18.650 – Fundamentals of Statistics

4. Parametric hypothesis testing

Waiting time in the ER

- The average waiting time in the Emergency Room (ER) in the US is 30 minutes according to the CDC
- Some patients claim that the new Princeton-Plainsboro hospital has a longer waiting time. Is it true?
- Collect a sample: $X_1, \dots, X_n$ (waiting time in minutes for n random patients) with unknown expected value $\mathbb{E}[X_1] = \mu$.
- We want to know if $\mu > 30$.

$$H_0: \quad \mu \leq 30$$
$$H_1: \quad \mu > 30$$

© 2018 ealth ap, All rights reserved

Statistical formulation

- Consider a sample $X_1, \dots, X_n$ of i.i.d. random variables and a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$.
- Let Θ_0 and Θ_1 be a *partition* of Θ.
- Consider the two hypotheses: $\begin{cases} H_0 : & \theta \in \Theta_0 \\ H_1 : & \theta \in \Theta_1 \end{cases}$
- H_0 is the *null hypothesis*, H_1 is the *alternative hypothesis*.
- We say that we *test* H_0 *against* H_1.

Testing lexicon

- For $k = 0$ (H_0) or $k = 1$ (H_1), we say that
  - Θ_k is a *simple hypothesis* if $\Theta_k = \{\theta_k\}$
  - Θ_k is a *composite hypothesis* if Θ_k is of the following three forms

$$\Theta_k = \{\theta : \ \theta > \theta_k\} \quad \Theta_k = \{\theta : \ \theta < \theta_k\} \quad \Theta_k = \{\theta : \ \theta \neq \theta_k\}$$

- A test is typically either *one-sided* or *two-sided*

Two-sided

$$\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta \neq \theta_0 \end{cases}$$

One-sided

$$\begin{cases} H_0 : & \theta \leq \theta_0 \\ H_1 : & \theta > \theta_0 \end{cases} \quad \text{or} \quad \begin{cases} H_0 : & \theta \geq \theta_0 \\ H_1 : & \theta < \theta_0 \end{cases}$$

Examples

1. Waiting time in the ER

$$\begin{aligned} H_0 &: \quad \mu \leq 30 \\ H_1 &: \quad \mu > 30 \end{aligned}$$

   Both hypotheses are composite. The test is one-sided

2. In the Kiss example, we want to test

$$\begin{aligned} H_0 &: \quad p = .5 \\ H_1 &: \quad p \neq .5 \end{aligned}$$

   H_0 is simple, H_1 is composite hypotheses. The test is two-sided

Clinical trials

- Pharmaceutical companies use hypothesis testing to test if a new drug is efficient.
- To do so, they administer a drug to a group of patients (test group) and a placebo to another group (control group).
- We consider testing a drug that is supposed to lower LDL (low-density lipoprotein), a.k.a "bad cholesterol" among patients with a high level of LDL (above 200 mg/dL)

Notation and modelling

- Let $\mu_{\mathrm{d}} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the drug.
- Let $\mu_{\mathrm{c}} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the placebo.
- Hypothesis testing problem:

$$\begin{aligned} H_0 : &\quad \mu_{\mathrm{d}} \leq \mu_{\mathrm{c}} \\ H_1 : &\quad \mu_{\mathrm{d}} > \mu_{\mathrm{c}} \end{aligned}$$

- We observe two independent samples:
  - $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu_{\mathrm{d}}, \sigma_{\mathrm{d}}^2)$ from the test group and
  - $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathcal{N}(\mu_{\mathrm{c}}, \sigma_{\mathrm{c}}^2)$ from the control group.
- This is a *two-sample* test: these are very common (A/B testing).

Asymmetry in the hypotheses

- We want to decide whether to *reject* H_0 (look for evidence against H_0 in the data).

- H_0 and H_1 do not play a symmetric role: the data is only used to try to disprove H_0

$$\begin{aligned} H_0 &: \quad \text{status quo} \\ H_1 &: \quad \text{a (scientific) discovery} \end{aligned}$$

- In particular lack of evidence, does not mean that H_0 is true ("innocent until proven guilty")

Examples

1. Waiting time in the ER

$$\begin{aligned} H_0 &: \quad \mu \leq 30 \\ H_1 &: \quad \mu > 30 \end{aligned}$$

   Status quo: CDC statement. We collect data to show that Princeton-Plainsboro is different

2. Kiss

$$\begin{aligned} H_0 &: \quad p = .5 \\ H_1 &: \quad p \neq .5 \end{aligned}$$

   Status quo: our intuition tells us there should be no preference. We collect data to show that there is one.

3. Clinical trials

$$\begin{aligned} H_0 &: \quad \mu_{\mathrm{d}} \leq \mu_{\mathrm{c}} \\ H_1 &: \quad \mu_{\mathrm{d}} > \mu_{\mathrm{c}} \end{aligned}$$

   Status quo: The drug is not more effective than a placebo. We collect data to prove that the drug is effective.

What is a test?

- A *test* is a statistic $\psi \in \{0,1\}$ that does not depend on unknown quantities and such that:
    - If $\psi = 0$, H_0 is not rejected;
    - If $\psi = 1$, H_0 is rejected.

  Important remark: Can always write $\psi = \mathbb{1}\{R\}$, where R is an *event* called *rejection region*

- Waiting time in the ER:

$$\begin{aligned} H_0 :& \quad \mu \le 30 \\ H_1 :& \quad \mu > 30 \end{aligned} \qquad \psi = \mathbb{1}\{\bar{X}_n > C\}$$

- Kiss:

$$\begin{aligned} H_0 :& \quad p = .5 \\ H_1 :& \quad p \neq .5 \end{aligned} \qquad \psi = \mathbb{1}\{|\bar{X}_n - .5| > C\}$$

- Clinical trials

$$\begin{aligned} H_0 :& \quad \mu_{\mathrm{d}} \le \mu_{\mathrm{c}} \\ H_1 :& \quad \mu_{\mathrm{d}} > \mu_{\mathrm{c}} \end{aligned} \qquad \psi = \mathbb{1}\{\bar{X}_n - \bar{Y}_m > C\}$$

Errors

A test can make two types of errors:

	Fail to reject Null	Reject Null
H_0 true ($\theta \in \Theta_0$)		
H_1 true ($\theta \in \Theta_1$)		

Both errors can be computed from the *power function*

$$\beta(\theta) = \mathbb{P}_\theta[\psi = 1]$$

- If $\theta \in \Theta_0$,

  $$\beta(\theta) = \mathbb{P}_\theta[\psi \text{ makes an error of type I }]$$

  We want $\beta(\theta)$ to be small
- If $\theta \in \Theta_1$,

  $$\beta(\theta) = 1 - \mathbb{P}_\theta[\psi \text{ makes an error of type II }]$$

  We want $\beta(\theta)$ to be large

The Neyman-Pearson paradigm

Recall the waiting time in the ER example

$$\begin{aligned} H_0 &: \quad \mu \le 30 \\ H_1 &: \quad \mu > 30 \end{aligned} \qquad \psi = \mathbb{1}\{\bar{X}_n > C\}$$

How to choose C ?

