"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." H. G. Wells

Credit: wikipedia
Objectives

- Hypothesis test
- Maximum Likelihood Estimation
A hypothesis

Ms. Smith’s vote percentage is 55%

This is what we want to test, often called null hypothesis $H_0$

Should we reject this hypothesis given the poll data?
Fraction of “less extreme” samples

- Assuming the hypothesis $H_0$ is true
- Define a test statistic

\[ x = \frac{(\text{sample mean}) - (\text{hypothesized value})}{\text{standard error}} \]

- Since $N>30$, we assume $x$ comes from a standard normal

- So, the fraction of “less extreme” samples is:

\[ f = \frac{1}{\sqrt{2\pi}} \int_{-|x|}^{|x|} \exp\left(-\frac{u^2}{2}\right)du \]
Rejection region of null hypothesis $H_0$

- Assuming the hypothesis $H_0$ is true
- Define a test statistic
  \[ x = \frac{(\text{sample mean}) - (\text{hypothesized value})}{\text{standard error}} \]
- Since $N>30$, assume $x$ comes from a standard normal

Credit: J. Orloff et al
It is conventional to report the p-value of a hypothesis test

\[ p = 1 - f = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|x|}^{|x|} \exp\left(-\frac{u^2}{2}\right) du \]

By convention:
\[ 2\alpha = 0.05 \]
That is:
If \( p < 0.05 \), reject \( H_0 \)
**p-value: election polling**

- \( H_0: \) Ms. Smith’s vote percentage is 55%
- The sample mean is 51% and stderr is 1.44%
- The test statistic \( x = \frac{51 - 55}{1.44} = -2.7778 \)
- And the p-value for the test is:
  \[
p = 1 - \frac{1}{\sqrt{2\pi}} \int_{-2.7778}^{2.7778} \exp(-\frac{u^2}{2})du = 0.00547 < 0.05
  \]
- So we reject the hypothesis
Q: what distribution should we use to test the hypothesis of sample mean if N<30?

A. Normal distribution
B. t-distribution with degree = 30
C. t-distribution with degree = N
D. t-distribution with degree = N-1
The use and misuse of p-value

- p-value use in scientific practice
  - Usually used to reject the null hypothesis that the data is random noise
  - Common practice is $p < 0.05$ is considered significant evidence for something interesting

- Caution about p-value hacking
  - Rejecting the null hypothesis doesn’t mean the alternative is true
  - $P < 0.05$ is arbitrary and often is not enough for controlling false positive phenomenon
Be wary of one tailed p-values

- The one tailed p-value should only be considered when the realized sample mean or differences will for sure fall only to one size of the distribution.

- Sometimes scientist are tempted to use one tailed test because it’ll give smaller p-val. But this is bad statistics!
Maximum likelihood estimation
Suppose we have a dataset that we know comes from a distribution (i.e. Binomial, Geometric, or Poisson, etc.)

What is the best estimate of the parameters ($\theta$ or $\theta$s) of the distribution?

Examples:
- For binomial and geometric distribution, $\theta = p$ (probability of success)
- For Poisson and exponential distributions, $\theta = \lambda$ (intensity)
- For normal distributions, $\theta$ could be $\mu$ or $\sigma^2$. 
Motivation: Poisson example

- Suppose we have data on the number of babies born each hour in a large hospital

<table>
<thead>
<tr>
<th>hour</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td># of babies</td>
<td>$k_1$</td>
<td>$k_2$</td>
<td>...</td>
<td>$k_N$</td>
</tr>
</tbody>
</table>

- We can assume the data comes from a Poisson distribution

- What is your best estimate of the intensity $\lambda$?

Credit: David Varodayan
Maximum likelihood estimation (MLE)

- We write the probability of seeing the data $D$ given parameter $\theta$

  \[ L(\theta) = P(D|\theta) \]

- The **likelihood function** $L(\theta)$ is **not** a probability distribution.

