Hypothesis testing and confidence intervals

For the maize data:

- the 95% (equal tail) confidence interval for $\mu_X - \mu_Y$ is $(3.34, 38.53)$ (see Sheet 2, Question 5)
- when testing $\mu_x = \mu_Y$ against $\mu_x \neq \mu_Y$, the p-value is 0.021.

So, observe that

(i) the p-value less than 0.05

(ii) the 95% confidence interval does not contain 0 (= the value of $\mu_X - \mu_Y$ under H_0).

(i) and (ii) both being true is not a coincidence – there is a connection between hypothesis tests and confidence intervals.

3.3 Hypothesis testing and confidence intervals

Example $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, μ, σ^2 unknown.

(i) A $1-\alpha$ C.I. for μ is

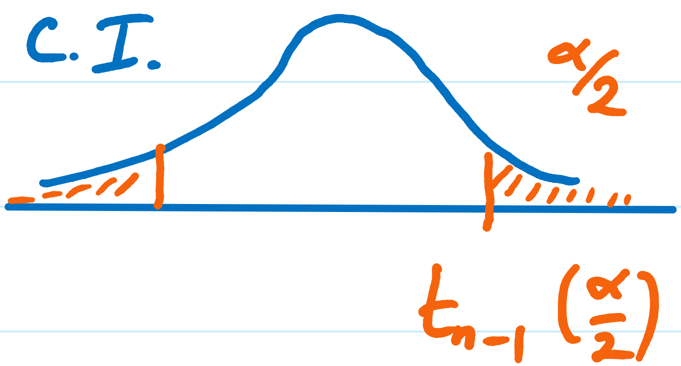
$$\left(\bar{x} \pm t_{n-1}\left(\frac{\alpha}{2}\right) \cdot \frac{s}{\sqrt{n}} \right) \qquad (1)$$

(ii) For t-test of $\mu = \mu_0$ against $\mu \neq \mu_0$,

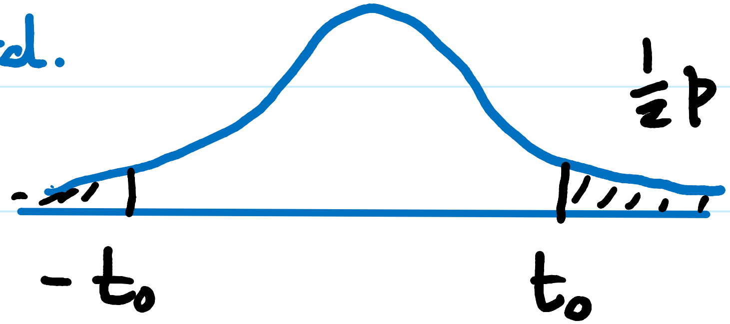
p-value is $p = P(|t_{n-1}| \geq t_0)$

where $t_0 = |t(\underline{x})| = \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right|$.

For C.I.



p-val.



So $p < \alpha \iff t_0 > t_{n-1}\left(\frac{\alpha}{2}\right)$

$$\iff t(\underline{x}) > t_{n-1}\left(\tfrac{\alpha}{2}\right) \quad \text{or} \quad t(\underline{x}) < -t_{n-1}\left(\tfrac{\alpha}{2}\right)$$

$$\iff \mu_0 < \bar{x} - t_{n-1}\left(\tfrac{\alpha}{2}\right)\frac{s}{\sqrt{n}} \quad \text{or}$$

$$\mu_0 > \bar{x} + t_{n-1}\left(\tfrac{\alpha}{2}\right)\frac{s}{\sqrt{n}}$$

That is: $p < \alpha \iff$ C.I. ① does not contain μ_0.

3.4 Hypothesis testing general setup

Let $X_1, \dots, X_n$ be iid from $f(x;\theta)$ where $\theta \in \Theta$ is a vector or scalar parameter.

Consider testing:
- the null hypothesis $H_0: \theta \in \Theta_0$
- against the alternative hypothesis $H_1: \theta \in \Theta_1$

where $\Theta_0 \cap \Theta_1 = \emptyset$ and possibly but not necessarily $\Theta_0 \cup \Theta_1 = \Theta$.

Suppose we can construct a test statistic $t(\underset{\sim}{X})$ such that large values of $t(\underset{\sim}{X})$ indicate a departure from H_0 in the direction of H_1.

Let $t_{obs} = t(\underset{\sim}{x})$, the value of $t(\underset{\sim}{X})$ observed.

Then the p-value or significance level is

$$p = P\left(t(\underset{\sim}{X}) \geq t_{obs} \mid H_0\right).$$

A small p is an indicator that H_0 and the data are inconsistent.

Warning: The p-value is NOT the probability that H_0 is true.

Rather: assuming H_0 is true, it is the probability of $t(\underline{X})$ taking a value at least as extreme as the value t_{obs} that we actually observed.

A hypothesis which completely determines f is called **simple**, e.g. $\theta = \theta_0$.

Otherwise a hypothesis is called **composite**, e.g. $\theta > \theta_0$ or $\theta \neq \theta_0$.

Example $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, μ, σ^2 unknown.