We are facing a dilemma: both errors should be small!

- To make Type I error → 0, take $C \to +\infty$
- To make Type II error → 0, take $C \to -\infty$

Cannot make both small at the same time.

The *Neyman-Pearson paradigm*:

- Make sure that $\mathbb{P}[\text{Type I error}] \le \alpha$ (e.g., $\alpha = 5\%, 1\%, \ldots$)
- Minimize $\mathbb{P}[\text{Type II error}]$ subject to this constraint

Level

The value of $\alpha \in (0,1)$ chosen in the Neyman-pearson paradigm is called level of a test

For which $\theta \in \Theta_0$ should we compute $\mathbb{P}_\theta[\psi = 1]$ (probability of Type I error)?

- A test ψ has *level* α if

$$\mathbb{P}_\theta[\psi = 1] \leq \alpha, \quad \forall\ \theta \ \in \Theta_0.$$

$$\iff \max_{\theta \in \Theta_0} \mathbb{P}_\theta[\psi = 1] \leq \alpha$$

- A test $\psi = \psi_n$ has *asymptotic level* α if

$$\lim_{n \to \infty} \max_{\theta \in \Theta_0} \mathbb{P}_\theta[\psi_n = 1] \leq \alpha,$$

Building a test from a confidence interval

Given a confidence interval, we can often build a test (and vice versa).

- Let $I = [A, B]$ be a confidence interval at level $1 - \alpha$ for a parameter θ:
$$\mathbb{P}_\theta(\theta \in [A, B]) \geq 1 - \alpha$$
- We want to use this I to build a test at level α for
$$\begin{aligned} H_0 : &\quad \theta = \theta_0 \\ H_1 : &\quad \theta \neq \theta_0 \end{aligned}$$
- Natural candidate:
$$\psi = \mathbb{1}\{ \quad \theta_0 \notin [A, B] \quad \}$$
- Level of test:
$$\mathbb{P}_{\theta_0}[\psi = 1] = \mathbb{P}_{\theta_0}\left[\theta_0 \notin I\right] = 1 - \mathbb{P}_{\theta_0}\left[\theta_0 \in I\right] \leq 1 - (1 - \alpha) = \alpha$$
- Therefore ψ is a test with level α

A test for the Kiss example

We want to test:

$$\begin{aligned} H_0: &\quad p = 0.5 \\ H_1: &\quad p \neq 0.5 \end{aligned}$$

We observe $R_1, \ldots, R_n \overset{iid}{\sim} \mathsf{Ber}(p)$.

- Recall that

$$\mathcal{I}_{\mathsf{conserv}} = \left[\bar{R}_n \pm \frac{1.96}{2\sqrt{n}}\right]$$

  is a confidence interval of asymptotic level $1 - \alpha$ for p.

- Consider the test:

$$\psi = \mathbb{1}\left\{.5 \notin \mathcal{I}_{\mathsf{conserv}}\right\}$$

- We have

$$\lim_{n\to\infty} \mathbb{P}_{.5}[\psi = 1] = 1 - \lim_{n\to\infty} \mathbb{P}_{.5}\left[.5 \in \mathcal{I}_{\mathsf{conserv}}\right] \le 1 - (1 - \alpha) = \alpha$$

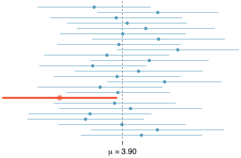
- Therefore ψ is a test with asymptotic level α

Meaning of the level

- Recall that

  $\mathcal{I}$ *is a CI at level* 95% *for* θ

  means that if we repeat the experiment many times, at least 95% confidence intervals will contain the true parameter θ.



Confidence interval coverage plot, (c)OpenIntro.org

- Similarly:

  ψ *is a test at level* 5% *for* H_0 *vs* H_1

  means that if we repeat the experiment many times, at most 5% of the tests will make an error of type I.

What if we change the level?

With our data $\mathcal{I}_{\textsf{conserv}} = [0.56, 0.73]$ so we reject H_0 at level 5%.

α	$q_{\alpha/2}$	$\mathcal{I}_{\textsf{conserv}}$	decision
10%	1.64	$[0.57, 0.72]$	Reject
5%	1.96	$[0.56, 0.73]$	Reject
1%	2.76	$[0.52, 0.77]$	Reject
$.1\%$	3.29	$[0.497, 0.79]$	Fail to reject
$.01\%$	3.89	$[0.47, 0.82]$	Fail to reject

The value of α across which we switch from "reject" to "fail to reject" is called the p-value.

p-value

Definition

The (asymptotic) *p-value* of a test ψ is the smallest (asymptotic) level α at which ψ rejects H_0.

Golden rule

p-value $\leq \alpha \quad \Leftrightarrow \quad H_0$ is rejected by ψ, at the (asymptotic) level α.

Kiss example: we need to find α_0 such that $\bar{R}_n - \frac{q_{\alpha_0/2}}{2\sqrt{n}} = 0.5$

If $\bar{R}_n = .645, \; n = 124$ we get $q_{\alpha_0/2} = 3.23$. To find α_0:

$$\frac{\alpha_0}{2} = \mathbb{P}[Z > 3.23] = 1 - 0.9994 = 0.06\% \qquad \Rightarrow \; \alpha_0 = 0.12\%$$

where $Z \sim \mathcal{N}(0,1)$ and $\mathbb{P}(Z \leq 3.24) = 0.9994$ (read from table).

The evidence scale

- Statisticians, and more generally researchers, are used to communicating directly in terms of p-values rather than "reject/fail to reject at level..."
- The mental conversion is as follows:

p-value	evidence against H_0
$> 10\%$	almost none
$[5\%, 10\%]$	weak
$[1\%, 5\%]$	strong
$[.1\%, 1\%]$	very strong
$< .1\%$	undisputable

Parametric hypothesis testing

Parametric hypothesis testing

- Given the duality between confidence intervals (CI) and tests, it is not surprising that the same tools will be used.
- A simple approach: first build CI, then deduce a test is nice but *limited*: one/two-sided, two sample tests are more common than confidence intervals for (say) $\mu_{\mathrm{d}} > \mu_{\mathrm{c}}$.
- Easier to unfold the same machinery: this is the principle behind the Wald test.
- Wald's test only guarantees *asymptotic* level. An alternative is the T-test.

The Wald test (1)

- Statistical model $(E, \{\mathbb{P}_\theta\}_{\theta \in \Theta})$
- Estimator $\hat{\theta}$ such that $\dfrac{\hat{\theta} - \theta}{\sqrt{\widehat{\mathsf{var}}(\hat{\theta})}} \xrightarrow[n \to \infty]{(d)} \mathcal{N}(0, 1)$

  where $\widehat{\mathsf{var}}(\hat{\theta})$ is an estimator of the variance of $\hat{\theta}$
- For example, in the Bernoulli case, $\widehat{\mathsf{var}}(\hat{p}) = \dfrac{\hat{p}(1 - \hat{p})}{n}$.

