Significance Testing

B39AX — Fall 2023

Heriot-Watt University

Motivation

- Can someone detect whether milk is poured over tea or tea is poured over milk? (Lady tasting tea real-life example by Ron Fisher)
- A coin is tossed repeatedly and independently. Is the coin fair?
- We observe a sequence of i.i.d. normal random variables $X_1, \ldots, X_n$ Are they standard normal?
- Is a drug treatment effective?
- How do you test for someone claiming to be clairvoyant?
- How to detect whether or not a signal is present? (detection theory)

Null Hypothesis and Alternative Hypothesis

Example

A friend claims to be clairvoyant. He can guess what a coin flip will be.

We test his powers by performing 30 trials. There are two hypotheses:

Null hypothesis H_0: he is not clairvoyant; his success rate is $\leq 50\%$

(the simplest, default hypothesis)

Alternative hypothesis H_1: he is clairvoyant; his success rate is $> 50\%$

He correctly predicted 20 flips. *Shall we conclude he's clairvoyant?*

Null Hypothesis and Alternative Hypothesis

Let X denote the RV representing "number of correct flip guesses"

$X \sim \mathsf{Bin}(n\,,\,p)$

Null hypothesis H_0: $p \leq 50\%$

Alternative hypothesis H_1: $p > 50\%$

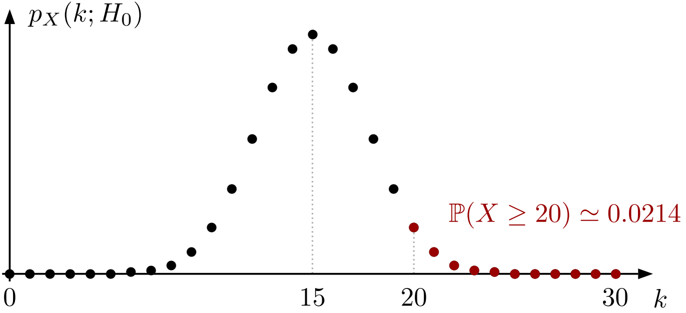
What is the probability of obtaining ≥ 20 correct flips under H_0?

$$\mathbb{P}\big(X \geq 20\,;\,H_0\big) \leq \sum_{k=20}^{30} \binom{30}{k} 0.5^k\,(1-0.5)^{30-k} \simeq \underbrace{0.0214}_{p\text{-value}}$$

Since 0.0214 is small, we would typically reject H_0.

Null Hypothesis and Alternative Hypothesis

$$X \sim \mathsf{Bin}(n\,,\,p) \qquad \Longrightarrow \qquad \mathbb{E}[X] = np = 15$$



If H_0 is true and we repeat the experiment several times, only $\simeq 2.14\%$ of times our friend would correctly predict 20 or more flips.

Error types

True hypothesis	Accept H_0	Accept H_1
H_0 is true	✓	Type I error $\boldsymbol{\alpha}$
H_1 is true	Type II error $\boldsymbol{\beta}$	✓

Both errors cannot be made arbitrarily small (e.g., $\downarrow \beta \implies \uparrow \alpha$)

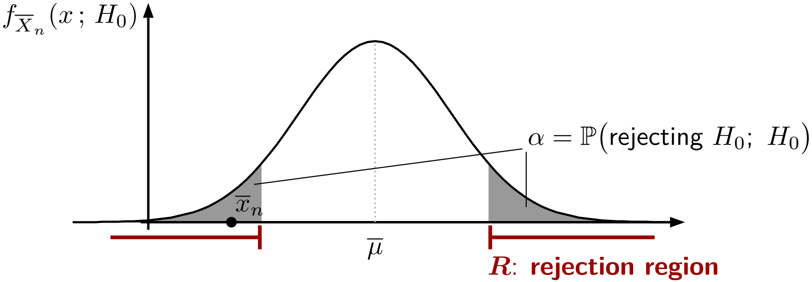
It is common practice to specify only α (e.g., $\alpha = 0.05$)

Example

X is a normal RV: $X \sim \mathcal{N}\big(\mu,\, \sigma^2\big)$

$X_1, \ldots, X_n$; i.i.d. copies of X. Estimator of μ: $\widehat{\Theta}_n = \overline{X}_n \sim \mathcal{N}(\mu,\, \frac{\sigma^2}{n})$

$$H_0 :\ \mu = \overline{\mu} \qquad\qquad H_1 :\ \mu \neq \overline{\mu}$$



If $\overline{x}_n = \dfrac{x_1 + \cdots + x_n}{n} \in R$, we reject H_0 with significance α

Otherwise, we do not reject H_0 (or accept H_0) with significance α

Significance testing

Philosophy: Suppose we want to test a new drug.