$H_0 : \mu = \mu_0$ is composite because it corresponds to $\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0, \ \sigma^2 > 0\}$ ← this set contains more than one point

Here σ^2 is called a **nuisance parameter**.

Suppose we want to make a definite decision:

either reject H_0

or don't reject H_0.

Then we can define a test in terms of a critical region $C \subset \mathbb{R}^n$:

- if $\underline{x} \in C$ then we reject H_0
- if $\underline{x} \notin C$ then we don't reject H_0

Errors in hypothesis testing

There are two possible types of error:

	don't reject H_0	reject H_0
H_0 true	✓	type I error
H_0 false	type II error	✓

✓ = correct decision

Consider simple $H_0: \theta = \theta_0$ versus simple $H_1: \theta = \theta_1$.

The <u>type I error probability</u> α, also called the <u>size</u> of the test, is defined by

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

$$= P(\underset{\sim}{X} \in C \mid \theta_0)$$

The <u>type II error probability</u> β is defined by

$$\beta = P(\text{don't reject } H_0 \mid H_1 \text{ true})$$

$$= P(\underset{\sim}{X} \notin C \mid \theta_1)$$

$1-\beta = P(\text{reject } H_0 \mid H_1 \text{ true})$ is called the **power** of the test.

Note: $\text{power} = 1-\beta = P(\underset{\sim}{X} \in C \mid \theta_1)$

$= $ probability of correctly detecting that H_0 is false.

If H_0 is composite, $H_0 : \theta \in \Theta_0$ say, then the size is defined by

$$\alpha = \sup_{\theta \in \Theta_0} P(\underset{\sim}{X} \in C \mid \theta)$$

If H_1 is composite then we have to define the power as a function of θ: the power function $w(\theta)$ is defined by

$$w(\theta) = P(\text{reject } H_0 \mid \theta \text{ is the true value})$$

$$= P(\underset{\sim}{X} \in C \mid \theta)$$

Ideally we'd like

$w(\theta)$ to be near 1 for H_1-values of θ

- - - - - - - - - 0 for H_0-values of θ.

3.5 The Neyman–Pearson Lemma

Consider testing simple $H_0: \theta = \theta_0$ against simple $H_1: \theta = \theta_1$. $(*)$

Suppose we choose a small type I error probability α (e.g. $\alpha = 0.05$). Then, among all tests of this size we could aim to:

$$\begin{cases} \text{minimise the type II error probability } \beta \\ \text{i.e. maximise the power } 1-\beta \end{cases}$$

This approach treats H_0 and H_1 asymmetrically.

Theorem 3.1 (N-P Lemma) Let $L(\theta;\underset{\sim}{x})$ be the likelihood. Define the critical region C by

$$C = \left\{ \underset{\sim}{x} : \frac{L(\theta_0;\underset{\sim}{x})}{L(\theta_1;\underset{\sim}{x})} \leq k \right\}$$

and suppose constants k and α are such that

$P(\underset{\sim}{X} \in C \mid H_0) = \alpha.$ ← "C has size α"

Then among all tests of (*) of size $\leq \alpha$, the test with critical region C has maximum power.

Proof (for cts random variables – for discrete replace $\int$ by Σ)

Consider any test of size $\leq \alpha$, with critical region A say.

Then $P(\underline{X} \in A \mid H_0) \leq \alpha$ ①.

(C is one possibility for A).

Define $\phi_A(\underline{x}) = \begin{cases} 1 & \text{if } \underline{x} \in A \\ 0 & \text{otherwise} \end{cases}$

and let C and k be as in statement of theorem.

Then $\quad 0 \leq \{\phi_C(\underline{x}) - \phi_A(\underline{x})\}.\left[L(\theta_1; \underline{x}) - \frac{1}{k} L(\theta_0; \underline{x})\right]$

since $\{\ldots\}$ and $[\ldots]$ are both ≥ 0 if $\underline{x} \in C$

and both ≤ 0 if $\underline{x} \notin C$

So $$0 \leq \int_{\mathbb{R}^n} \{\phi_C(\underline{x}) - \phi_A(\underline{x})\}\left[L(\theta_1;\underline{x}) - \frac{1}{k}L(\theta_0;\underline{x})\right] d\underline{x}$$

$$= P(\underline{X}\in C \mid H_1) - P(\underline{X}\in A \mid H_1) - \frac{1}{k}\Big[\underbrace{P(\underline{X}\in C \mid H_0)}_{\alpha} - \underbrace{P(\underline{X}\in A \mid H_0)}_{\leq \alpha \text{ by } (1)}\Big]$$

$$\leq P(\underline{X}\in C \mid H_1) - P(\underline{X}\in A \mid H_1).$$

That is, $P(\underline{X}\in C \mid H_1) \geq P(\underline{X}\in A \mid H_1)$ (2)

power of region C power of region A

And (2) says the power is maximised by using region C. $\square$

Example $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma_0^2)$, σ_0^2 known.

Find most powerful test of $H_0 : \mu = 0$ against $H_1 : \mu = \mu_1$, where $\mu_1 > 0$.