The Wald test (2)

	$H_0: \quad \theta = \theta_0$ $H_1: \quad \theta \neq \theta_0$	$H_0: \quad \theta \leq \theta_0$ $H_1: \quad \theta > \theta_0$	$H_0: \quad \theta \geq \theta_0$ $H_1: \quad \theta < \theta_0$
Wald Test ψ	$\mathbb{1}(\lvert W \rvert > q_{\alpha/2})$	$\mathbb{1}(W > q_{\alpha})$	$\mathbb{1}(W < -q_{\alpha})$

$$W := \frac{\hat{\theta} - \theta_0}{\sqrt{\widehat{\mathsf{var}}(\hat{\theta})}}$$

Asymptotic level of the Wald test (case 1)

If

$$\begin{aligned} H_0 : &\quad \theta = \theta_0 \\ H_1 : &\quad \theta \neq \theta_0 \end{aligned}$$

Then, for any $\theta = \theta_0$,

$$\lim_{n\to\infty} \mathbb{P}_{\theta_0}[\psi = 1] = \lim_{n\to\infty} \mathbb{P}_{\theta_0}[|W| > q_{\alpha/2}] = \mathbb{P}[|Z| > q_{\alpha/2}] = \alpha$$

Note that it is important to take the same θ_0 in $\mathbb{P}_{\theta_0}$ and W!

Asymptotic level of the Wald test (case 2 & 3)

If

$$\begin{aligned} H_0 : &\quad \theta \le \theta_0 \\ H_1 : &\quad \theta > \theta_0 \end{aligned}$$

Then, for any $\theta \le \theta_0$,

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}_\theta[\psi = 1] &= \lim_{n\to\infty} \mathbb{P}_\theta[W > q_\alpha] \\ &= \lim_{n\to\infty} \mathbb{P}_\theta\left[\frac{\hat\theta - \theta_0}{\sqrt{\widehat{\mathsf{var}}(\hat\theta)}} > q_\alpha\right] \\ &\le \lim_{n\to\infty} \mathbb{P}_\theta\left[\frac{\hat\theta - \theta}{\sqrt{\widehat{\mathsf{var}}(\hat\theta)}} > q_\alpha\right] \\ &= \mathbb{P}[Z > q_\alpha] = \alpha \end{aligned}$$

Example 1: News

More than 2/3 of Americans get news on social media

Is this quote from a 2018 Pew Research Center study justified?
$X_1, \dots, X_n \overset{iid}{\sim} \mathsf{Ber}(p)$, $p \in [0,1]$,

$$\begin{array}{ll} H_0: & p \leq 2/3 \\ H_1: & p > 2/3 \end{array}$$

This claim is based on $n = 4,581$ randomly sampled U.S., $\hat{p} = .68$.

$$W^{\text{obs}} = \sqrt{4,581}\frac{.68 - 2/3}{\sqrt{.68(1-.68)}} = 1.93 \quad > 1.645 \text{ so Reject}$$

The p-value is α_0 such that

$$q_{\alpha_0} = 1.93 \quad \Longleftrightarrow \quad \alpha_0 = \mathbb{P}(Z > 1.93) = 1 - 0.9732 = 2.68\%$$

Fail to reject at asymptotic level $\alpha = 1\%$.

p-values for the Wald test

- Recall that $W := \dfrac{\hat{\theta} - \theta_0}{\sqrt{\widehat{\mathsf{var}}(\hat{\theta})}}$.
- Denote by W^{obs} the realization (observed value) of W in a given example. For the News example, $W^{\mathrm{obs}} = 1.93$
- Then p-values and asymptotic p-values are given by

	$H_0: \quad \theta = \theta_0$ $H_1: \quad \theta \neq \theta_0$	$H_0: \quad \theta \leq \theta_0$ $H_1: \quad \theta > \theta_0$	$H_0: \quad \theta \geq \theta_0$ $H_1: \quad \theta < \theta_0$
Wald test	$\lvert W\rvert > q_{\alpha/2}$	$W > q_\alpha$	$W < -q_\alpha$
p-value	$\mathbb{P}(\lvert W\rvert > \lvert W^{\mathrm{obs}}\rvert)$	$\mathbb{P}(W > W^{\mathrm{obs}})$	$\mathbb{P}(W < W^{\mathrm{obs}})$
asymp. p-value	$\mathbb{P}(\lvert Z\rvert > \lvert W^{\mathrm{obs}}\rvert)$	$\mathbb{P}(Z > W^{\mathrm{obs}})$	$\mathbb{P}(Z < W^{\mathrm{obs}})$

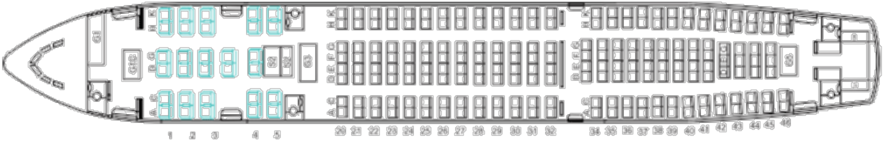
where $Z \sim \mathcal{N}(0, 1)$

Example 2: How to board a plane?

What is the fastest method to board a plane?

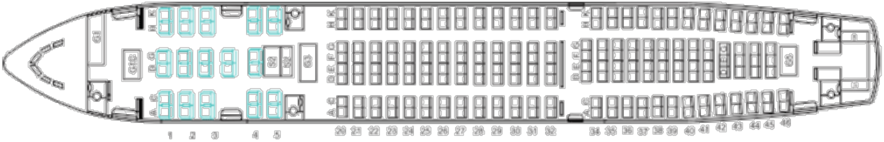
R2F or WiIMA?

- R2F= Rear to Front (JetBlue)



(c) airbus.com

- WiIMA=Window, Middle, Aisle (United)



(c) airbus.com

Model and Assumptions

- X: boarding time of a random JetBlue flight.

$$\mathbb{E}[X] = \mu_1, \qquad \mathsf{var}[X] = \sigma_1^2$$

- Y: boarding time of a random United flight.

$$\mathbb{E}[Y] = \mu_2, \qquad \mathsf{var}[Y] = \sigma_2^2$$

- We have $X_1, \ldots, X_n$ independent copies of X and $Y_1, \ldots Y_m$ independent copies of Y.
- We further assume that the two samples are independent.

Is there a difference between the two boarding methods:

$$\begin{array}{rl} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{array}$$

Equivalently, write $\theta = \mu_1 - \mu_2$, we get

$$\begin{array}{rl} H_0: & \theta = 0 \\ H_1: & \theta \neq 0 \end{array}$$

We have two samples: this is a **two-sample** testing problem.