If you want to prove the new drug works, you do it by showing the data is inconsistent with the drug not working.

$$H_0: \text{ Drug does not work}$$

$$H_1: \text{ Drug works}$$

Outline of the procedure:

- Build an estimator (statistic) of what we want to test
- Set a significance level α (probability of rejecting H_0 when H_0 is true)
- Find the rejection region R using α (use pdf or pmf of estimator)
- If the realization of the estimator falls into R, then reject H_0 with significance α; otherwise, accept it with significance α

Significance testing procedure

Let X be a RV, and $X_1, \ldots, X_n$ independent copies of X (observations)

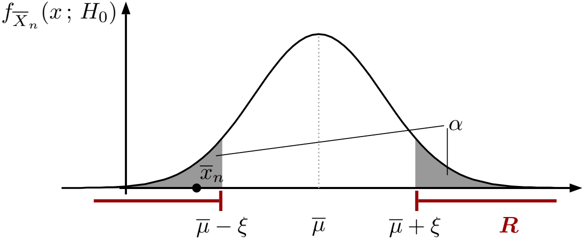
Procedure (before observing the data)

- Formulate the *null hypothesis* H_0 and the *alternative hypothesis* H_1
- Select an estimator (or statistic) $\widehat{\Theta}_n = g(X_1, \ldots, X_n)$
- Determine the shape of the *rejection region* R of H_0 as a function of a critical value ξ (e.g., one-sided or two-sided intervals)
- Choose the significance level α (probability of false rejection of H_0)
- Compute ξ as a function of α (this completely determines R)

Significance testing procedure

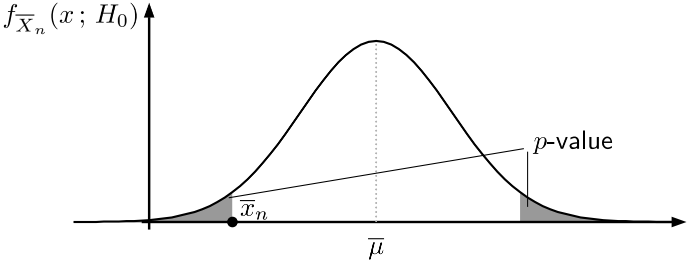
Procedure (after observing the data)

- Calculate value of statistic $\widehat{\theta}_n = g(x_1, \ldots, x_n)$ of $\widehat{\Theta}_n$ (e.g., $\widehat{\Theta}_n = \overline{X}_n$)
- Reject the hypothesis H_0 if it belongs to the rejection region R



Significance testing procedure

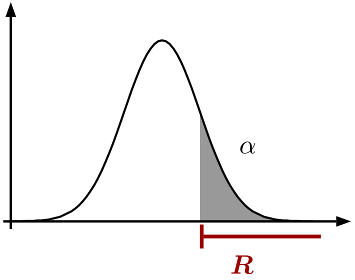
It is common to bypass the selection of α, and just present the p-value



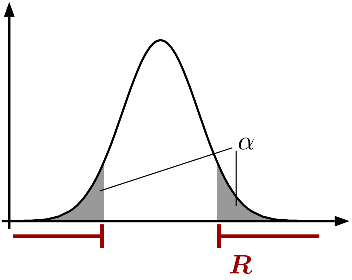
p-value: probability under H_0 of obtaining $\overline{x}_n$ or a more extreme value

If p-value is small (e.g., < 0.05), then H_0 is rejected

One-sided vs two-sided rejection regions



One-sided



Two-sided

Exercise

Let $X \sim \mathcal{N}(\mu, 1)$.

We want to test the hypothesis $\mu \neq 0$ at 5% significance level.

We observed 100 samples of X and their sample average was 0.2.

Perform the significance test, and compute the p-value.