Likelihood $L(\mu; \underline{x}) = (2\pi\sigma_0^2)^{-n/2} \exp\left[-\frac{1}{2\sigma_0^2}\sum(x_i - \mu)^2\right]$

Step 1 H_0, H_1 both simple, so N-P applies and most powerful test is of the form

$$\text{reject } H_0 \iff \frac{L(0; \underline{x})}{L(\mu_1; \underline{x})} \le k_1$$

k_1 a constant, i.e. doesn't depend on $\underline{x}$.

$$\iff \exp\left[-\frac{1}{2\sigma_0^2}\sum x_i^2\right]\exp\left[\frac{1}{2\sigma_0^2}\sum(x_i-\mu_1)^2\right] \leq k_1$$

$$\iff \exp\left[\frac{1}{2\sigma_0^2}\left(-\sum x_i^2 + \sum x_i^2 - 2\mu_1\sum x_i + n\mu_1^2\right)\right] \leq k_1$$

$$\iff \frac{1}{2\sigma_0^2}\left(-2\mu_1 n\bar{x} + n\mu_1^2\right) \leq k_2 \qquad (k_2 = \log k_1)$$

$$\iff -\mu_1\bar{x} \leq k_3$$

$$\iff \bar{x} \geq c$$

where k_1, k_2, k_3, c are constants that don't depend on $\underline{x}$ (they can depend on $n, \sigma_0^2, \ldots$).

Step 2 Choose c so that the test has size α.

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

$$= P(\bar{X} \geq c \mid H_0) \qquad \text{and under } H_0,\ \bar{X} \sim N\left(0, \frac{\sigma_0^2}{n}\right) \quad (3)$$

$$= P\left(\frac{\bar{X}}{\sigma_0/\sqrt{n}} \geq \frac{c}{\sigma_0/\sqrt{n}} \,\middle|\, H_0\right)$$

$$= P\left(N(0,1) \geq \frac{c}{\sigma_0/\sqrt{n}}\right) \qquad \text{by } (3)$$

Hence $\frac{c}{\sigma_0/\sqrt{n}} = z_\alpha$. So most powerful critical region is $C = \left\{\underset{\sim}{x} : \bar{x} \geq z_\alpha \frac{\sigma_0}{\sqrt{n}}\right\}$.

Let's also calculate the power function of this test.

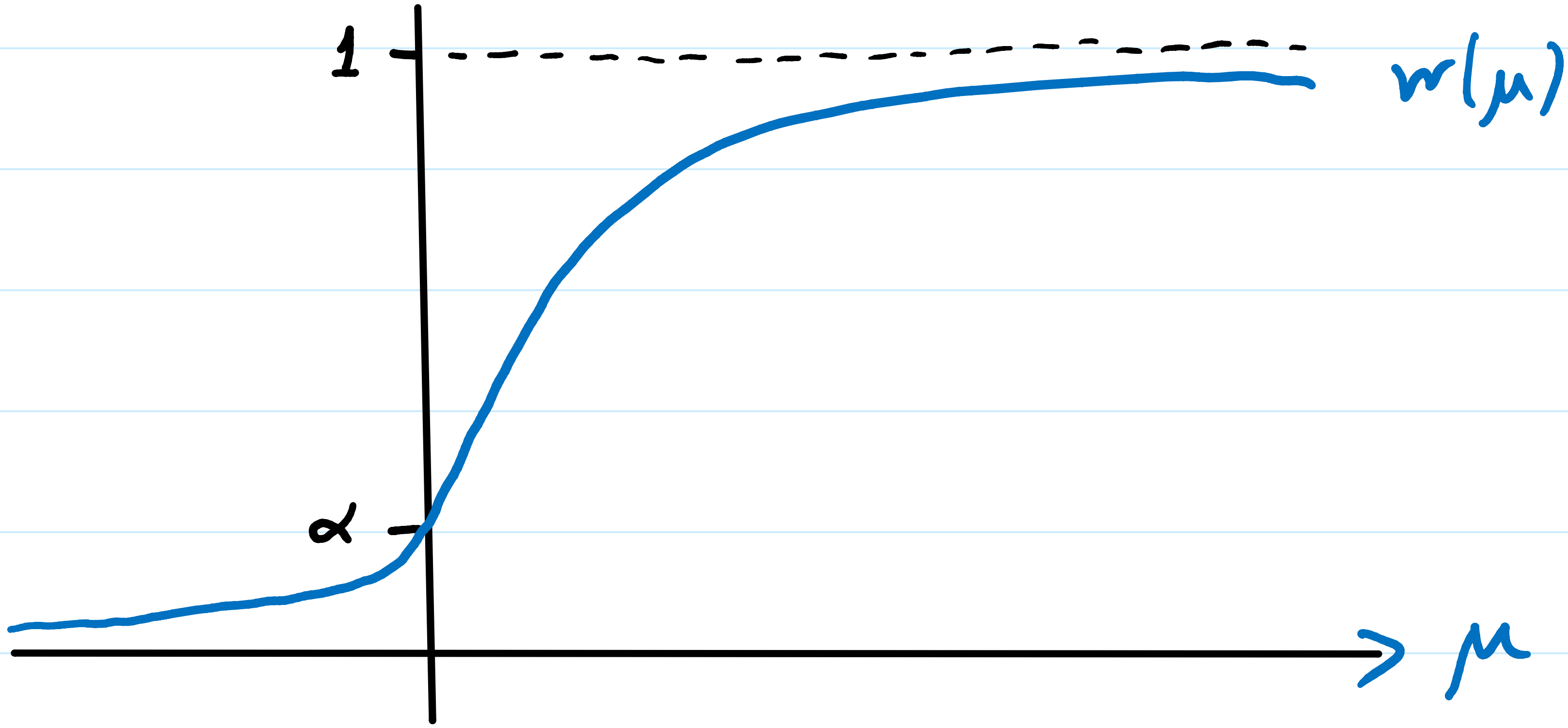
$$W(\mu) = P(\text{reject } H_0 \mid \mu \text{ is the true value})$$