- The **maximum likelihood estimate (MLE)** of $\theta$ is

  \[ \hat{\theta} = \arg \max_\theta L(\theta) \]
Why is $L(\theta)$ not a probability distribution?

A. It doesn’t give the probability of all the possible $\theta$ values.

B. Don’t know whether the sum or integral of $L(\theta)$ for all possible $\theta$ values is one or not.

C. Both.
Likelihood function: Binomial example

- Suppose we have a coin with unknown probability of coming up heads
- We toss it \( N \) times and observe \( k \) heads
- We know that this data comes from a binomial distribution
- What is the likelihood function \( L(\theta) = P(D|\theta) \) ?
Likelihood function: binomial example

- Suppose we have a coin with unknown probability of coming up heads
- We toss it \( N \) times and observe \( k \) heads
- We know that this data comes from a binomial distribution
- What is the likelihood function \( L(\theta) = P(D|\theta) \)?

\[
L(\theta) = \binom{N}{k} \theta^k (1 - \theta)^{N-k}
\]
MLE derivation: binomial example

\[ L(\theta) = \binom{N}{k} \theta^k (1 - \theta)^{N-k} \]

In order to find: \[ \hat{\theta} = \arg \max_{\theta} L(\theta) \]

We set: \[ \frac{dL(\theta)}{d\theta} = 0 \]
MLE derivation: binomial example

\[ L(\theta) = \binom{N}{k} \theta^k (1 - \theta)^{N-k} \]
MLE derivation: binomial example

\[ L(\theta) = \binom{N}{k} \theta^k (1 - \theta)^{N-k} \]

\[
\frac{d}{d\theta} L(\theta) = \binom{N}{k} (k\theta^{k-1}(1 - \theta)^{N-k} - \theta^k(N - k)(1 - \theta)^{N-k-1}) = 0
\]
MLE derivation: binomial example

\[ L(\theta) = \binom{N}{k} \theta^k (1 - \theta)^{N-k} \]

\[
\frac{d}{d\theta} L(\theta) = \binom{N}{k} \left( k\theta^{k-1} (1 - \theta)^{N-k} - \theta^k (N - k)(1 - \theta)^{N-k-1} \right) = 0
\]

\[ k\theta^{k-1} (1 - \theta)^{N-k} = \theta^k (N - k)(1 - \theta)^{N-k-1} \]
MLE derivation: binomial example

\[ L(\theta) = \binom{N}{k} \theta^k (1 - \theta)^{N-k} \]

\[
\frac{d}{d\theta} L(\theta) = \binom{N}{k} (k\theta^{k-1}(1 - \theta)^{N-k} - \theta^k (N - k)(1 - \theta)^{N-k-1}) = 0
\]

\[
k\theta^{k-1}(1 - \theta)^{N-k} = \theta^k (N - k)(1 - \theta)^{N-k-1}
\]

\[
k - k\theta = N\theta - k\theta
\]
MLE derivation: binomial example

\[
L(\theta) = \binom{N}{k} \theta^k (1 - \theta)^{N-k}
\]

\[
\frac{d}{d\theta} L(\theta) = \binom{N}{k} \left( k\theta^{k-1} (1 - \theta)^{N-k} - \theta^k (N - k)(1 - \theta)^{N-k-1} \right) = 0
\]

\[
k\theta^{k-1} (1 - \theta)^{N-k} = \theta^k (N - k)(1 - \theta)^{N-k-1}
\]

\[
k - k\theta = N\theta - k\theta
\]

\[
\hat{\theta} = \frac{k}{N}
\]

