SIGNIFICANCE TESTS
STATISTICAL HYPOTHESIS TESTING

Purpose:
test, whether a certain hypothesis is likely to be true

Procedure:
1. Define null hypothesis $H_0$ and alternative hypothesis $H_1$
2. Define level of significance
3. Reject hypothesis or not: type 1 and 2 error

Example: Test, whether a coin is fair
i.e. $H_0$: probability $p = 0.5$ $H_1 = p \neq 0.5$
STATISTICAL HYPOTHESIS TESTING
EXAMPLE: FLIPPING A COIN

**Assumptions:** Bernoulli trial: Binominal \((n,p)\) random variable \(X \rightarrow \) Normal distribution

```python
def normal_approximation_to_binomial(n, p):
    """finds \(\mu\) and \(\sigma\) corresponding to a Binominal(n, p)""
    mu = p * n
    sigma = math.sqrt(p * (1 - p) * n)
    return mu, sigma
```

# probability a normal lies in an interval

```python
# the normal cdf _is_ the probability the variable is below a threshold
normal_probability_below = normal_cdf

# it's above the threshold if it's not below the threshold
def normal_probability_above(lo, mu=0, sigma=1):
    return 1 - normal_cdf(lo, mu, sigma)

# it's between if it's less than hi, but not less than lo
def normal_probability_between(lo, hi, mu=0, sigma=1):
    return normal_cdf(hi, mu, sigma) - normal_cdf(lo, mu, sigma)

# it's outside if it's not between
def normal_probability_outside(lo, hi, mu=0, sigma=1):
    return 1 - normal_probability_between(lo, hi, mu, sigma)
```
STATISTICAL HYPOTHESIS TESTING
EXAMPLE: FLIPPING A COIN

Example: \( n = 1000 \) coin flippings; \( p = 0.5 \)

for \( H_0 \): mean \( \mu = 50 \); standard deviation \( \sigma = 15.8 \);
significance(\text{Irrtumswahr\ss\text{e}nlichkeit}) = 5\%

\[
\mu_0, \sigma_0 = \text{normal\_approximation\_to\_binomial}(1000, 0.5)
\]

Errors: \[
\text{normal\_two\_sided\_bounds}(0.95, \mu_0, \sigma_0) \quad \# (469, 531)
\]

- **Type 1 error** \( (H_0 \) right, but rejected) \( \rightarrow \) significance
- **Type 2 error** \( (H_0 \) wrong, not rejected) \( \rightarrow \) power

\[
\text{# actual } \mu \text{ and } \sigma \text{ based on } p = 0.55
\]
\[
\mu_1, \sigma_1 = \text{normal\_approximation\_to\_binomial}(1000, 0.55)
\]

\[
\text{# a type 2 error means we fail to reject the null hypothesis}
\]
\[
\text{# which will happen when } X \text{ is still in our original interval}
\]
\[
\text{type\_2\_probability} = \text{normal\_probability\_between}(\text{lo}, \text{hi}, \mu_1, \sigma_1)
\]
\[
\text{power} = 1 - \text{type\_2\_probability} \quad \# 0.887
\]
STATISTICAL HYPOTHESIS TESTING
EXAMPLE: FLIPPING A COIN

P-Value:

For a two-sided test whether the coin is fair, we compute:

```python
def two_sided_p_value(x, mu=0, sigma=1):
    if x >= mu:
        # if x is greater than the mean, the tail is what's greater than x
        return 2 * normal_probability_above(x, mu, sigma)
    else:
        # if x is less than the mean, the tail is what's less than x
        return 2 * normal_probability_below(x, mu, sigma)
```

e.g. If we see 530 heads:
```python
two_sided_p_value(529.5, mu_0, sigma_0)  # 0.062
```
→greater than 5% significance: don’t reject the null $H_0$

e.g. If we see 532 heads:
```python
two_sided_p_value(531.5, mu_0, sigma_0)  # 0.0463
```
→smaller than 5% significance: reject the null $H_0$
CONFIDENCE INTERVAL

A confidence interval under a given statistical significance level $\alpha$ in an infinite number of independent experiments is a range of values, which contains in $(1- \alpha)$ percent of all observations the true value of the desired parameter $p$.

Typical values for $\alpha$: 0.05 or 0.01

Confidence level $= 1 - \alpha$
Flipping an unfair coin **1000** times

We observe **525** heads

⇒ estimate the probability for head $p_{est} = \frac{525}{1000} = 0.525$

$\alpha = 0.05$

**confidence level** $= 1 - \alpha = 0.95$

$mean = p_{est} = 0.525$

$\sigma = \text{sigma} = \sqrt{\frac{p_{est} \cdot (1 - p_{est})}{1000}} = \sqrt{\frac{0.525 \cdot (1 - 0.525)}{1000}} = 0.0158$
def normal_two_sided_bounds(probability, mu=0, sigma=1):
    """returns the symmetric (about the mean) bounds that contain the specified probability""
    tail_probability = (1 - probability) / 2
    # returns the z for which P(Z <= z) = probability
    upper_bound = inverse_normal_cdf(tail_probability, mu, sigma)
    # returns the z for which P(Z >= z) = probability
    lower_bound = inverse_normal_cdf(1 - tail_probability, mu, sigma)
    return lower_bound, upper_bound