$$= P\left(\bar{X} \geq z_\alpha \frac{\sigma_0}{\sqrt{n}} \;\middle|\; \mu\right) \qquad \text{if } \mu \text{ is true value, } \bar{X} \sim N\left(\mu, \frac{\sigma_0^2}{n}\right) \; (4)$$

$$= P\left(\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \geq z_\alpha - \frac{\mu}{\sigma_0/\sqrt{n}} \;\middle|\; \mu\right)$$

$$= P\left(N(0,1) \geq z_\alpha - \frac{\mu}{\sigma_0/\sqrt{n}}\right) \qquad \text{by } (4)$$

$$= 1 - \Phi\left(z_\alpha - \frac{\mu}{\sigma_0/\sqrt{n}}\right)$$



1
α
π(μ)
μ

<u>Last example</u>: $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma_0^2)$, σ_0^2 known.

We were testing $H_0 : \mu = 0$ against $H_1 : \mu = \mu_1$,

where μ_1 was a single value satisfying $\mu_1 > 0$.

Critical region was $\bar{x} \geqslant c$, or $\sum x_i \geqslant k$ (where $k = nc$)

Equation linking k and α was $\alpha = P(\sum X_i \geqslant k \mid H_0)$.

$\sum X_i$ was normal, so any value of α possible by choosing k appropriately.

If e.g. the $X_i \sim$ Poisson, then not all values of α possible as $P(\sum X_i \geqslant k \mid H_0)$ will decrease in jumps as k increases.

3.6 Uniformly most powerful tests

Consider $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Theta_1$.

When testing simple $\theta = \theta_0$ against simple $\theta = \theta_1 \in \Theta_1$, the critical region from N-P lemma may be the same for each $\theta_1 \in \Theta_1$. Then C is said to be <u>uniformly most powerful</u> (UMP) for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1$.

Previous example: $N(\mu, \sigma^2)$, σ_0^2 known.

The critical region C we found for $\mu=0$ versus $\mu=\mu_1$ was the same for all $\mu_1>0$.

Hence our C is UMP for testing $\mu=0$ against $\mu>0$.

$$C = \left\{ \underset{\sim}{x} : \bar{x} \geqslant z_\alpha \frac{\sigma_0}{\sqrt{n}} \right\}$$

Insect traps

33 insect traps were set out across sand dunes and the numbers of insects caught in a fixed time were counted (Gilchrist, 1984). The number of traps containing various numbers of the taxa *Staphylinoidea* were as follows.

Count	0	1	2	3	4	5	6	$\geqslant 7$
Frequency	10	9	5	5	1	2	1	0

Suppose $X_1, \ldots, X_{33} \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$.

Consider testing $H_0 : \lambda = 1$ against $H_1 : \lambda = \lambda_1$, where $\lambda_1 > 1$.

The NP lemma leads to a test of the form

$$\text{reject } H_0 \iff \sum x_i \geqslant c.$$

If the test has size α, then $\alpha = P(\sum X_i \geqslant c \mid H_0)$.

Under H_0, we have $\sum X_i \sim \text{Poisson}(33)$ exactly. However, instead of using this we can use a normal approximation:

$$\alpha = P\left(\frac{\sum X_i - 33}{\sqrt{33}} \geqslant \frac{c - 33}{\sqrt{33}} \,\middle|\, H_0 \right)$$

and, by the CLT, if H_0 is true then $\frac{\sum X_i - 33}{\sqrt{33}} \overset{\text{D}}{\approx} N(0,1)$, so

$$\alpha \approx 1 - \Phi\left(\frac{c - 33}{\sqrt{33}} \right).$$

Hence $\frac{c-33}{\sqrt{33}} \approx z_\alpha$, so $c \approx 33 + z_\alpha \sqrt{33}$.

So we have a critical region

$$C = \{\mathsf{x} : \sum x_i \geqslant 33 + z_\alpha \sqrt{33}\}.$$

Note that C does not depend on which value of $\lambda_1 > 1$ we are considering, so we actually have a UMP test of $\lambda = 1$ against $\lambda > 1$.

If $\alpha = 0.01$ then $c \approx 47$; if $\alpha = 0.001$ then $c \approx 51$.

The observed value of $\sum x_i$ is 54.

So in both cases the observed value of 54 is $\geqslant c$, so in both cases we'd reject H_0.