The MLE of \( p \)
Likelihood function: geometric example

- Suppose we have a die with unknown probability of coming up six
- We roll it and it comes up six for the first time on the kth roll
- We know that this data comes from a geometric distribution
- What is the likelihood function \( L(\theta) = P(D|\theta) \)? Assume \( \theta \) is \( p \).
MLE derivation: geometric example

\[ L(\theta) = (1 - \theta)^{k-1}\theta \]
MLE derivation: geometric example

\[ L(\theta) = (1 - \theta)^{k-1}\theta \]

\[ \frac{d}{d\theta} L(\theta) = (1 - \theta)^{k-1} - (k - 1)(1 - \theta)^{k-2}\theta = 0 \]
MLE derivation: geometric example

\[ L(\theta) = (1 - \theta)^{k-1} \theta \]

\[ \frac{d}{d\theta} L(\theta) = (1 - \theta)^{k-1} - (k - 1)(1 - \theta)^{k-2} \theta = 0 \]

\[ (1 - \theta)^{k-1} = (k - 1)(1 - \theta)^{k-2} \theta \]
MLE derivation: geometric example

\[ L(\theta) = (1 - \theta)^{k-1}\theta \]

\[
\frac{d}{d\theta} L(\theta) = (1 - \theta)^{k-1} - (k - 1)(1 - \theta)^{k-2}\theta = 0
\]

\[
(1 - \theta)^{k-1} = (k - 1)(1 - \theta)^{k-2}\theta
\]

\[
1 - \theta = k\theta - \theta
\]
MLE derivation: geometric example

\[ L(\theta) = (1 - \theta)^{k-1} \theta \]

\[
\frac{d}{d\theta} L(\theta) = (1 - \theta)^{k-1} - (k - 1)(1 - \theta)^{k-2}\theta = 0
\]

\[ (1 - \theta)^{k-1} = (k - 1)(1 - \theta)^{k-2}\theta \]

\[ 1 - \theta = k\theta - \theta \]

\[ \hat{\theta} = \frac{1}{k} \quad \text{The MLE of } p \]
MLE with data from IID trials

- If the dataset $D = \{x\}$ comes from IID trials

\[ L(\theta) = P(D|\theta) = \prod_{x_i \in D} P(x_i|\theta) \]

- Each $x_i$ is one observed result from an IID trial
Q: MLE with data from IID trials

If the dataset \( D = \{x\} \) comes from IID trials

\[
L(\theta) = P(D|\theta) = \prod_{x_i \in D} P(x_i|\theta)
\]

Why is the above function defined by the product?

A. IID samples are independent

B. Each trial has identical probability function

C. Both.
MLE with data from IID trials

• If the dataset $D = \{x\}$ comes from IID trials

\[
L(\theta) = P(D|\theta) = \prod_{x_i \in D} P(x_i|\theta)
\]

• The likelihood function is hard to differentiate in general, except for the binomial and geometric cases.

• Clever trick: take the (natural) log
Log-likelihood function

Since log is a strictly increasing function

\[
\hat{\theta} = \arg \max_\theta L(\theta) = \arg \max_\theta \log L(\theta)
\]

So we can aim to maximize the log-likelihood function

\[
\log L(\theta) = \log P(D|\theta) = \log \prod_{x_i \in D} P(x_i|\theta) = \sum_{x_i \in D} \log P(x_i|\theta)
\]

The log-likelihood function is usually much easier to differentiate
Suppose we have data on the number of babies born each hour in a large hospital.

<table>
<thead>
<tr>
<th>hour</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td># of babies</td>
<td>$k_1$</td>
<td>$k_2$</td>
<td>...</td>
<td>$k_N$</td>
</tr>
</tbody>
</table>

We can assume the data comes from a Poisson distribution $\lambda$.

What is the log likelihood function $\text{Log}L(\theta)$?
Log-likelihood function: Poisson example

\[ L(\theta) = \prod_{i=1}^{N} \frac{e^{-\theta} \theta^{k_i}}{k_i!} \]

\[ \log L(\theta) = \log \left( \prod_{i=1}^{N} \frac{e^{-\theta} \theta^{k_i}}{k_i!} \right) = \sum_{i=1}^{N} \log \left( \frac{e^{-\theta} \theta^{k_i}}{k_i!} \right) \]

\[ = \sum_{i=1}^{N} \left( -\theta + k_i \log \theta - \log k_i! \right) \]
MLE: Poisson example