Asymptotically normal estimator for θ

- Define the estimator $\hat{\theta} = \bar{X}_n - \bar{Y}_m$.
- We have by the CLT:

$$\frac{\hat{\theta} - \theta}{\sqrt{\mathsf{var}(\hat{\theta})}} \xrightarrow[n \to \infty]{(d)} \mathcal{N}(0,1)$$

- But: $\mathsf{var}(\hat{\theta}) = \mathsf{var}(\bar{X}_n) + \mathsf{var}(\bar{Y}_m) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$
- We can estimate σ_1^2 by $\hat{\sigma}_1^2$ and σ_2^2 by $\hat{\sigma}_2^2$ where

$$\hat{\sigma}_1^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \qquad \hat{\sigma}_2^2 := \frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2$$

- Both estimators are consistent so by Slutsky

$$\frac{\hat{\theta} - \theta}{\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}} \xrightarrow[\substack{n \to \infty \\ m \to \infty}]{(d)} \mathcal{N}(0,1)$$

Applying the Wald test

$$W = \frac{\hat{\theta} - 0}{\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}} \qquad \psi = \mathbb{1}\{|W| > q_{\alpha/2}\}$$

Data from JetBlue (R2F) and United (WiIMA):

	R2F	WiIMA
Average (mins)	24.2	25.9
Std. Dev (mins)	5.1	4.3
Sample size	72	56

$$W = \frac{24.2 - 25.9}{\sqrt{\frac{5.1^2}{72} + \frac{4.3^2}{56}}} = -2.04$$

The (asymptotic) p-value is given by

$$\alpha_0 = \mathbb{P}[|Z| > 2.04] = 2\mathbb{P}[Z < -2.04] = 2 \cdot 0.0207 = 4.14\%$$

Example 3: Waiting for the T

Waiting times for the T: $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$.

$$\begin{aligned} H_0 : &\quad \lambda \geq 1 \\ H_1 : &\quad \lambda < 1 \end{aligned}$$

- Recall that using the Delta-method, we got for $\hat{\lambda} = 1/\bar{X}_n$,

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \lambda^2)$$

- Therefore, by Slutsky

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

- Test statistic

$$W = \sqrt{n}\frac{1}{\hat{\lambda}}(\hat{\lambda} - 1)$$

- Reject at 5% if $W < 1.645$.

Example 4: MLE and the Wald test

- Recall that under some regularity conditions, we have:

$$\sqrt{n}(\hat{\theta}^{\mathsf{MLE}} - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \frac{1}{I(\theta)})$$

  where $I(\theta)$ is the Fisher information.

- Using Slutsky, we get

$$\sqrt{nI(\hat{\theta}^{\mathsf{MLE}})}(\hat{\theta}^{\mathsf{MLE}} - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, 1)$$

- Therefore, we can use the Wald test with test statistic given by

$$W = \sqrt{nI(\hat{\theta}^{\mathsf{MLE}})}(\hat{\theta}^{\mathsf{MLE}} - \theta_0)$$

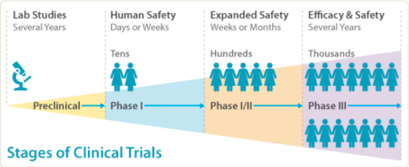
Who was Wald?

- Abraham Wald (1902-1950) was a very influential statistician and mathematician
- First (correct) proof of consistency of MLE under general conditions
- Introduced the first notion of curvature of a metric space
- Worked on aircraft damage during WWII
- Died in a plane crash in India while giving lectures across the country

Photo by onrad acobs, Erlangen, (c) he Mathematisches orschungsinstitut Oberwolfach gGmb (M O). CC B -SA 2.0 Germany

Small sample sizes

- Sometimes, sample sizes are too small to apply CLT/Slutsky which is central to the Wald test.
- This is often the case for early phases clinical trials
- No magic: we have to assume that our data is **Gaussian**



Home wind turbines

- The DoE recommends a minimum average wind speed of 10 miles an hour for a grid-connected wind turbine
- A candidate home was monitored once a month for a year and 12 measurements $X_1, \ldots, X_{12}$ were collected
- Assume that

$$X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$$

- Can we conclude that there is enough wind at this home?

Photograph © 2022 by Rob Cardillo

What goes wrong?

$$X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$$

- We don't even need the CLT since $\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)$
- But to replace σ^2 with a consistent estimator $\hat{\sigma}^2$, we need Slutsky, which is not true for small sample sizes:
- We carefully choose an unbiased estimator $\hat{\sigma}^2$:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

- We are going to use the representation:

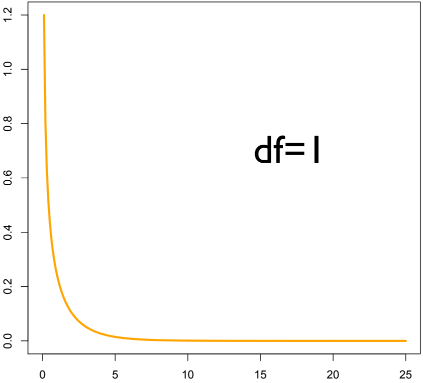
$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} = \frac{\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}}{\sqrt{\frac{S_n^2}{\sigma^2}}}$$

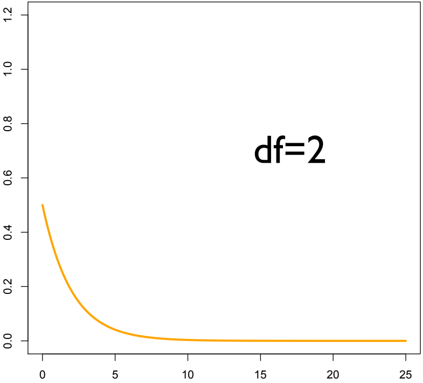
The χ^2 distribution

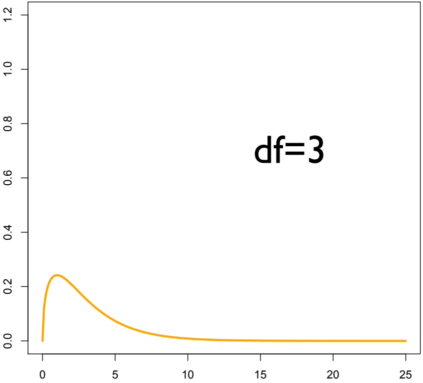
Definition

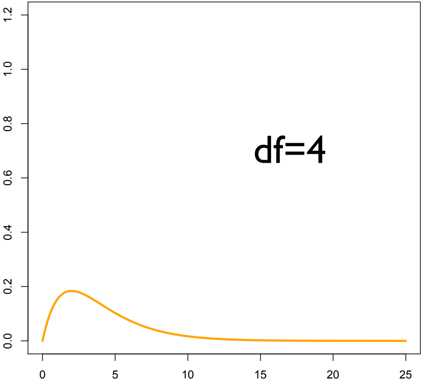
For a positive integer k, the χ^2 *(pronounced "Kai-squared") distribution with* k *degrees of freedom* is the law of the random variable $Z_1^2 + Z_2^2 + \ldots + Z_k^2$, where $Z_1, \ldots, Z_k \overset{iid}{\sim} \mathcal{N}(0,1)$.