An alternative way of thinking about this is to calculate the p-value:

$$
\begin{aligned}
p &= P(\text{we observe a value at least as extreme as } 54 \mid H_0) \\
&= P\big(\textstyle\sum X_i \geqslant 54 \mid H_0\big) \\
&\approx 0.0005
\end{aligned}
$$

which is very strong evidence for rejecting H_0.

Note that a test of size α rejects H_0 if and only if $\alpha \geqslant p$. That is, the p-value is the smallest value of α for which H_0 would be rejected. (This is true generally, not just in this particular example.)

In practice, no-one tells us a value of α, we have to judge the situation for ourselves. Our conclusion here is that there is very strong evidence for rejecting H_0.

3.6 Likelihood ratio tests

Now consider testing $H_0 : \theta \in \Theta_0$ against the general alternative $H_1 : \theta \in \Theta$ (where $\Theta_0 \subset \Theta$).

So now H_0 is a special case of H_1.

H_0 is "nested within" H_1.

We test to see if simplifying to the H_0-model is reasonable.

The likelihood ratio $\lambda(\underline{x})$ is defined by

$$\lambda(\underline{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta; \underline{x})}{\sup_{\theta \in \Theta} L(\theta; \underline{x})} = \frac{\text{TOP}}{\text{BOTTOM}} \quad (1)$$

A (generalised) likelihood ratio test (LRT) has critical region of the form

$$C = \{ \underline{x} : \lambda(\underline{x}) \leq k \}.$$

Sometimes we can calculate the distribution of a function of $\lambda(\underline{X})$,

more often we will approximate the distribution of a function of $\lambda(\underline{X})$.

Example $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, μ, σ^2 unknown.

Let $H_0 : \mu = \mu_0$ (and any $\sigma^2 > 0$)

$H_1 : \mu \in (-\infty, \infty)$ (and any $\sigma^2 > 0$).

Likelihood $L(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2\right]$.

For TOP of ①: max L over σ^2 with $\mu = \mu_0$ fixed.

Max is at $\sigma^2 = \hat{\sigma}_0^2 = \frac{1}{n} \sum (x_i - \mu_0)^2$.

For BOTTOM of ①: max L over μ and σ^2.

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

Substitute into ① to get

$$\lambda(\underline{x}) = \frac{L(\mu_0, \hat{\sigma}_0^2)}{L(\hat{\mu}, \hat{\sigma}^2)} \longleftarrow (2\pi\hat{\sigma}^2)^{-n/2} \exp\left[-\frac{1}{2\hat{\sigma}^2}\sum(x_i-\hat{\mu})^2\right]$$

$-n/2$

$$= \frac{\left[\frac{2\pi}{n}\sum(x_i-\mu_0)^2\right]^{-n/2} e^{-n/2}}{\left[\frac{2\pi}{n}\sum(x_i-\bar{x})^2\right]^{-n/2} e^{-n/2}}$$

Now note $\sum(x_i-\mu_0)^2 = \sum(x_i-\bar{x})^2 + n(\bar{x}-\mu_0)^2$.

Substitute into $\lambda(x)$ to find

$$\lambda(\underline{x}) = \left[1 + \frac{n(\bar{x}-\mu_0)^2}{\sum(x_i-\bar{x})^2}\right]^{-n/2}$$

So LRT is reject $H_0 \iff \lambda(\underline{x}) \leq k$

$$\iff \left|\frac{\bar{x}-\mu_0}{s/\sqrt{n}}\right| \geq k_1.$$

This is the t-test, so take $k_1 = t_{n-1}(\alpha/2)$ for a test of size α. i.e. we know the exact distribution of a fn. of $\lambda(\underline{X})$.

Likelihood ratio statistic

$\Lambda(\underset{\sim}{X}) = -2\log\lambda(\underset{\sim}{X})$ is called the <u>likelihood ratio statistic</u>.

The critical region $\{\underset{\sim}{x} : \lambda(\underset{\sim}{x}) \leq k\}$ becomes $\{\underset{\sim}{x} : \Lambda(\underset{\sim}{x}) \geq c\}$.

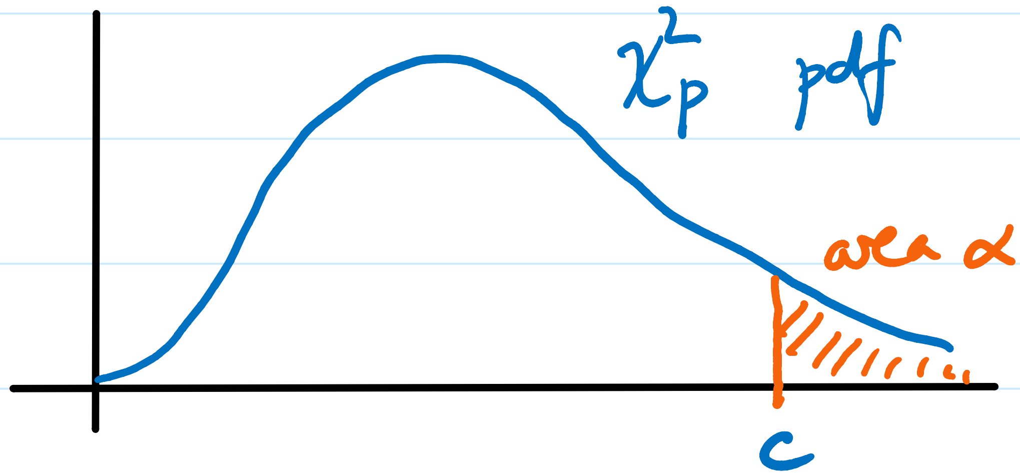
If H_0 is true then, under regularity conditions, as $n \to \infty$, we have $\Lambda(\underset{\sim}{X}) \xrightarrow{D} \chi^2_p$ ⊛

where $p = \dim H_1 - \dim H_0$.

