HW4 has been posted.
Two-sided Confidence Interval on the Population Mean, $\mu$, Variance, $\sigma$, is known

$$x - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < x + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$f(z) \sim \exp(-z^2/2)$$
Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

\[
\bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2, n-1} \frac{\sigma}{\sqrt{n}}
\]

\[
f(t) \sim \left(1 + \frac{t^2}{n - 1}\right)^{-n/2}
\]
Hypothesis testing: 
one sample
Is P53 gene expressed at a lower level in cancer patients than in healthy people?

- We are interested if a P53 gene expression is lowered in population of cancer patients compared to the healthy population.

- We know that mean gene expression in the healthy population is $\mu_h=50$ mRNAs/cell. We are interested in deciding whether or not the mean expression in cancer population is lower than in healthy population. Let’s call hypothesis $H_1$. Here $H_1$ is one-sided.

- If we asked: cancer is not equal to healthy $H_1$ would be a two-sided hypothesis.

- Assume we have a sample of 100 cancer patients with sample mean $\bar{x}=48$ mRNAs/cell and standard deviation $\sigma=10$ mRNA/cell.

- Can we use our sample to reject the “business as usual” or null hypothesis $H_0$: cancer = healthy and select one-sided hypothesis $H_1$: cancer < healthy.
Two types of errors

<table>
<thead>
<tr>
<th></th>
<th>decide $H_0$</th>
<th>decide $H_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>true $H_0$ probability</td>
<td>Correct action $1 - \alpha$</td>
<td>Type I error $\alpha$</td>
</tr>
<tr>
<td>true $H_1$ probability</td>
<td>Type II error $\beta$</td>
<td>Correct action $\text{power} = 1 - \beta$</td>
</tr>
</tbody>
</table>

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

Sometimes the **type I error probability $\alpha$** is called the **significance level**, or the **$\alpha$-error**

**Instructions:** get $\alpha$ from your boss or PI (e.g., 5% or 1%)

$\text{Prob}(H_0 \text{ is true given the sample data}) < \alpha$  
$\rightarrow$ reject $H_0$ and accept $H_1$

$\text{Prob}(H_0 \text{ is true given the sample data}) > \alpha$  
$\rightarrow$ accept $H_0$ and reject $H_1$

**Type II error** is much harder to estimate. Will deal with it later
P-Values of Hypothesis Tests

• **P-value**: what is the probability to get the observed value of sample mean of \( \bar{x} =48 \text{ mRNAs/cell} \) (or even smaller) and \( \sigma=10 \text{ mRNAs/cell} \) in a healthy population with \( \mu_h=50 \text{ mRNAs/cell} \)

• If **P-value** is small – the null hypothesis is likely wrong and thus, the probability of making a type I error (incorrectly rejecting the null hypothesis) is small

• P-value answers the question: if I reject the null hypothesis \( H_0 \) based on the sample, what is the probability that I am making a type I error?
P-Value vs \( \alpha \) in Hypothesis Testing

- Problem with using a predefined \( \alpha \): you don’t know by how much you exceeded it
- Another approach is to calculate \( \text{Prob}(H_0 \text{ is true given the sample data}) \) referred to as P-value. It the smallest \( \alpha \) that would lead to rejection of null hypothesis
- You give your boss the P-value and let him/her decide if it is good enough
- Routinely with big datasets in genomics and systems biology P-values can be \( 10^{-\text{large number}} \sim 10^{-100} \). This number is used to judge the quality of the hypothesis
\[ \mu_h = 50 \]
\[ H_0: \mu_c = \mu_h \]

One-sided hypothesis \( H_1: \mu_c < \mu_h \)

\[ n = 100, \quad X = 48, \quad s = 10 \]

\[ Z_0 = 1.64 \]

\[ \text{Area} = 0.05 \]

\[ \mu_h = 50 \]

\[ X = 48 \]

\[ Z_X = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{48 - 50}{10 / \sqrt{100}} = 1 \]

\[ P\text{-Value}(X = 48 \mid H_0: \mu_c = \mu_h) = \text{Prob}(X \leq 48) = 2.5\% = 0.025 \]
\( \mu_h = 50 \)

\[ H_0: \mu_c = \mu_h \]

\[ n = 100, \ X = 48, \ \sigma = 10 \]

\[ H_1: \mu_c < \mu_h \]

Set p-value threshold: \( \alpha = 5\% \)

\( \bar{x} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1 \)

\( \beta = P(\text{Accept } H_0 \mid H_1 \text{ is true}) = \int_{-\infty}^{48.36} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-47)^2}{2}\right) dx = 8.8\% \)

\( \alpha = 1 - 0.05 = 5\% \)
Generalizations

• What if $H_1$ is a two-sided hypothesis?
  
  A: P-value is $2(1-\Phi(|Z|))$, where $Z=(\bar{X}-\mu_0)/[S/\sqrt{n}]$.
  
  Compare it to: For one sized $\mu_1 > \mu_0$ it is $1-\Phi(Z)$
  For one sized $\mu_1 < \mu_0$ it is $\Phi(Z)$.

