
3. Use the relevant distribution: $(\bar{X} - \mu)/(S/\sqrt{n-1}) \sim t_{(n-1)}$ with $n = 30$
4. If $(\bar{X} - \mu)/(S/\sqrt{n-1}) \sim t_{n-1}$ is "too large" I should reject H_0
5. One reads $\mathbb{P}(t_{(29)} \leq t_c) = 0.95$.

   t_c is the critical value. Here, $t_c = 1.70$.
6. $(\bar{X} - \mu)/(S/\sqrt{n-1}) = (22.2 - 20)/\sqrt{52/29} = 1.64 < 1.7$
7. The null hypothesis H_0 is not rejected.

   "The sample was not able to prove H_0 was guilty"

Testing the mean: assessing the power

Example: testing the mean

$$H_0 : \ \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu = \mu_1 > \mu_0$$

$X_1 = X_0 + (\mu_1 - \mu_0) \sim \mathcal{N}(\mu_0 + (\mu_1 - \mu_1), \sigma^2)$

Some mathematics

$$\begin{aligned} \mathbb{P}(H_1 \mid H_1) &= \mathbb{P}\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n-1}} \geq t_c\right) \\ &= \mathbb{P}\left(\frac{\bar{X}_0}{S/\sqrt{n-1}} \geq t_c - \frac{\mu_1 - \mu_0}{S/\sqrt{n-1}}\right) \\ &= 1 - F_{t_{n-1}}\left(t_c - \frac{\mu_1 - \mu_0}{S/\sqrt{n-1}}\right) \end{aligned}$$

Testing the mean: assessing the power

Example: unilateral tests for the mean

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu = \mu_1 > \mu_0$$

$$X_1 = X_0 + (\mu_1 - \mu_0) \sim \mathcal{N}(\mu_0 + (\mu_1 - \mu_1), \sigma^2)$$

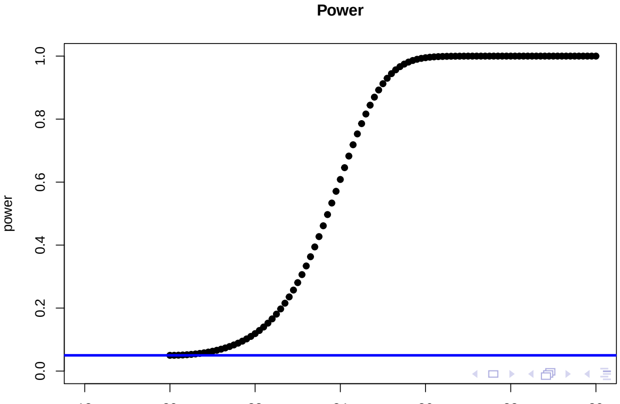
```
> delta = seq(0,10,by=0.1)
> tc    = qt(0.95,df=29)
> pow   = 1 - pt(tc - delta*delta/sqrt(var(Ni)),df=29)
> plot(20+delta,pow,main="Power",xlim=c(18,30),ylim=c(0,1),
>      xlab=expression(mu[1]),ylab="power",pch=19)
> abline(h=0.05,lwd=3,col="blue")
```

Testing the mean: assessing the power

Example: unilateral tests for the mean

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu = \mu_1 > \mu_0$$

$X_1 = X_0 + (\mu_1 - \mu_0) \sim \mathcal{N}(\mu_0 + (\mu_1 - \mu_1), \sigma^2)$



p-value

We do not set a level beforehand. Instead, one computes the probability of rejecting H_0, given the data.

Definition of the p-value

The probability of obtaining an "equal or more extreme" test statistics than what was actually observed, assuming H_0 is true.

- A small p-value (≤ 0.05) indicates strong evidence against the null hypothesis, so it is rejected.
- A large p-value (> 0.05) indicates weak evidence against the null hypothesis (fail to reject).
- p-values very close to the cutoff (~ 0.05) are considered to be marginal (need attention).

Back to our first example

The average of Ni should be $\mu = 20$. In a sample of size 30, it is found that $\bar{X} = 22.2$ and $S^2 = 52$.

Should H_0 be rejected ?

$$\begin{aligned} p &= 1 - \mathbb{P}\left(t_{(n-1)} \leq (\bar{X} - \mu)/(S/\sqrt{n-1})\right) \\ &= 1 - \mathbb{P}\left(t_{29} \leq (22.2 - 20)/\sqrt{52/29}\right) \\ &= 1 - \mathbb{P}\left(t_{29} \leq 1.643\right) \\ &= 0.0556 \end{aligned}$$

Fail to reject, but not by much. Requires attention.

Testing the variance

The variance of Ni should be $\sigma^2 = 65.5$. In a sample of size 30, it is found that $S^2 = 90.2$.