$\dim H_1 = \#$ independent parameters in Θ

$\dim H_0 = \text{-- -- -- -- -- -- --}\ \Theta_0.$

Since $\Lambda(\underline{X}) \approx \chi^2_p$ for large n, under H_0, we get an approx test of size α by choosing c such that $P(\chi^2_p > c) = \alpha$.



Why is 2 true?

Sketch proof for scalar θ, so $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Theta$ with $\dim \Theta = 1$.

So here $p = \dim \Theta - \dim \Theta_0 = 1 - 0 = 1$.

Taylor expansion: $\ell(\theta_0) \approx \ell(\hat{\theta}) + (\hat{\theta} - \theta_0)\ell'(\hat{\theta})$
$$+ \frac{1}{2}(\hat{\theta} - \theta_0)^2 \ell''(\hat{\theta})$$

$$= \ell(\hat{\theta}) - \frac{1}{2}(\hat{\theta} - \theta_0)^2 J(\hat{\theta}) \quad 3$$

assuming $\ell'(\hat{\theta}) = 0$.

So $\Lambda(\underset{\sim}{X}) = -2\log\left(\dfrac{L(\theta_0)}{L(\hat{\theta})}\right)$

$$= 2\left[l(\hat{\theta}) - l(\theta_0)\right]$$

$$\approx \underbrace{(\hat{\theta}-\theta_0)^2 I(\theta_0)}_{\approx [N(0,1)]^2} \cdot \underbrace{\frac{J(\hat{\theta})}{I(\theta_0)}}_{\approx 1}$$

using 3

under H_0, for large n

$$\approx \chi^2_1.$$

We now write the LR statistic as

$$\Lambda = -2 \log \lambda = -2 \log \left(\frac{\sup_{H_0} L}{\sup_{H_1} L} \right) \qquad (1)$$

Goodness of fit tests

Hardy–Weinberg equilibrium

In a sample from the Chinese population of Hong Kong, blood types occurred with the following frequencies (Rice, 1995):

	Blood type			
	M	MN	N	Total
Frequency	342	500	187	1029

If gene frequencies are in Hardy–Weinberg equilibrium, then the probability of an individual having blood type M, MN, or N should be

$$\begin{aligned} P(M) &= (1-\theta)^2 \\ P(MN) &= 2\theta(1-\theta) \\ P(N) &= \theta^2. \end{aligned}$$

Consider n independent observations, each in one of categories $1, \ldots, k$.

Let n_i = # observations in category i (frequency),

so $\sum_{i=1}^{k} n_i = n$

π_i = probability of an observation being in category i, so $\sum_{i=1}^{k} \pi_i = 1$.

Let $\pi = (\pi_1, \ldots, \pi_k)$

Likelihood $L(\pi) = \dfrac{n!}{n_1! \dots n_k!} \pi_1^{n_1} \cdots \pi_k^{n_k}$ multinomial distribution

Log-lik $\ell(\pi) = \sum n_i \log \pi_i + \text{constant}$

Consider $H_0: \pi_i = \pi_i(\theta)$ for $i = 1, \dots, k$, where $\theta \in \Theta$

(e.g. $\pi_1 = (1-\theta)^2$, $\pi_2 = 2\theta(1-\theta)$, $\pi_3 = \theta^2$, $\theta \in (0,1)$)

versus $H_1: \pi_i$ unrestricted except for $\sum \pi_i = 1$.

Then $\dim H_1 = k-1$,
and suppose $\dim H_0 = q < k-1$.

$$\Lambda = -2\log\left(\frac{\sup_{H_0} L}{\sup_{H_1} L}\right) \qquad (1)$$

The degrees of freedom for Λ are:

$$p = \dim H_1 - \dim H_0 = (k-1) - q.$$

(i) For TOP in (1): maximise over θ to get MLE $\theta = \hat{\theta}$

(ii) For BOTTOM in (1): maximise $f(\pi) = \sum n_i \log \pi_i$

subject to the constraint $g(\pi) = \sum \pi_i - 1 = 0.$

With Lagrange multiplier λ, we need

$$\frac{\partial f}{\partial \pi_i} = \lambda \frac{\partial g}{\partial \pi_i} \qquad i = 1 \dots k$$

i.e. $$\frac{n_i}{\pi_i} = \lambda . 1$$

So $\pi_i = \dfrac{n_i}{\lambda}$ and then $1 = \sum \pi_i = \dfrac{\sum n_i}{\lambda} = \dfrac{n}{\lambda}$

and so $\lambda = n$.