>>> normal_two_sided_bounds(0.95, 0.525, 0.0158)
>>>(0.4940, 0.5560)
Proofed hypothesis:
- Political extremists only distinguish black / white
- Moderate students see more distinct shades of grey

First study: p < 0.01
Replicated study: p < 0.59

„False alarm rates“:
11% for p < 0.01, 29% for p < 0.05

P-Value was originally not intended as definite test

Source: Nature
http://www.nature.com/news/scientific-method-statistical-errors-1.14700
On a 0.05 significance level, in 1000 experiments, 46 times a fair coin is classified as unfair.
- Decide hypothesis before looking at data
- Use common sense to check hypothesis

- Don’t clean up data with hypothesis in mind, which would mean removing outliers that disprove the hypothesis
RUNNING AN A/B TEST

Experience optimization
→ trying to get people to click on advertisements

Example: new energy drink
→ choosing between advertisement A („tastes great!“) and advertisement B („less bias!“)

Experiment: randomly showing one of the two ads + tracking clicks

990 out of 1,000 A-viewers vs. 10 out of 1,000 B-viewers
→ A is the better ad

What if the differences are not so stark?
→ Statistical inference
RUNNING AN A/B TEST

NA people see ad A and nA of them click it.

Each ad view → Bernoulli trial where pA is the probability that someone clicks ad A.

nA / NA is approximately a normal random variable with mean pA and standard deviation $\sigma_A = \sqrt{p_A(1 - p_A) / N_A}$.

nB / NB is approximately a normal random variable with mean pB and standard deviation $\sigma_B = \sqrt{p_B(1 - p_B) / N_B}$.

def estimated_parameters(N, n):
    p = n / N
    sigma = math.sqrt(p * (1 - p) / N)
    return p, sigma
Assume those two normals are independent → their difference should also be normal with mean \( p_B - p_A \) and standard deviation \( \sqrt{\sigma_A^2 + \sigma_B^2} \).

Test the null hypothesis that \( p_A \) and \( p_B \) are the same (i.e. \( p_A - p_B \) is zero), using the statistic:

```python
def a_b_test_statistic(N_A, n_A, N_B, n_B):
    p_A, sigma_A = estimated_parameters(N_A, n_A)
    p_B, sigma_B = estimated_parameters(N_B, n_B)
    return (p_B - p_A) / math.sqrt(sigma_A ** 2 + sigma_B ** 2)
```
RUNNING AN A/B TEST

Ad A („tastes great!“) gets 200 clicks out of 1,000 views
Ad B („less bias!“) gets 180 clicks out of 1,000 views

\[ z = \text{a_b_test_statistic}(1000, 200, 1000, 180) \quad \# -1.14 \]

Probability of such a large difference if means were actually equal:

\[ \text{two_sided_p_value}(z) \quad \# 0.254 \]

→ large enough that you can’t conclude there’s much of a difference

Ad B („less bias!“) gets 150 clicks out of 1,000 views

\[ z = \text{a_b_test_statistic}(1000, 200, 1000, 150) \quad \# -2.94 \]
\[ \text{two_sided_p_value}(z) \quad \# 0.003 \]

→ only a 0.003 probability you’d see such a large difference if ads were equally effective
BAYESIAN INFERENCE

prior distribution of the parameters

using observed data and Bayes‘s Theorem

updated posterior distribution of the parameters
BAYESIAN INFERENCE

def B(alpha, beta):
    """a normalizing constant so that the total probability is 1"""
    return math.gamma(alpha) * math.gamma(beta) / math.gamma(alpha + beta)

def beta_pdf(x, alpha, beta):
    if x < 0 or x > 1:  # no weight outside of [0, 1]
        return 0
    return x ** (alpha - 1) * (1 - x) ** (beta - 1) / B(alpha, beta)

center of the weight at

alpha / (alpha + beta)

Figure 7-1. Example Beta distributions
Example: Flipping a Coin

prior distribution on $p$: assuming 55% heads
$\rightarrow$ $\rightarrow$ alpha = 55, beta = 45
$\rightarrow$ $\rightarrow$ Beta (55, 45), centered around 0.55

gathering data: flipping the coin 10 times
$\rightarrow$ $\rightarrow$ $h$ heads & $t$ tails
$\rightarrow$ $\rightarrow$ e.g. $h = 3$, $t = 7$

posterior distribution on $p$:
$\rightarrow$ $\rightarrow$ Beta with parameters alpha + $h$ & beta + $t$
$\rightarrow$ $\rightarrow$ Beta (58, 52), centered around 0.53
Figure 7-2. Posterior distributions arising from different priors.
THANK YOU FOR YOUR ATTENTION.

HOLY SHIT, MAN!! LOOK AT THIS!!

"STUDY FINDS 50% OF PEOPLE BORED BY STATISTICS."

Cyanide and Happiness © Explosm.net