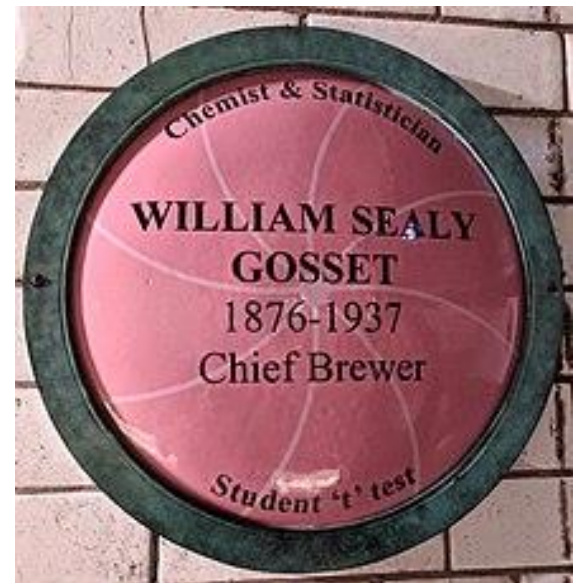


# Hypothesis Testing: Review



# Hypothesis Testing Framework

1. Formulate null-hypothesis
2. Collect data
3. Check model assumptions
4. Compute statistic
5. Compute p-value based on the statistic
6. (Optionally) check p-value against a threshold and reject the null hypothesis if the p-value is smaller than the threshold

# Binomial distribution

- Sample Null hypothesis: the probability of 1 is 0.5
  - Another null hypothesis could be that the probability of 1 is e.g. 0.2
- Model assumption: the trials are independent
  - If you got data that reads 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0 you might suspect something is amiss
  - Obviously the outcomes must be just 1s and 0s
- Statistic: number of 1's nH
- P-value for the null hypothesis that  $\text{Prob}(1) = \text{prob}$   
**`expect = size*prob`**  
**`pbinom(q = expect - abs(nH-expect), size =`**  
**`my.size, prob = prob) +`**  
**`1- pbinom(q = expect + abs(nH-expect) -1,`**  
**`size = my.size, prob = prob)`**
- Can also compute P-value using Gaussian approximation

# Normal Distribution, known s.d.

- Sample null hypothesis: the mean of the population is  $\mu = 0.7$
- Model assumption: the individuals in the population are normally distributed
  - Plot the sample (density, histogram, boxplot) to verify that the distribution is normal
    - Histogram and density approximately bell-shaped
    - Almost no datapoints outside of  $[\mu - 3\sigma, \mu + 3\sigma]$
    - Estimate  $\mu$  and  $\sigma$  using the sample mean  $\bar{x}$  and sample standard deviation  $s$
- Data: observations  $x_1, \dots, x_n$
- Statistic: the sample mean  $\bar{x} = \frac{x_1 + \dots + x_n}{n}$
- If the null hypothesis holds,  $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- P-value:  
**`2 * pnorm(mu - abs(mu - mean(x)), mean = mu, sd = sigma/sqrt(n))`**

# Normal distribution, unknown s.d.

- Sample null hypothesis: the mean of the population is  $\mu = 0.7$
- Model assumption: the individuals in the population are normally distributed with some standard deviation
  - Plot the sample (density, histogram, boxplot) to verify that the distribution is normal, with some standard deviation
    - Histogram and density approximately bell-shaped
    - Almost no datapoints outside of  $[\mu - 3\sigma, \mu + 3\sigma]$
    - Estimate  $\mu$  using the sample mean  $\bar{x}$
- Data: observations  $x_1, \dots, x_n$
- Statistic: the t-statistic  $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
- If the null hypothesis holds,  $t \sim t(n - 1)$
- P-value:  
 **$2 * pt(-abs(t), df = n - 1)$**

# Two samples from normal distributions

- Null hypothesis: the difference between the mean of population A and the mean of population B is 0
- Model assumption: the individuals in the two populations are normally distributed, possibly with different standard deviations
  - Plot the both samples (density, histogram, boxplot) to verify that the distributions are normal
    - Histograms and densities approximately bell-shaped
    - Almost no datapoints outside of  $[\mu_i - 3\sigma_i, \mu + 3\sigma_i]$
    - Estimate  $\mu_a, \mu_b, \sigma_a, \sigma_b$  using the sample means and sample standard deviations
- Data: two sets of observations
- Statistic: the t-statistic  $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
- If the null hypothesis holds,  $t \sim t(\nu), \nu = \dots$
- P-value:  
**`2 * pt(-abs(t), df = nu)`**

# P-values using fake data

- Simulate fake data that conforms to the assumption of the null hypothesis
- For each fake dataset, compute the statistic
- Compute the proportion of the time that the statistic for the fake dataset is more extreme than the statistic you actually observe in your data