• If $\alpha$ is given, use $\mu_0 +/\!/- z_{\alpha/2} \times S$ as thresholds to reject the null hypothesis.

• What if the sample size $n$ is small (say $n<10$):
  
  A: Use t-distribution with $n-1$ degrees of freedom for 2-sided P-value $=2(1-CDF_{Tdist}(|T|))$ where $T=(\bar{X}-\mu_0)/[S/\sqrt{n}]$.
  
  For given $\alpha$ use $\mu_0 +/\!/- t_{\alpha/2,n-1} \times T$ to reject the null hypothesis.
Type II Error and Choice of Sample Size

Assume you know the minimum $\delta = |\mu_1 - \mu_0|$ that you care about. What is the minimal sample you should use to separate $H_0$ and $H_1$ hypotheses if your tolerance to type I and type II errors is $\alpha$ and $\beta$?

$$\frac{\delta \sqrt{n}}{\sigma} = z_{\alpha/2} + z_\beta$$

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2} \quad \text{where} \quad \delta = \mu - \mu_0$$ (9-22)
Standard notation to indicate P-value with *, **, ***

Table 11.1: A commonly adopted convention for reporting $p$ values: in many places it is conventional to report one of four different things (e.g., $p < .05$) as shown below. I’ve included the “significance stars” notation (i.e., a * indicates $p < .05$) because you sometimes see this notation produced by statistical software. It’s also worth noting that some people will write n.s. (not significant) rather than $p > .05$.

<table>
<thead>
<tr>
<th>Usual notation</th>
<th>Signif. stars</th>
<th>English translation</th>
<th>The null is...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p &gt; .05$</td>
<td></td>
<td>The test wasn’t significant</td>
<td>Retained</td>
</tr>
<tr>
<td>$p &lt; .05$</td>
<td>*</td>
<td>The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$.</td>
<td>Rejected</td>
</tr>
<tr>
<td>$p &lt; .01$</td>
<td>**</td>
<td>The test was significant at $\alpha = .05$ but not at $\alpha = .001$.</td>
<td>Rejected</td>
</tr>
<tr>
<td>$p &lt; .001$</td>
<td>***</td>
<td>The test was significant at all levels</td>
<td>Rejected</td>
</tr>
</tbody>
</table>
Happy Halloween!
(belated)

Credit: Trust me,
I’m a “Biologist”
Facebook community

Credit: XKCD
comics

\[ P = 0.06 \]
A peculiar prevalence of $p$ values just below .05

E. J. Masicampo¹, and Daniel R. Lalande²

¹Department of Psychology, Wake Forest University, Winston-Salem, NC, USA
²Department of Health Sciences, Université du Québec à Chicoutimi, Chicoutimi, QC, Canada

MASICAMPO AND LALANDE
Hypothesis testing: two samples
10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

**Figure 10-1** Two independent populations.

- Population 1
  - \( \mu_1 \)
  - \( \sigma_1^2 \)
  - Sample 1: \( x_{11}, x_{12}, \ldots, x_{1n_1} \)

- Population 2
  - \( \mu_2 \)
  - \( \sigma_2^2 \)
  - Sample 2: \( x_{21}, x_{22}, \ldots, x_{2n_2} \)
10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

Assumptions

1. $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from population 1.
2. $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample from population 2.
3. The two populations represented by $X_1$ and $X_2$ are independent.
4. Both populations are normal.

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$
10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

The quantity

\[ Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \]  

(10-1)

has a \( N(0, 1) \) distribution.
10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

10-2.1 Hypothesis Tests for a Difference in Means, Variances Known

Null hypothesis: \( H_0: \mu_1 - \mu_2 = \Delta_0 \)

Test statistic: \( Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \) \hspace{1cm} (10-2)

<table>
<thead>
<tr>
<th>Alternative Hypotheses</th>
<th>( P )-Value</th>
<th>Rejection Criterion For Fixed-Level Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_1: \mu_1 - \mu_2 \neq \Delta_0 )</td>
<td>Probability above (</td>
<td>z_0</td>
</tr>
<tr>
<td>( H_1: \mu_1 - \mu_2 &gt; \Delta_0 )</td>
<td>Probability above (z_0), (P = 1 - \Phi(z_0))</td>
<td>(z_0 &gt; z_{\alpha})</td>
</tr>
<tr>
<td>( H_1: \mu_1 - \mu_2 &lt; \Delta_0 )</td>
<td>Probability below (z_0), (P = \Phi(z_0))</td>
<td>(z_0 &lt; -z_{\alpha})</td>
</tr>
</tbody>
</table>

\(\Delta_0 = 0\) usually
10-2.1 Hypotheses Tests on the Difference in Means, Variances Unknown

Case 2: $\sigma_1^2 \neq \sigma_2^2$

If $H_0: \mu_1 - \mu_2 = \Delta_0$ is true, the statistic

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

is distributed as t-distribution with degrees of freedom given by

$$\nu = n_1 + n_2 - 2,$$

or more generally
Plasma antioxidants from chocolate

Dark chocolate may offer its consumers health benefits the milk variety cannot match.