\[ LogL(\theta) = \sum_{i=1}^{N} (-\theta + k_i \log \theta - \log k_i!) \]
MLE: Poisson example

\[ \log L(\theta) = \sum_{i=1}^{N} \left( -\theta + k_i \log \theta - \log k_i! \right) \]

\[ \frac{d}{d\theta} \log L(\theta) = 0 \implies \sum_{i=1}^{N} \left( -1 + \frac{k_i}{\theta} - 0 \right) = 0 \]
MLE : Poisson example

\[ \text{Log} L(\theta) = \sum_{i=1}^{N} (-\theta + k_i \log \theta - \log k_i!) \]

\[ \frac{d}{d\theta} \log L(\theta) = 0 \Rightarrow \sum_{i=1}^{N} (-1 + \frac{k_i}{\theta} - 0) = 0 \]

\[ -N + \frac{\sum_{i}^{N} k_i}{\theta} = 0 \]
MLE : Poisson example

\[ LogL(\theta) = \sum_{i=1}^{N} (-\theta + k_i \log \theta - \log k_i!) \]

\[ \frac{d}{d\theta} \log L(\theta) = 0 \Rightarrow \sum_{i=1}^{N} \left(-1 + \frac{k_i}{\theta} - 0\right) = 0 \]

\[ -N + \frac{\sum_{i}^{N} k_i}{\theta} = 0 \]

\[ \hat{\theta} = \frac{\sum_{i}^{N} k_i}{N} \]

The MLE of \( \lambda \)
MLE for normal distribution

Suppose we model the dataset $D = \{x\}$ as normally distributed.

What should be the likelihood function? Is the method of modeling the same as for the Poisson distribution?

A. Yes  B. No
MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed

- What should be the likelihood function? Is the method of modeling the same as for the Poisson distribution? **Yes and No.** The idea is similar but the normal distribution is continuous, we need to use the **probability density** instead.
MLE for normal distribution

Suppose we model the dataset $D = \{x\}$ as normally distributed.

The likelihood function of a normal distribution:

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$
Suppose we model the dataset $D = \{x\}$ as normally distributed.

There are two parameters to estimate: $\mu$ and $\sigma$.

If we fix $\sigma$ and set $\theta = \mu$,
$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

If we fix $\mu$ and set $\theta = \sigma$,
$$\hat{\theta} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$
Drawbacks of MLE

- Maximizing some likelihood or log-likelihood function is mathematically hard

- If there are very few data items, the MLE estimate maybe very unreliable
  - If we observe 3 heads in 10 coin tosses, should we accept that $p(\text{heads})= 0.3$?
  - If we observe 0 heads in 2 coin tosses, should we accept that $p(\text{heads})= 0$?
Confidence intervals for MLE estimates

- An MLE parameter estimate \( \hat{\theta} \) depends on the data that was observed.

- We can construct a confidence interval for \( \hat{\theta} \) using the parametric bootstrap:
  - Use the distribution with parameter \( \hat{\theta} \) to generate a large number of bootstrap samples.
  - From each “synthetic” dataset, re-estimate the parameter using MLE.
  - Use the histogram of these re-estimates to construct a confidence interval.
Assignments

- Finish Chapter 7 of the textbook
- Next time: Maximum likelihood estimate, Bayesian inference
Additional References


Morris H. Degroot and Mark J. Schervish "Probability and Statistics"
Chi-square distribution

- If $Z_i$ are independent variables of standard normal distribution, 
  \[ X = Z_1^2 + Z_2^2 + \ldots + Z_m^2 = \sum_{i=1}^{m} Z_i^2 \]

  has a Chi-square distribution with degree of freedom $m$, $X \sim \chi^2(m)$

- We can test the goodness of fit for a model using a statistic $C$ against this distribution, where

  \[
  C = \sum_{i=1}^{m} \frac{(f_0(\varepsilon_i) - f_t(\varepsilon_i))^2}{f_t(\varepsilon_i)}
  \]
Independence analysis using Chi-square