Exercise: solution

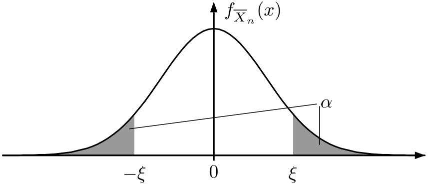
- Formulate the hypotheses:

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

- Estimator for μ: $\overline{X}_n = \frac{X_1+\cdots+X_n}{n} \sim \mathcal{N}\big(\mu, \frac{1}{n}\big)$

- Determine the shape of R, assuming H_0. Two-sided rejection region:



We'll reject H_0 if $|\overline{x}_n| \geq \xi$

Exercise: solution

- Compute ξ from $\alpha = 0.05$. Under H_0, $\overline{X}_n \sim \mathcal{N}\big(0,\, \frac{1}{n}\big)$, so

$$\alpha \geq \mathbb{P}\big(\big|\overline{X}_n\big| \geq \xi\big) \;\;\Leftrightarrow\;\; \Phi\Big(\xi\sqrt{n}\Big) \geq 1 - \frac{\alpha}{2} = 0.975 \;\;\Leftrightarrow\;\; \xi = \frac{1.96}{\sqrt{n}}$$

  So the rejection region will be

$$\begin{aligned} R &= \left(-\infty\,,\, -\frac{1.96}{\sqrt{n}}\right] \bigcup \left[\frac{1.96}{\sqrt{n}}\,,\, +\infty\right) \\ &= \big(-\infty\,,\, -0.196\big] \cup \big[0.196\,,\, +\infty\big) \end{aligned}$$

- We observe the data: $\overline{x}_n = 0.2$, which belongs to the rejection region, i.e., $\overline{x}_n \in R$. So, we reject H_0 at 5% significance level.

- p-value: $\mathbb{P}\big(\big|\overline{X}_n\big| \geq 0.2\big) = 2\mathbb{P}\big(Z \geq 0.2\sqrt{100}\big) = 0.0456$

Exercise

Similar problem, but with unknown variance.

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

We want to test the hypothesis $\mu \neq 0$ at 5% significance level.

We observed 10 samples of X

- their sample average was 0.2
- their sample (unbiased) variance was 1

Perform the significance test.

Exercise: solution

- Same hypotheses: $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$
- Same estimator: $\overline{X}_n = \dfrac{X_1 + \cdots + X_n}{n}$
- Rejection region with same format: $\left|\overline{X}_n\right| \geq \xi$
- To compute ξ (under H_0), we note that $n = 10$ is small, so

$$T_n = \frac{\overline{X}_n - 0}{\widehat{S}_n/\sqrt{n}} \sim t\text{-Student}(9)\,,$$

where $\widehat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$. By symmetry, we have

$$\mathbb{P}\big(|\overline{X}_n| \geq \xi\big) = 2\,\mathbb{P}\big(\overline{X}_n \geq \xi\big) = 2\,\mathbb{P}\big(T_n \geq \xi\sqrt{n}/\widehat{S}_n\big) = 0.05\,,$$

so $\mathbb{P}\big(T_n \geq \xi\sqrt{n}/\widehat{S}_n\big) = 0.025$.

Exercise: solution

From the table, $\mathbb{P}\big(T_n \geq \xi\sqrt{n}/\widehat{S}_n\big) = 0.025$ gives $\xi\sqrt{n}/\widehat{S}_n = 2.262$.

The rejection region is then

$$\begin{aligned} R &= (-\infty\,,\, -\xi] \cup [\xi\,,\, +\infty) \\ &= \Big(-\infty\,,\, -2.262\,\frac{\widehat{S}_n}{\sqrt{n}}\Big] \cup \Big[2.262\,\frac{\widehat{S}_n}{\sqrt{n}}\,,\, +\infty\Big) \\ &= \Big(-\infty\,,\, -0.72\Big] \cup \Big[0.72\,,\, +\infty\Big) \end{aligned}$$

Since $\overline{x}_n = 0.2 \notin R$, we do not reject H_0 with significance 5%.

Exercise

Similar problem, but with one-sided rejection region.

Let $X \sim \mathcal{N}\big(\mu,\, 1\big)$.