So the MLEs under H_1 are $\hat{\pi}_i = \dfrac{n_i}{n}$.

So $\quad \Lambda = -2 \log\left(\dfrac{L(\pi(\hat{\theta}))}{L(\hat{\pi})}\right)$

$$= 2\left[l(\hat{\pi}) - l(\pi(\hat{\theta}))\right]$$

$$= 2\left[\sum n_i \log \hat{\pi}_i - \sum n_i \log \pi_i(\hat{\theta})\right]$$

$$= 2 \sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n\,\pi_i(\hat{\theta})}\right) \qquad \text{since } \hat{\pi}_i = \frac{n_i}{n}.$$

Compare this Λ to a χ^2_p where $p = k-1-q$ to carry out the test.

Pearson's chi-squared statistic

$$\Lambda = 2\sum_{i=1}^{k} o_i \log\left(\frac{o_i}{e_i}\right)$$

where $o_i = n_i$ observed

$e_i = n.\pi_i(\hat{\theta})$ expected under H_0

Using $x \log \frac{x}{a} \approx x - a + \frac{(x-a)^2}{2a}$ gives

$$\Lambda \approx 2\sum\left[o_i - e_i + \frac{(o_i - e_i)^2}{2e_i}\right]$$

$$= \sum \frac{(o_i - e_i)^2}{e_i} = P$$ Pearson's χ^2 statistic

Hardy–Weinberg equilibrium

In a sample from the Chinese population of Hong Kong, blood types occurred with the following frequencies (Rice, 1995):

	Blood type			
	M	MN	N	Total
Frequency	342	500	187	1029

If gene frequencies are in Hardy–Weinberg equilibrium, then the probability of an individual having blood type M, MN, or N should be

$$
\begin{aligned}
P(M) &= (1-\theta)^2 \\
P(MN) &= 2\theta(1-\theta) \\
P(N) &= \theta^2.
\end{aligned}
$$

The observed frequencies are $(n_1, n_2, n_3) = (342, 500, 187)$, with total $n = n_1 + n_2 + n_3 = 1029$.

The likelihood is

$$L(\theta) \propto [(1-\theta)^2]^{n_1} \times [\theta(1-\theta)]^{n_2} \times [\theta^2]^{n_3}$$

so the log-likelihood is

$$\ell(\theta) = (2n_1 + n_2)\log(1-\theta) + (n_2 + 2n_3)\log\theta + \text{constant}$$

from which we obtain

$$\widehat{\theta} = \frac{n_2 + 2n_3}{2n} = 0.425.$$

So $\pi_1(\widehat{\theta}) = (1-\widehat{\theta})^2$, $\pi_2(\widehat{\theta}) = 2\widehat{\theta}(1-\widehat{\theta})$, $\pi_3(\widehat{\theta}) = \widehat{\theta}^2$ and

$$\Lambda = 2\sum_i n_i \log\left(\frac{n_i}{n\pi_i(\widehat{\theta})}\right) = 0.032.$$

We compare Λ to a χ_p^2 where $p = \dim\Theta - \dim\Theta_0 = (3-1)-1 = 1$.

The value $\Lambda = 0.032$ is much less than $E(\chi_1^2) = 1$. The p-value is $P(\chi_1^2 \geqslant 0.032) = 0.86$, so there is no reason to doubt the Hardy–Weinberg model.

Pearson's chi-squared statistic leads to the same conclusion

$$P = \sum \frac{[n_i - n\pi_i(\widehat{\theta})]^2}{n\pi_i(\widehat{\theta})} = 0.0319.$$

Insect counts (Bliss and Fisher, 1953)

[Example from Rice (1995).] From each of 6 apple trees in an orchard that had been sprayed, 25 leaves were selected. On each of the leaves, the number of adult female red mites was counted.

Number per leaf	0	1	2	3	4	5	6	7	8+
Observed frequency	70	38	17	10	9	3	2	1	0

Does a Poisson(θ) model fit these data?

As usual for a Poisson, $\widehat{\theta} = \overline{x} = 1.147$, and

$$\pi_i(\widehat{\theta}) = \widehat{\theta}^i e^{-\widehat{\theta}}/i!, \quad i = 0, 1, \ldots, 7$$

$$\pi_8(\widehat{\theta}) = 1 - \sum_{i=0}^{7} \pi_i(\widehat{\theta}).$$

The expected frequency in cell i is $n\pi_i(\widehat{\theta})$.

Some expected frequencies are very small:

# per leaf	0	1	2	3	4	5	6	7	8+
Observed	70	38	17	10	9	3	2	1	0
Expected	47.7	54.6	31.3	12.0	3.4	0.8	0.2	0.02	0.004

The χ^2 approximation for the distribution of Λ applies when there are large counts.

The usual rule-of-thumb is that the χ^2 approximation is good when the expected frequency in each cell is at least 5.

To ensure this, we should pool some cells before calculating Λ or P.

After pooling cells $\geqslant 3$:

# per leaf	0	1	2	$\geqslant 3$
Observed	70	38	17	25
Expected	47.7	54.6	31.3	16.4

Then $\Lambda = 2 \sum O_i \log\left(\frac{O_i}{E_i}\right) = 26.60$, and $P = \sum (O_i - E_i)^2 / E_i = 26.65$.