Should H_0 be rejected ?

1. Define the hypotheses $H_0 : \sigma^2 = 65.5$; $H_1 : \sigma^2 > 65.5$
2. Use the relevant distribution:

$$S^2/(\sigma^2/n) \sim \chi_{(n-1)},$$

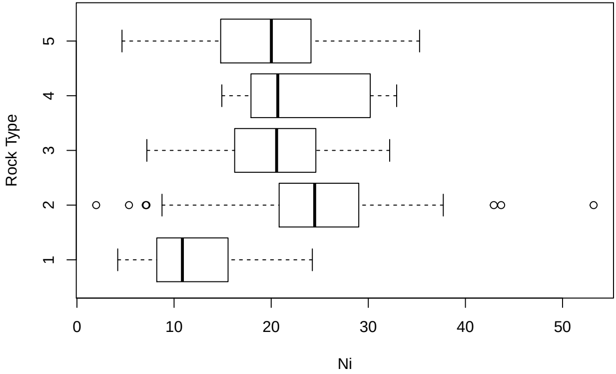
   with $n = 30$
3. One reads $\mathbb{P}(\chi_{(29)} \leq 90.2 * 30/65.5) = 0.935$.

   The p-value is 0.065
4. The null hypothesis H_0 is not rejected.

   But close

Back to data

```
> boxplot(Ni ~ jura$rt, horizontal=T, xlab="Ni",ylab="Rock Type")
```



- Different means?
- Different variances?

Testing two means

Test

$$H_0: \mu_1 = \mu_2; \quad H_1: \mu_1 \neq \mu_2$$

i.e.

$$H_0: \mu_1 - \mu_2 = 0; \quad H_1: \mu_1 - \mu_2 \neq 0$$

with $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Under Gaussian hypothesis, we have

$$\frac{n_1 S_1^2}{\sigma^2} \sim \chi^2_{n_1-1}; \qquad \frac{n_2 S_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$$

and

$$\bar{X}_1 \sim \mathcal{N}(\mu_1, \sigma^2/\sqrt{n_1}) \qquad \bar{X}_2 \sim \mathcal{N}(\mu_2, \sigma^2/\sqrt{n_2})$$

Testing two means

Hence,

$$\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

and

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2 = 0, \sigma^2(1/n_1 + 1/n_2)\right).$$

Therefore, the test statistics is

$$T = \frac{\bar{X}_1 - \bar{X}_2}{(n_1 S_1^2 + n_2 S_2^2)(1/n_1 + 1/n_2)} \sqrt{n_1 + n_2 - 2} \sim t_{(n_1+n_2-2)}$$

Testing two means: example

Jura data:

rt	1	2	3	4	5
$\bar{X}$	12.3	25.0	20.4	22.9	18.8
S^2	31.0	54.6	32.0	50.5	57.2
n	76	124	89	6	64

Mean, variance and number of data, according to rock type

T-tests:

	2	3	4	5
1	0	0	$1.6\,10^{-5}$	$1.5\,10^{-5}$
2	–	$1.1\,10^{-5}$	0.25	$1.2\,10^{-7}$
3	–	–	0.16	0.07
4	–	–	–	0.10

p-value of T tests, assuming identical variance

Testing two variances

Test

$$H_0 : \sigma_1^2 = \sigma_2^2; \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$$

i.e.

$$H_0 : \sigma_1^2/\sigma_2^2 = 1; \qquad H_1 : \sigma_1^2/\sigma_2^2 \neq 1$$

Under Gaussian hypothesis, we have

$$\frac{n_1 S_1^2}{\sigma^2} \sim \chi^2_{n_1-1}; \qquad \frac{n_2 S_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$$

and

$$\frac{n_1 S_1^2}{n_1 - 1} \Big/ \frac{n_2 S_2^2}{n_2 - 1} \sim F_{n_1-1, n_2-1}.$$

Testing two variances: example

Jura data:

rt	1	2	3	4	5
S^2	31.0	54.6	32.0	50.5	57.2
n	76	124	89	6	64

Variance and number of data, according to rock type

F-tests:

	2	3	4	5
1	0.003	0.437	0.280	0.006
2	–	0.004	0.520	0.423
3	–	–	0.298	0.007
4	–	–	–	0.353

p-value of F tests

Comparison test(s)

Gaussian hypothesis: not a problem for large n, thanks to CLT

- $\bar{X} \to \mathcal{N}$ as $n \to \infty$
- $nS^2/\sigma^2 \to \chi^2_{n-1} \to \mathcal{N}$ as $n \to \infty$
- $t_{n-1} \to \mathcal{N}$ as $n \to \infty$

For moderate n, (say $n \leq 30$), the order is important:

1. Test for equal variance first;
2. If not rejected, test means

Otherwise use Welch's t test