We want to test the hypothesis $\mu < 0$ at 5% significance level.

We observed 100 samples of X and their sample average was $\overline{x}_n = -0.2$.

Perform the significance test.

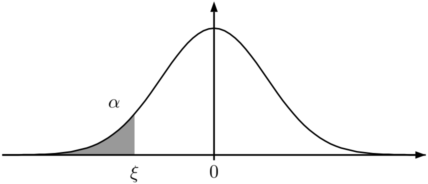
Exercise: solution

- Now the hypotheses are

$$H_0 : \mu \geq 0$$

$$H_1 : \mu < 0$$

- Same estimator for μ: $\overline{X}_n = \dfrac{X_1 + \cdots + X_n}{n}$
- Determine the shape of R. Now we have a one-sided rejection region:



We'll reject H_0 if $\overline{X}_n \leq \xi$

Example: one-sided rejection region

- Compute ξ as a function of $\alpha = 0.05$. Under H_0, $\overline{X}_n \sim \mathcal{N}\big(0,\, \frac{1}{n}\big)$, so

$$\alpha = \mathbb{P}\Big(\overline{X}_n \leq \xi\Big) \;\Leftrightarrow\; \Phi\Big(-\xi\sqrt{n}\Big) = 1 - \alpha = 0.95 \;\Leftrightarrow\; \xi = -\frac{1.65}{\sqrt{n}}$$

  So the rejection region will be

$$R = \left(-\infty\,,\, -\frac{1.65}{\sqrt{n}}\right] = \Big(-\infty\,,\, -0.165\Big]\,.$$

- Since $\overline{x}_n = -0.2 \in R$, we reject H_0 at 5% significance level.

Comparing means

We are testing a medicine for a cold. We select 200 people with a cold.

- To $n_X = 100$ randomly selected people we give the medicine
- To the $n_Y = 100$ remaining people we give a placebo

Assuming the duration of a cold is normal distributed, we want to test whether the medicine is effective with 5% significance level.

Comparing means

Let X_i be the duration of the cold of person i from the *medicine* group.

Let Y_i be the duration of the cold of person i from the *placebo* group.

$$X_i \sim \mathcal{N}\big(\mu_X \,,\, \sigma_X^2\big)\,, \qquad Y_i \sim \mathcal{N}\big(\mu_Y \,,\, \sigma_Y^2\big)$$

Hypotheses: $H_0 \,:\, \mu_X \geq \mu_Y \,,\qquad H_1 \,:\, \mu_X < \mu_Y$

Estimators:

$$\overline{X} = \frac{1}{n_X}\sum_{i=1}^{n_X} X_i \qquad \overline{Y} = \frac{1}{n_Y}\sum_{i=1}^{n_Y} Y_i$$

Rejection region: Reject H_0 if $\overline{Y} - \overline{X} > \xi$

How to compute ξ such that $\mathbb{P}\big(\overline{Y} - \overline{X} > \xi \,;\, H_0\big) \leq \alpha$?

Comparing means

Because sums of independent Gaussians are Gaussian,

$$\overline{Y} - \overline{X} \sim \mathcal{N}\left(\mu_Y - \mu_X\,,\ \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$$

Under H_0, $\mu_Y = \mu_X$. But how do we estimate σ_X^2 and σ_Y^2?

- First estimate the common value of $\mu_Y = \mu_X$ (under H_0):

$$\widehat{\mu} = \frac{\sum_{i=1}^{n_Y} Y_i + \sum_{i=1}^{n_X} X_i}{n_Y + n_X}$$

- Then, because $n_Y = n_X = 100 \gg 50$, use the sample variance

$$\begin{aligned}\mathsf{Var}\big(\overline{Y} - \overline{X}\big) &= \mathsf{Var}\big(\overline{Y}\big) + \mathsf{Var}\big(\overline{X}\big)\\ &\simeq \frac{1}{n_Y - 1}\sum_{i=1}^{n_Y}(Y_i - \widehat{\mu})^2 + \frac{1}{n_X - 1}\sum_{i=1}^{n_X}(X_i - \widehat{\mu})^2\end{aligned}$$

Then, $\mathbb{P}\big(\overline{Y} - \overline{X} > \xi\big) \le \alpha$ can be computed using the normal table.