Given the two way table, test whether the column and row are independent

<table>
<thead>
<tr>
<th></th>
<th>Boy</th>
<th>Girl</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grades</td>
<td>117</td>
<td>130</td>
<td>247</td>
</tr>
<tr>
<td>Popular</td>
<td>50</td>
<td>91</td>
<td>141</td>
</tr>
<tr>
<td>Sports</td>
<td>60</td>
<td>30</td>
<td>90</td>
</tr>
<tr>
<td>Total</td>
<td>227</td>
<td>251</td>
<td>478</td>
</tr>
</tbody>
</table>
Independence analysis using Chi-square

The theoretical expected values if independent

<table>
<thead>
<tr>
<th></th>
<th>Boy</th>
<th>Girl</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grades</td>
<td>117.29916</td>
<td>129.70084</td>
<td>247</td>
</tr>
<tr>
<td>Popular</td>
<td>66.96025</td>
<td>74.03975</td>
<td>141</td>
</tr>
<tr>
<td>Sports</td>
<td>42.74059</td>
<td>47.25941</td>
<td>90</td>
</tr>
<tr>
<td>Total</td>
<td>227</td>
<td>251</td>
<td>478</td>
</tr>
</tbody>
</table>
The degree of the chi-square distribution for the two way table

The degree of freedom for the chi-square distribution for a $r$ by $c$ table is

$$(r-1) \times (c-1) \text{ where } r>1 \text{ and } c>1$$

Because the degree $df = n-1-p$

\[
= r - 1 - (r-1) - (c-1) \\
= (r-1) \times (c-1) \\
= 2
\]

$n$ is the number of cells of data;
$p$ is the number of unknown parameters

See textbook Pg 171-172
Chi-square test for the popular kid data

The Chi-statistic: 21.455

chisq.test(data_BG)

Pearson's Chi-squared test

data: data_BG
X-squared = 21.455, df = 2, p-value = 2.193e-05

P-value: 2.193e-05

It’s very unlikely the two categories are independent
Q. What is the degree of freedom for this?

The following 2-way table for chi-square test has a degree of freedom equal to:

<table>
<thead>
<tr>
<th></th>
<th>Number of lectures attended</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Freshmen</td>
<td>10</td>
</tr>
<tr>
<td>Sophomores</td>
<td>14</td>
</tr>
<tr>
<td>Juniors</td>
<td>15</td>
</tr>
<tr>
<td>Seniors</td>
<td>19</td>
</tr>
</tbody>
</table>

A. 20  B. 9  C. 12  D. 4
Chi-square test is very versatile

- Chi-square test is so versatile that it can be utilized in many ways either for discrete data or continuous data via intervals.

- Please check out the worked-out examples in the textbook and read more about its applications.
We are interested in comparing sample means

- Are the average daily body temperature of the two beavers the same?
- We need to model the difference between two sample means.
How do we model the difference between two samples means?

- We know when the sample size $N$ is large, the sample mean random variable approaches normal 

- So our problem became finding the model of the difference between two normally distributed random variables.

* Assume the daily temperature at different times are independent.
Background: sum of independent normals

- We know
  \[ X_1 \sim \text{normal}(\mu_1, \sigma_1^2) \]
  \[ X_2 \sim \text{normal}(\mu_2, \sigma_2^2) \]
  \[ X_1 + X_2 \sim ? \]

- The sum of \( X_1 \) and \( X_2 \) is still normal (proof omitted, ref. ...)
Background: sum of independent normals

We know

\[ X_1 \sim normal(\mu_1, \sigma_1^2) \]

\[ X_2 \sim normal(\mu_2, \sigma_2^2) \]

So \[ X_1 + X_2 \sim normal(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \]

By the linearity of expected value and the sum rule of variance of the sum of two independent random variables.
Background: sum of independent normals