Example: If $Z \sim \mathcal{N}_k(\mathbf{0}, I_k)$, then $\|Z\|_2^2 \sim \chi_k^2$.

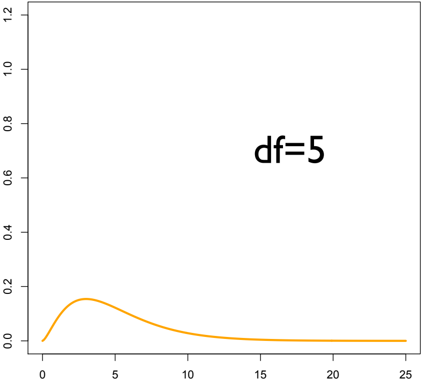

1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
df=1

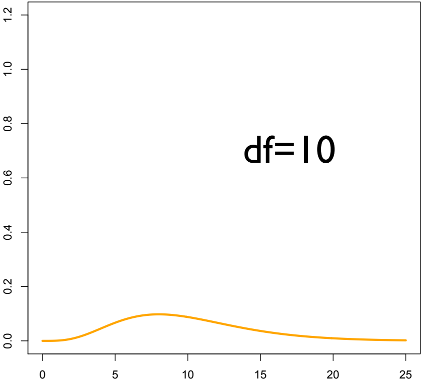

df=2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25

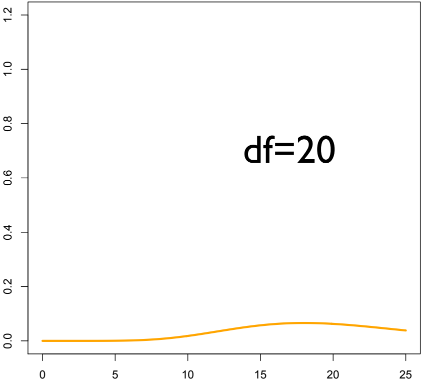

df=3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25



df=4
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25


df=5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25


df=10
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25


df=20
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25

Properties of the χ^2 distribution

Definition

For a positive integer k, the χ^2 *(pronounced "Kai-squared") distribution with k degrees of freedom* is the law of the random variable $Z_1^2 + Z_2^2 + \ldots + Z_k^2$, where $Z_1, \ldots, Z_k \overset{iid}{\sim} \mathcal{N}(0,1)$.

Properties: If $V \sim \chi_k^2$, then

- $\mathbb{E}[V] = k$
- $\text{var}[V] = 2k$

Important example: the sample variance

- Recall that $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$
- **Cochran's theorem:** If $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then
  - $\frac{(n-1)S_n^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}_n}{\sigma}\right)^2 \sim \chi^2_{n-1}.$
  - $\bar{X}_n$ and S_n^2 are independent r.v.;
- Therefore

$$\frac{\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}}{\sqrt{\frac{S_n^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{V}{n-1}}}$$

where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{n-1}$ are independent

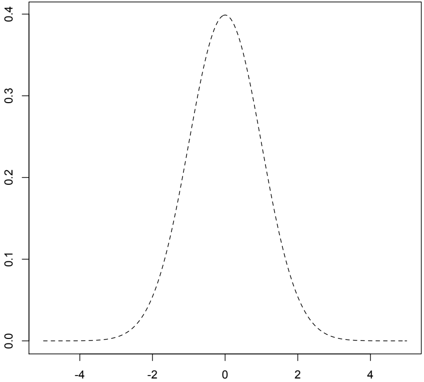
Student's T distribution

Definition

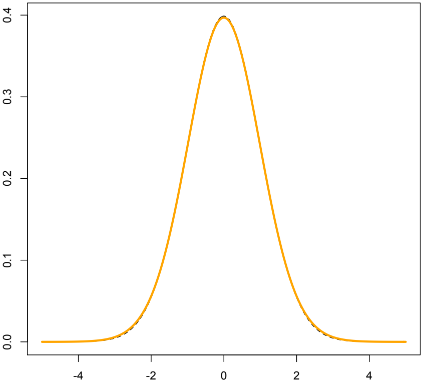
For a positive integer k, *Student's T distribution with k degrees of freedom* (denoted by t_k) is the law of the random variable

$$\frac{Z}{\sqrt{V/k}}$$

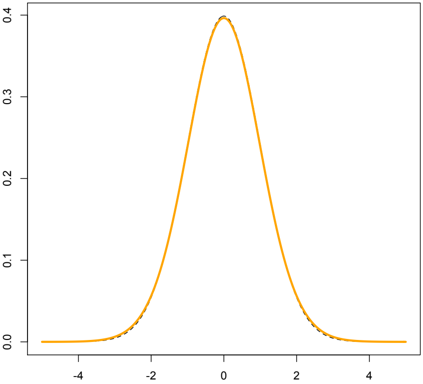
where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi_k^2$ are independent r.v.