Pitfalls of significance testing

Recall the example $X \sim \mathcal{N}(\mu,\,1)$ with hypotheses

$$H_0\;:\;\mu = 0\,,\qquad H_1\;:\;\mu \neq 0\,,$$

which we want to test with 5% significance level. We obtained

$$R = \left(-\infty\,,\,-\frac{1.96}{\sqrt{n}}\right] \bigcup \left[\frac{1.96}{\sqrt{n}}\,,\,+\infty\right)$$

Exercise: If the true mean is $\mu = 0.1$ and $n = 100$, what is the probability of accepting H_0? (answer: $83\%\,!$)

Detecting small effects requires many samples

Pitfalls of significance testing

Avoid using significance tests

In the 80's, K. Rothman, editor of the American Journal of Public Health, started rejecting papers that performed significance tests.

Significance tests depend on **size of the effect** you're trying to measure, **number of samples**, and **measurement noise** (e.g., IQ tests)

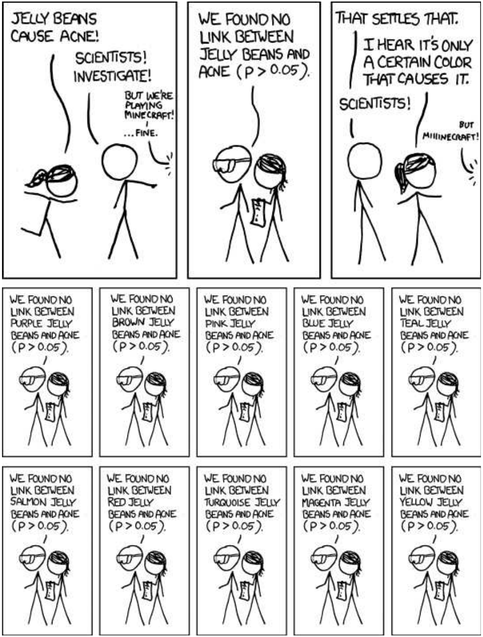
Misinterpreted p-values and poorly executed significance tests abound in literature (even in the journals Nature and Science) and public policy.

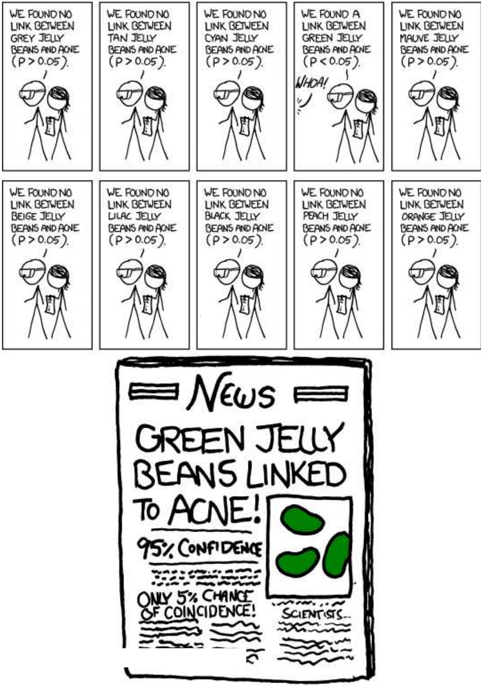
Pitfalls of significance testing

- Right turns on red lights in the US
  (underpowered significance tests did not have enough data to detect the increase in the # of accidents, roughly 20%)

- In $\sim 50\%$ of cancer research studies that report statistical insignificant results, there was not enough data to measure the effect they were trying to find.

- Many times scientists collect data until they obtain a statistical significant result, and stop the collection after that.

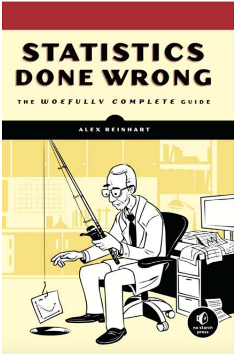

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Pitfalls of significance testing

Underpowered significance tests have created several myths:

- Beautiful parents have more daughters
- Engineers have more sons, nurses have more daughters
- Increasing salt consumption increases blood pressure



Pitfalls of significance testing

Alternatives

- Confidence intervals
- Bayesian inference