☀️ We know

\[ X_1 \sim \text{normal}(\mu_1, \sigma^2_1) \]

\[ X_2 \sim \text{normal}(\mu_2, \sigma^2_2) \]

☀️ So

\[ X_1 + X_2 \sim \text{normal}(\mu_1 + \mu_2, \sigma^2_1 + \sigma^2_2) \]

☀️ By properties:

\[ E[X_1 + X_2] = E[X_1] + E[X_2] \]

\[ \text{var}[X_1 + X_2] = \text{var}[X_1] + \text{var}[X_2] \]
Difference of independent normals

We know

\[ X_1 \sim normal(\mu_1, \sigma_1^2) \]

\[ X_2 \sim normal(\mu_2, \sigma_2^2) \]

\[ X_1 - X_2 \sim ? \]

The difference of \( X_1 \) and \( X_2 \) is still normal (proof omitted)
Difference of independent normals

We know

\[ X_1 \sim \text{normal}(\mu_1, \sigma_1^2) \]

\[ X_2 \sim \text{normal}(\mu_2, \sigma_2^2) \]

So

\[ X_1 - X_2 \sim \text{normal}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2) \]

By the linearity of expected value and the sum rule of variance of the sum of two independent random variables and the scaling property of variance.
Derivation of the mean and variance of difference of independent normals

Because

\[ X_1 - X_2 \sim normal(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2) \]
Derivation of the mean and variance of difference of independent normals

Because

\[ E[X_1 - X_2] = E[X_1] - E[X_2] \]
\[ = \mu_1 - \mu_2 \]
Derivation of the mean and variance of difference of independent normals

Because

\[ E[X_1 - X_2] = E[X_1] - E[X_2] \]
\[ = \mu_1 - \mu_2 \]

\[ var[X_1 - X_2] = var[X_1 + (-X_2)] \]
Derivation of the mean and variance of difference of independent normals

Because

\[ E[X_1 - X_2] = E[X_1] - E[X_2] \]
\[ = \mu_1 - \mu_2 \]

\[ \text{var}[X_1 - X_2] = \text{var}[X_1 + (-X_2)] \]
\[ = \text{var}[X_1] + \text{var}[-X_2] \]
Because

\[ E[X_1 - X_2] = E[X_1] - E[X_2] \]
\[ = \mu_1 - \mu_2 \]

\[ \text{var}[X_1 - X_2] = \text{var}[X_1 + (-X_2)] \]
\[ = \text{var}[X_1] + \text{var}[-X_2] \]
\[ = \text{var}[X_1] + \text{var}[X_2] \]

\[ \text{var}[c \cdot X_2] = c^2 \text{var}[X_2] \]
Derivation of the mean and variance of difference of independent normals

Because

\[ E[X_1 - X_2] = E[X_1] - E[X_2] \]
\[ = \mu_1 - \mu_2 \]

\[ \text{var}[X_1 - X_2] = \text{var}[X_1 + (-X_2)] \]
\[ = \text{var}[X_1] + \text{var}[-X_2] \]
\[ = \text{var}[X_1] + \text{var}[X_2] \]
\[ = \sigma_1^2 + \sigma_2^2 \]
Because

\[ E[X_1 - X_2] = E[X_1] - E[X_2] \]
\[ = \mu_1 - \mu_2 \]

\[ \text{var}[X_1 - X_2] = \text{var}[X_1 + (-X_2)] \]
\[ = \text{var}[X_1] + \text{var}[-X_2] \]
\[ = \text{var}[X_1] + \text{var}[X_2] \]
\[ = \sigma_1^2 + \sigma_2^2 \]

\[ X_1 - X_2 \sim \text{normal}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2) \]
Now we are ready to check the differences between sample means.