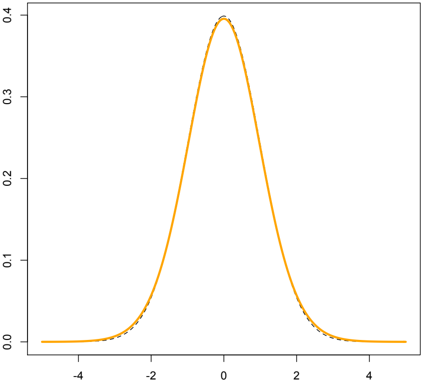
The standard normal pdf



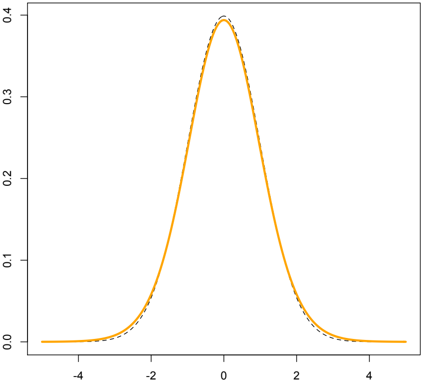
The t_{50} pdf



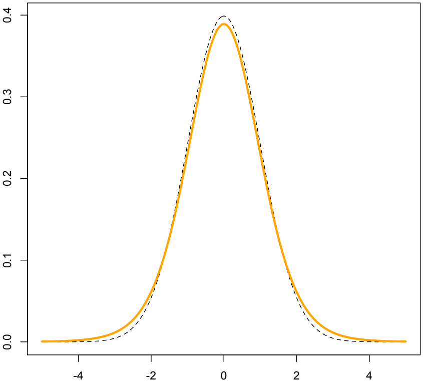
The t_{40} pdf



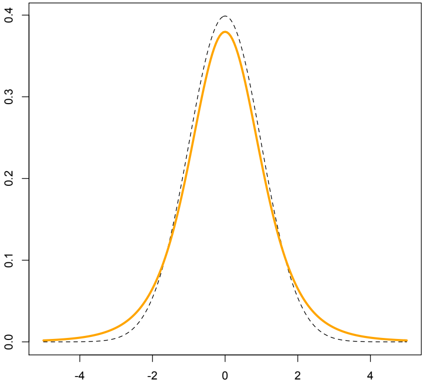
The t_{30} pdf



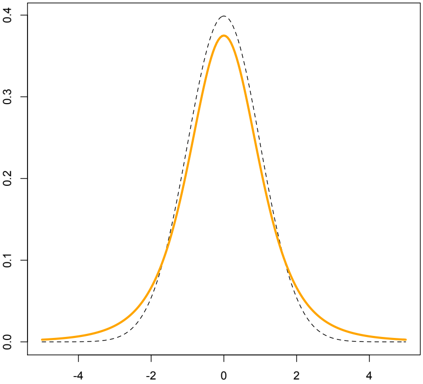
The t_{20} pdf



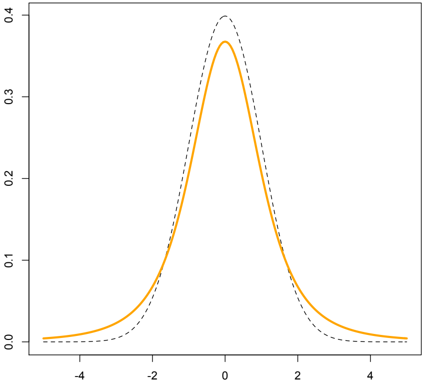
The t_{10} pdf



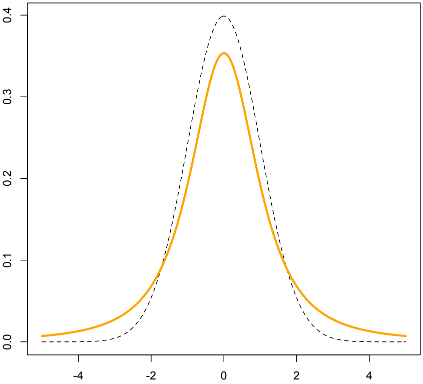
The t_5 pdf



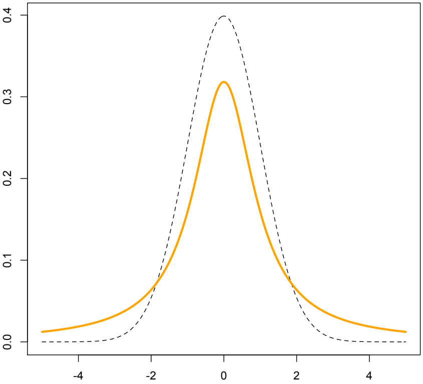
The t_4 pdf



The t_3 pdf



The t_2 pdf



The t_1 pdf

Student's T test (one sample)

- Gaussian model: $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ both μ and σ^2 unknown
- We know that

$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \sim t_{n-1}$$

- Then T-tests are given by:

	$H_0: \quad \mu = \mu_0$ $H_1: \quad \mu \neq \mu_0$	$H_0: \quad \mu \leq \mu_0$ $H_1: \quad \mu > \mu_0$	$H_0: \quad \mu \geq \mu_0$ $H_1: \quad \mu < \mu_0$
T-Test ψ	$\mathbb{1}\{\lvert T\rvert > q_{\alpha/2}^{t_{n-1}}\}$	$\mathbb{1}\{T > q_{\alpha}^{t_{n-1}}\}$	$\mathbb{1}\{T < -q_{\alpha}^{t_{n-1}}\}$

where

$$T = \sqrt{n}\frac{\bar{X}_n - \mu_0}{S_n} \qquad \text{and} \qquad \mathbb{P}[t_{n-1} > q_{\alpha}^{t_{n-1}}] = \alpha$$

- Using the same analysis as for Wald's test, it can be shown that the T-test has (non-asymptotic) level α

Example: Home wind turbines

- We observe: $n = 12, \qquad \bar{X}_n = 14.3 \qquad S_n = 4.7 \quad \mu_0 = 10$

$$\begin{aligned} H_0: &\quad \mu \leq 10 \\ H_1: &\quad \mu > 10 \end{aligned}$$

- We get $T = \sqrt{12}\dfrac{14.3 - 10}{4.7} = 3.17$
- From table: $q^{t_{11}}_{5\%} = 1.80$ so we reject since $3.17 > 1.80$

p-values for the T-test

	$H_0: \quad \mu = \mu_0$ $H_1: \quad \mu \neq \mu_0$	$H_0: \quad \mu \leq \mu_0$ $H_1: \quad \mu > \mu_0$	$H_0: \quad \mu \geq \mu_0$ $H_1: \quad \mu < \mu_0$
T-Test	$\mathbb{1}\{\lvert T\rvert > q_{\alpha/2}^{t_{n-1}}\}$	$\mathbb{1}\{T > q_{\alpha}^{t_{n-1}}\}$	$\mathbb{1}\{T < -q_{\alpha}^{t_{n-1}}\}$
p-value	$\mathbb{P}[\lvert T\rvert > \lvert T^{\mathrm{obs}}\rvert]$	$\mathbb{P}[T > T^{\mathrm{obs}}]$	$\mathbb{P}[T < T^{\mathrm{obs}}]$

where $T \sim t_{n-1}$

Home wind turbines: p-value$=\mathbb{P}[t_{11} > 3.17] = 0.446\%$ (very strong evidence)

Comparison with the Wald test

- If this was a Wald test at asymptotic level 5%, we would compare this to $q_{5\%} = 1.645$ and reject
- In this setup, only difference is $q_{5\%}^{t_{11}} = 1.80$ vs. $q_{5\%} = 1.645$ (Gaussian quantiles).
- For n large enough, difference becomes very small.
- But in general, Wald test is more flexible (any θ, $\hat{\text{var}}(\hat{\theta})$, etc.)