There is some speculation that dietary flavonoids from chocolate, in particular (−)epicatechin, may promote cardiovascular health as a result of direct antioxidant effects or through antithrombotic mechanisms1−3. Here we show that consumption of plain, dark chocolate (Fig. 1) results in an increase in both the total antioxidant capacity and the (−)epicatechin content of blood plasma, but that these effects are markedly reduced when the chocolate is consumed with milk or if milk is incorporated as milk chocolate. Our findings indicate that milk may interfere with the absorption of antioxidants from chocolate in vivo and may therefore negate the potential health benefits that can be derived from eating moderate amounts of dark chocolate.

To determine the antioxidant content of different chocolate varieties, we took dark chocolate and milk chocolate prepared from the same batch of cocoa beans and defatted them with n-hexane before extracting them with a mixture of water, acetone and acetic acid (70.0:29.8:0.2 by volume). We measured their in vitro total antioxidant capacities using the ferric-reducing antioxidant potential (FRAP) assay4; FRAP reduced iron per 100 g for dark and milk chocolate, respectively. Volunteers must therefore consume twice as much milk chocolate as dark chocolate to receive a similar intake of antioxidants.

We recruited 12 healthy volunteers (7 women and 5 men with an average age of 32.2 ± 1.0 years (range, 25–35 years). Subjects were non-smokers, had normal blood lipid levels, were taking no drugs or vitamin supplements, and had an average weight of 65.8 ± 3.1 kg (range, 46.0–86.0 kg) and body-mass index of 21.9 ± 0.4 kg m⁻² (range, 18.6–23.6 kg m⁻²). On different days, following a crossover experimental design, subjects consumed 100 g dark chocolate, 100 g dark chocolate with 200 ml full-fat milk, or 200 g milk chocolate (containing the equivalent of up to 40 ml milk).

One hour after subjects had ingested the chocolate, or chocolate and milk, we measured the total antioxidant capacity of their plasma by FRAP assay. Plasma antioxidant levels increased significantly after consumption of dark chocolate alone, from 100 ± 3.5% to 118.4 ± 3.5% (t-test, P < 0.001), returning to baseline values (95.4 ± 3.6%) after 4 h (Fig. 2a). There was no significant difference in antioxidant capacity between the two chocolate treatments (Fig. 2a).

Mauro Serafini*, Rossana Bugiani*, Giuseppe Maiani*, Silvia Valtuena*, Simone De Santis*, Alan Crozier†

*Antioxidant Research Laboratory, Unit of Human Nutrition, National Institute for Food and Nutrition Research, Via Ardeatina 546, 00178 Rome, Italy
e-mail: seraﬁni@inran.it
†Plant Products and Human Nutrition Group, Graham Kerr Building, Division of Biochemistry and Molecular Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK

Figure 1 Stack of benefits? Unlike its milky counterpart, dark chocolate may provide more than just a treat for the taste buds.
Sweet matlab exercise #1

- Download `dark_vs_milk_chocolate_analysis_template.m` at the course website. **Correct all ?? In the file**
- `dark=[118.8 122.6 115.6 113.6 119.5 115.9 115.8 115.1 116.9 115.4 115.6 107.9];`
- `milk=[102.1 105.8 99.6 102.7 98.8 100.9 102.8 98.7 94.7 97.8 99.7 98.6]`
- Use Z-statistics to calculate *P-value* of the null hypothesis $H_0$ that $\text{milk} = \text{dark}$ against $H_1$ that $\text{dark} > \text{milk}$. $P\_value\_z=1-\text{normcdf}(Z)$
- Repeat using T-statistics. # of degrees of freedom is $\text{dof}=2*(n-1)$
  $P\_value\_t=1-\text{tcdf}(|T|, \text{dof})$
Sweet matlab exercise #1

- dark=[118.8 122.6 115.6 113.6 119.5 115.9 115.8 115.1 116.9 115.4 115.6 107.9];
- milk=[102.1 105.8 99.6 102.7 98.8 100.9 102.8 98.7 94.7 97.8 99.7 98.6]
- x_dark=mean(dark) % sample mean dark chocolate
- x_milk=mean(milk) % sample mean milk chocolate
- s_dark=std(dark) % sample std dark chocolate
- s_milk=std(milk) % sample std milk chocolate
- n=12 % sample size of both dark and milk
- std_xdiff=sqrt(s_dark.^2./2+s_milk.^2./n) % std diff x
- z_stat=(x_dark-x_milk)./std_xdiff % z-statistic
- P_value_z=erfc(z_stat./sqrt(2))./2 % P-value of null true
  % P_value_z=9.9629e-34
- dof=(n-1)+(n-1) % # of degrees of freedom
- P_value_t=tcdf(z_stat,dof,'upper') % P-value of null true
  %P_value_t= 1.8417e-11
Credit: XKCD comics