Because sample means are roughly normal when \( N \) is large.

\[
X_1 - X_2 \sim \text{normal}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)
\]
The difference between two sample means

- Suppose we draw samples from two populations \( \{x\} \) and \( \{y\} \)
  - From a sample of size \( k_x \) from \( \{x\} \), we get sample mean \( X^{(k_x)} \)
  - From a sample of size \( k_y \) from \( \{y\} \), we get sample mean \( Y^{(k_y)} \)
The difference between two sample means

- Define random variable $D = X^{(k_x)} - Y^{(k_y)}$ as the difference between the sample means.

- If we hypothesize that $\text{popmean}({x}) = \text{popmean}({y})$, then

$$E[D] = E[X^{(k_x)}] - E[Y^{(k_y)}] = 0$$
Standard error of the difference between two sample means

Recall the standard error is roughly the standard deviation of a sample mean

By the property of variance of the difference between two independent normals

\[
\text{var}[D] = \text{stderr}(\{x\})^2 + \text{stderr}(\{y\})^2
\]

\[
\text{std}[D] = \sqrt{\text{stderr}(\{x\})^2 + \text{stderr}(\{y\})^2} = \text{stderr}[D]
\]

\[
\text{std}[D] = \sqrt{\frac{\text{stdunbiased}(\{x\})^2}{k_x} + \frac{\text{stdunbiased}(\{y\})^2}{k_y}}
\]
P-value for testing the equality of two means

Define the test statistic

\[ g = \frac{\text{mean}\{x\} - \text{mean}\{y\}}{\text{stderr}(D)} \]

If \( k_x \geq 30 \) and \( k_y \geq 30 \)

\[ p = 1 - f = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|g|}^{|g|} \exp\left(-\frac{u^2}{2}\right) du \]
P-value: Rejection region- “The extreme fraction”

- It is conventional to report the p-value of a hypothesis test:

\[ p = 1 - f = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|g|}^{|g|} \exp\left(-\frac{u^2}{2}\right) du \]

- Since \( N>30 \), \( x \) should come from a standard normal distribution.

Rejection region (2\(\alpha\))

By convention:
\( 2\alpha = 0.05 \)
That is:
If \( p < 0.05 \), reject \( H_0 \)
Comparing the body temperatures of two beavers

$k_x = 114$ and $k_y = 100$

Mean($\{x\}$) = 36.86219

Mean($\{y\}$) = 37.5967

$\text{stderr}(\{x\}) = \frac{\text{stdunbiased}(\{x\})}{\sqrt{114}}$

$\text{stderr}(\{y\}) = \frac{\text{stdunbiased}(\{y\})}{\sqrt{100}}$

$\text{stderr}(D) = \sqrt{\text{stderr}(\{x\})^2 + \text{stderr}(\{y\})^2}$

$= 0.04821181$
Comparing the body temperatures of two beavers

- Hypothesis $H_0$: the mean temperatures of the two beavers are the same

- The test statistic $g = \frac{36.86219 - 37.5967}{0.04821181} = -15.235$

\[
p = 1 - f = 1 - \frac{1}{\sqrt{2\pi}} \int_{-15.235}^{15.235} \exp\left(-\frac{u^2}{2}\right) du
\]

$p \approx 0$

- So we can reject the hypothesis that the mean temperatures are the same
What if N < 30?

- There are general solutions for either N >= 30 or N < 30 if the data sets are random samples from normal distributed data.

- The difference between sample means can be either modeled as t-distribution with degree \((k_x + k_y - 2)\) when their population standard deviations are the same.

- Or the difference between sample means can be approximated with t-distribution with other proper degree of freedom.

- There are built-in t-test procedures in Python, R...
Compare the two mean temperatures of two beavers with \texttt{t.test}

- Hypothesis $H_0$: the mean temperatures of the two beavers are the same

```r
> t.test(beaver1$temp, beaver2$temp)

Welch Two Sample t-test

data:  beaver1$temp and beaver2$temp
t = -15.235, df = 131.12, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -0.8298806  -0.6391334
sample estimates:
mean of x mean of y
36.86219   37.59670
```

- $p < 2.2e-16$, also reject the hypothesis
See you next time

See You!