α	10%	5%	2.5%	1%	0.5%
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
$\vdots$					
30	1.31	1.70	2.04	2.46	2.75
$\vdots$					
50	1.30	1.68	2.01	2.40	2.68
∞	1.28	1.65	1.96	2.33	2.58

Clinical trial example (1)

- $\mu_{\mathrm{d}} > 0$: expected decrease of LDL in test group.
- $\mu_{\mathrm{c}} > 0$: expected decrease of LDL level in control group.
- Hypothesis testing problem:

$$\begin{array}{ll} H_0: & \mu_{\mathrm{d}} - \mu_{\mathrm{c}} \leq 0 \\ H_1: & \mu_{\mathrm{d}} - \mu_{\mathrm{c}} > 0 \end{array}$$

- We observe two independent samples:
  - $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu_{\mathrm{d}}\ , \sigma_{\mathrm{d}}^2)$ from the test group and
  - $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathcal{N}(\mu_{\mathrm{c}}\ , \sigma_{\mathrm{c}}^2)$ from the control group.
- We have

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_{\mathrm{d}} - \mu_{\mathrm{c}})}{\sqrt{\frac{\sigma_{\mathrm{d}}^2}{n} + \frac{\sigma_{\mathrm{c}}^2}{m}}} \sim \mathcal{N}(0, 1)$$

Clinical trial example (2)

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_{\mathrm{d}} - \mu_{\mathrm{c}})}{\sqrt{\frac{S_{\mathrm{d}}^2}{n} + \frac{S_{\mathrm{c}}^2}{m}}} = \frac{\frac{\bar{X}_n - \bar{Y}_m - (\mu_{\mathrm{d}} - \mu_{\mathrm{c}})}{\sqrt{\frac{\sigma_{\mathrm{d}}^2}{n} + \frac{\sigma_{\mathrm{c}}^2}{m}}}}{\sqrt{\frac{\frac{S_{\mathrm{d}}^2}{n} + \frac{S_{\mathrm{c}}^2}{m}}{\frac{\sigma_{\mathrm{d}}^2}{n} + \frac{\sigma_{\mathrm{c}}^2}{m}}}}$$

- Good news: $\frac{\bar{X}_n - \bar{Y}_m - (\mu_{\mathrm{d}} - \mu_{\mathrm{c}})}{\sqrt{\frac{\sigma_{\mathrm{d}}^2}{n} + \frac{\sigma_{\mathrm{c}}^2}{m}}} \sim \mathcal{N}(0,1)$
- Bad news: don't know the distribution of $\sqrt{\frac{\frac{S_{\mathrm{d}}^2}{n} + \frac{S_{\mathrm{c}}^2}{m}}{\frac{\sigma_{\mathrm{d}}^2}{n} + \frac{\sigma_{\mathrm{c}}^2}{m}}}$

Two-sample T-test

- We have approximately

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_\mathrm{d} - \mu_\mathrm{c})}{\sqrt{\frac{S_\mathrm{d}^2}{n} + \frac{S_\mathrm{c}^2}{m}}} \sim t_N$$

  where

$$N = \frac{\left(S_\mathrm{d}^2/n + S_\mathrm{c}^2/m\right)^2}{\frac{S_\mathrm{d}^4}{n^2(n-1)} + \frac{S_\mathrm{c}^4}{m^2(m-1)}} \gtrsim \min(n, m)$$

- This is Welch-Satterthwaite (WS) formula
- Sanity check: if $m \to \infty$ (one sample limit), we have $N \to n - 1.$

Non-asymptotic test

- Example $n = 12, m = 22$, $\bar{X}_n = 156.4$, $\bar{Y}_m = 132.7$, $S_{\mathrm{d}} = 22.5$, $S_{\mathrm{c}} = 8.7$,

$$T = \frac{156.4 - 132.7}{\sqrt{\frac{22.5^2}{12} + \frac{8.7^2}{22}}} = 3.51$$

- Using the WS formula:

$$N = \frac{\left(22.5^2/12 + 8.7^2/22\right)^2}{\frac{22.5^4}{12^2 \cdot 11} + \frac{8.7^4}{22^2 \cdot 21}} = 12.82$$

we round to 12.

- We get

$$\text{p-value} = \mathbb{P}(t_{12} > 3.51) \quad = 0.215\%$$

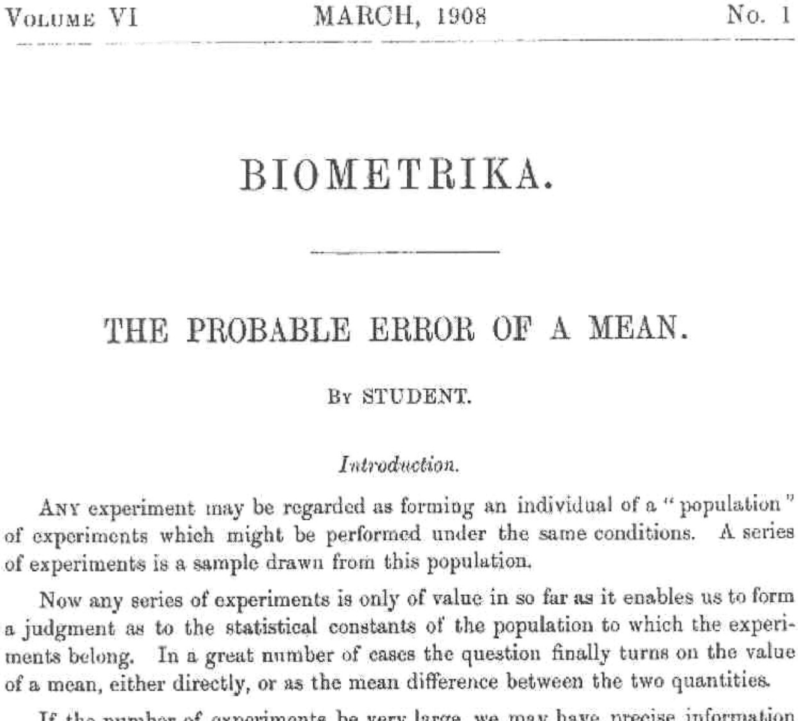
So we reject the test: there is strong evidence that the drug is effective.

Who was Student?

William Sealy Gosset.
Image Credit: Wikipedia.org

© Guinness & Co. 2022

This distribution was introduced by **William Sealy Gosset** (1876–1937) in 1908 while he was "head experimental brewer" for the Guinness brewery in Dublin, Ireland. He published his work under the pseudonym "Student" because Guinness forbid its employees to publish their results.



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information

Discussion

Advantage of Student's test: Non asymptotic / Can be run on small samples

Drawback of Student's test: It relies on the assumption that the sample is Gaussian (Next unit: we will see how to test this assumption)

The dead salmon experiment: setup

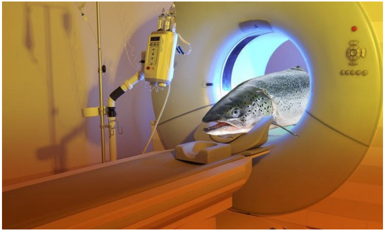


image (C) Craig Bennett (c) prefrontal.org., All rights reserved.

In 2009, neuroscientist Craig Bennett purchased a whole Atlantic salmon, took it to a lab at Dartmouth, and put it into an fMRI machine used to study the brain.

“The salmon was shown a series of $n = 15$ photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.”

The dead salmon experiment:results

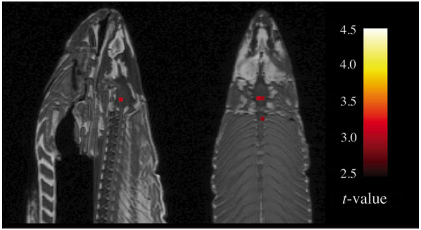


image (c) Craig Bennett (c) prefrontal.org., All rights reserved

- The salmon brain was split into $N = 8,064$ voxels
- In each voxel a statistical test at level $\alpha = .1\%$ was performed to see if activity was statistically significant
- 8 voxels* were found to be.
- Conclusion: these voxels contain salmon neurons involved in social perception. NO!