These are to be compared with a χ^2 with $(4-1) - 1 = 2$ degrees of freedom.

The p-value is $p = P(\chi_2^2 \geqslant 26.6) \approx 10^{-6}$, so there is clear evidence that a Poisson model is not suitable.

Two-way contingency tables

Hair and Eye Colour

The hair and eye colour of 592 statistics students at the University of Delaware were recorded (Snee, 1974) – dataset `HairEyeColor` in R.

	Eye colour			
Hair colour	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

Are hair colour and eye colour independent?

Cross-classify n individuals according to two sets of categories

(eye colour)

		1	2	$\cdots$	c	row sum
(hair colour)	1	n_{11}	$\cdots$	$\cdots$	n_{1c}	n_{1+}
	$\vdots$	$\vdots$			$\vdots$	$\vdots$
	r	n_{r1}	$\cdots$	$\cdots$	n_{rc}	n_{r+}
	col sum	n_{+1}		$\cdots$	n_{+c}	$\boxed{n_{++} = n}$

Let n_{ij} = frequency of (i, j) $\qquad \sum_i \sum_j n_{ij} = n$

π_{ij} = probability an individual falls into cell (i,j) $\qquad \sum_i \sum_j \pi_{ij} = 1$

Likelihood $\quad L(\pi) = n! \prod_{i=1}^{r} \prod_{j=1}^{c} \dfrac{\pi_{ij}^{n_{ij}}}{n_{ij}!}$

Log-lik $\quad \ell(\pi) = \sum_i \sum_j n_{ij} \log \pi_{ij} + \text{constant}$

Consider:

H_0: the two classifications are independent

(e.g. hair colour and eye colour are independent)

i.e. $\pi_{ij} = \alpha_i \beta_j$ where $\sum_i \alpha_i = 1$ and $\sum_j \beta_j = 1$

H_1: π_{ij} unrestricted except for $\sum_i \sum_j \pi_{ij} = 1$.

(i) Max under H_0 (Sheet 3): $\hat{\alpha}_i = \frac{n_{i+}}{n}$, $\hat{\beta}_j = \frac{n_{+j}}{n}$

(ii) Max under H_1 (done already): $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$.

We find $$\Lambda = 2 \sum_{i,j} n_{ij} \log\left(\frac{n_{ij} \cdot n}{n_{i+} \cdot n_{+j}}\right)$$

$$\approx \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where $o_{ij} = n_{ij}$ observed

$e_{ij} = n \hat{\alpha}_i \hat{\beta}_j$ expected # in (i,j) under H_0

Degrees of freedom of this Λ

$\dim H_1 = rc - 1$ probabilities $\pi_{11}, \dots, \pi_{rc}$ with $\sum_{i,j} \pi_{ij} = 1$.

$\dim H_0 = (r-1) + (c-1)$

$r-1$ for $\alpha_1 \dots \alpha_r$ with $\sum \alpha_i = 1$

$c-1$ for $\beta_1 \dots \beta_c$ with $\sum \beta_j = 1$

So $p = \dim H_1 - \dim H_0 = (r-1)(c-1)$

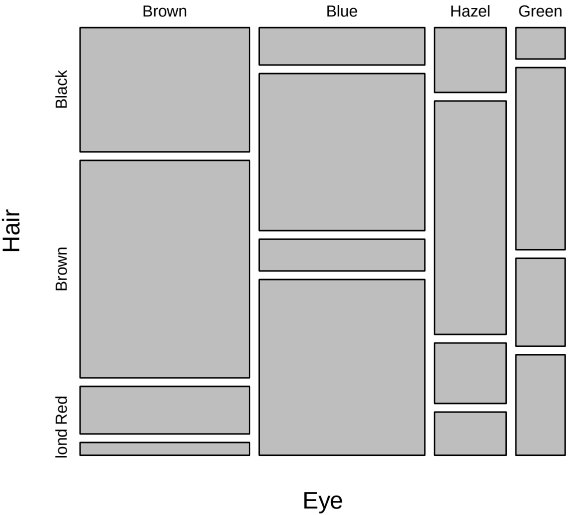
Hair and Eye Colour

The hair and eye colour of 592 statistics students at the University of Delaware were recorded (Snee, 1974) – dataset `HairEyeColor` in R.

	Eye colour			
Hair colour	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

Are hair colour and eye colour independent?

Relation between hair and eye colour



$$
\begin{aligned}
\Lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \log\left(\frac{n_{ij}\, n}{n_{i+}\, n_{+j}}\right) &= 146.4 \\
\dim H_1 &= 16 - 1 = 15 \\
\dim H_0 &= (4-1) + (4-1) = 6
\end{aligned}
$$

Hence we compare Λ to a χ_p^2 where $p = 15 - 6 = 9$.

The p-value is $P(\chi_9^2 \geqslant 146.4) \approx 0$.

So there is overwhelming evidence of an association between hair colour and eye colour (i.e. overwhelming evidence that they are not independent).

[Pearson's chi-squared statistic is $P = 138.3$.]