Statistical analysis

- Statistical model for a fixed voxel: observe (normalized) signal $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$, $n = 15$.
- Hypothesis testing problem:

$$\begin{aligned} H_0 : &\quad \mu = 0 \\ H_1 : &\quad \mu \neq 0 \end{aligned}$$

- Apply the T-test:
- Recall the interpretation of level $\alpha = .1\%$ of a test:

  If we repeat the experiment many times, at most .1% of the tests will make an error of type I.

- $.1\% \cdot 8,064 \simeq 8$

Multiple testing

- We cannot make conclusions about "all the voxels" at once: we are bound to make mistakes
- Two solutions:
  - Control Family Wise Error Rate (FWER): Find $C_1, \dots, C_N$ such that

$$\mathbb{P}_{\mu_i=0}\Big(\bigcup_{i=1}^{N}\{|T_i| > C_i\}\Big) \leq \alpha$$

  - Control False Discovery Rate (FDR): Find $C_1, \dots, C_N$ such that

$$\text{FDR} = \mathbb{E}\Big[\frac{\#\{i\,:\,|T_i| > C_i\ \&\ \mu_i = 0\}}{\#\{i\,:\,|T_i| > C_i\}}\Big] = \mathbb{E}_{\mu=0}\Big[\frac{\#\text{ of False discoveries}}{\#\text{ of discoveries}}\Big] \leq \alpha$$

- In both cases, it is easier to work with the p-values:

$$P_i = \mathbb{P}_{\mu_i=0}(|T| > |t_i^{\text{obs}}|)$$

where t_i^{obs} is the observed value of the test statistic for the ith test and $T \sim t_{n-1}$.

The Bonferroni method

- To control FWER, we use use the *Bonferroni correction*.
- Rather than rejecting each test at level α, we use the (much smaller) level α/N.
- In other words:

$$\text{Reject } i\text{th test} \quad \Longleftrightarrow \quad P_i < \frac{\alpha}{N}$$

- In the salmon example this means that each test is performed at level $\quad 0.001/8064 = 1.24 \cdot 10^{-7}$
- Often this is way to *conservative* (no discoveries).

Note that, with the Bonferroni method:

$$\begin{aligned} \text{FWER} &= \mathbb{P}_{\mu_i=0}\big(\bigcup_{i=1}^{N}\{P_i < \frac{\alpha}{N}\}\big) = \mathbb{P}_{\mu_i=0}\big(\bigcup_{i=1}^{N}\{|T_i| > q_{\frac{\alpha}{2N}}\}\big) \\ &\leq \sum_{i=1}^{N}\mathbb{P}_{\mu_i=0}\big(\{|T_i| > q_{\frac{\alpha}{2N}}^{t_{n-1}}\}\big) = N \cdot \frac{\alpha}{N} = \alpha \end{aligned}$$

The Benjamini-Hochberg method

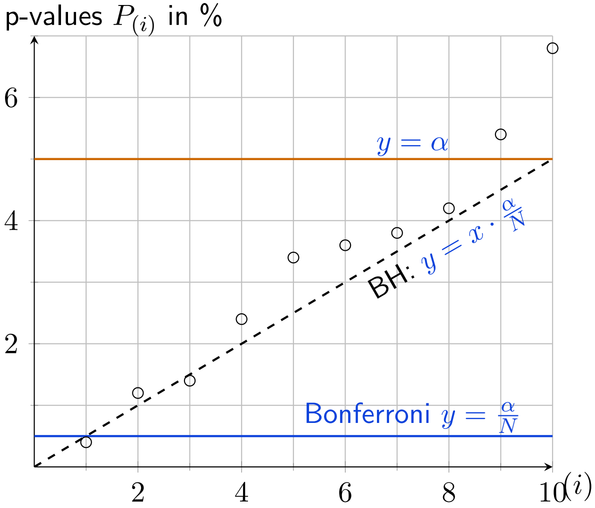
- To control FDR, we use use the *Benjamini-Hochberg (BH)* method.
- Intuitively, we should reject tests with the smallest p-values
- Order p-values: $P_{(1)} < P_{(2)} < P_{(3)} < P_{(4)} < \cdots < P_{(N)}$ and call "the (i)th test" the test with p-value $P_{(i)}$.
- Idea: reject all tests (i) such that $i \leq i_{\max}$
- Rule

$$i_{\max} := \max \left\{ i \, : \, P_{(i)} < i \cdot \frac{\alpha}{N} \right\}$$

- Benjamini and Hochberg (1995, 68K citations) have shown that with this procedure:

$$\text{FDR} \leq \alpha$$

- There are *many* variations of the BH procedure, in particular to account for correlations between p-values.



In this figure, $N = 10, \alpha = 5\%$.

- Bonferonni: $\alpha / N = .005$. Only test (1) is rejected.
- BH: $i_{\max} = 3$. Tests $(1), (2)$ and (3) are rejected.

Recap

- Given an asymptotically normal estimator, we can build the Wald test using quantiles of the Gaussian
- When the data is Gaussian, we can use the T-test even for small sample sizes (quantiles of Student's T distribution)
- For large sample sizes, the quantiles of Student's T distribution converge to those of the Gaussian distribution
- When performing multiple tests, one needs to be careful and apply a correction: Bonferroni controls the FWER but is very conservative, BH controls the FDR and is very popular because it is less conservative.
- The statements FWER $\leq \alpha$ and FDR $\leq \alpha$ are often erroneously conflated but they have different meanings that

ATTRIBUTION LIST

Slide #2
https://doctors.healthtap.com/hs-fs/hubfs/waiting_room.jpg?t=1511990105503&width=640&name=waiting_room.jpg
© 2018 HealthTap, All rights reserved.

Slide #16
Confidence interval coverage plot
© OpenIntro.org

Slide #28
Airbus A300-600, seating chart
https://www.airbus.com/en/who-we-are/our-history/commercial-aircraft-history/previous-generation-aircraft/a300-600
© airbus.com

Slide #34
Photograph of Abraham Wald by Konrad Jacobs, Erlangen,
© The Mathematisches Forschungsinstitut Oberwolfach gGmbH (MFO).
CC BY-SA 2.0 Germany

Slide # 35
Clinical Trials Approach
© Copyright 2002-2022. International Partnership
for Microbicides. All rights reserved

Slide #36
Photograph of Home
© 2022 by Rob Cardillo

Slides #50
Photograph,William Sealy Gosset.
https://en.wikipedia.org/wiki/William_Sealy_Gosset
Wikipedia.org

Guinness beer image
guinness.com/us   © Guinness & Co. 2022

Slides #53, Slides #54
images (C) Craig Bennett
(c) prefrontal.org., All rights reserved.