HW4 has been posted.
Two-sided Confidence Interval on the Population Mean, \( \mu \),

Variance, \( \sigma \), is known.

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

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

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

Student’s t distribution

\[f(t) \sim \left(1 + \frac{t^2}{n - 1}\right)^{-n/2}\]
Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

**Definition**

(Eq. 8-19)

If $s^2$ is the sample variance from a random sample of $n$ observations from a normal distribution with unknown variance $\sigma^2$, then a 100(1 $-$ $\alpha$)% confidence interval on $\sigma^2$ is

$$
\frac{(n - 1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n - 1)s^2}{\chi^2_{1-\alpha/2,n-1}}
$$

(Eq. 8-19)

where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are the upper and lower 100$\alpha$/2 percentage points of the chi-square distribution with $n - 1$ degrees of freedom, respectively. A confidence interval for $\sigma$ has lower and upper limits that are the square roots of the corresponding limits in Equation 8-19.
\( (1 - \alpha)100\% \) of \( \chi^2 \) values are in this interval

\[
\chi^2_{1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2}
\]

\[
\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}
\]
8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

Definition

(Eq. 8-19)

If $s^2$ is the sample variance from a random sample of $n$ observations from a normal distribution with unknown variance $\sigma^2$, then a $100(1 - \alpha)%$ confidence interval on $\sigma^2$ is

$$\frac{(n - 1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n - 1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$

(8-19)

where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are the upper and lower $100\alpha/2$ percentage points of the chi-square distribution with $n - 1$ degrees of freedom, respectively. A confidence interval for $\sigma$ has lower and upper limits that are the square roots of the corresponding limits in Equation 8-19.
Confidence estimates of the population proportion
Prevalence (with 95% CI bars) of obesity among New York City public elementary schoolchildren, by sex and race/ethnicity, 2003.

(source: CDC.GOV)

Collect a sample of BMI values. Obese means BMI > 30

What do those bars actually mean?
Large sample confidence estimate of population proportion

- Want to know the fraction $p$ of the population that belongs to a class, e.g. the class “obese” kids defined by BMI>30.
- Each variable is a Bernoulli trial with one parameter $p$. We can use moments or MLE estimator to estimate $p$.
- Both give the same estimate: sample fraction $\hat{P} = \text{(number of obese kids in the sample)} / \text{(sample size n)}$.
- How to put confidence bounds on $p$ based on $\hat{P}$.
- Expected # of successes is $np$ $\rightarrow$ Expected fraction of successes is $p$.
- Standard deviation of # of successes is $\sqrt{np(1-p)}$ $\rightarrow$ Standard deviation of fraction of successes is $\sqrt{p(1-p)/n}$. 
8-5 A Large-Sample Confidence Interval For a Population Proportion (Eq. 8-23)

If $\hat{p}$ is the proportion of observations in a random sample of size $n$ that belongs to a class of interest, an approximate 100$(1 - \alpha)$% confidence interval on the proportion $p$ of the population that belongs to this class is

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (8-23)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

This interval is known as the Wald interval (Wald and Wolfowitz, 1939).
Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: **24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown**?

http://www.scientificamerikeren.com/candy5.asp

“To our surprise M&Ms met our demand to review their procedures in determining candy ratios. It is, however, noted that the figures presented in their email differ from the information provided from their website (http://us.mms.com/us/about/products/milkchocolate/). An email was sent back informing them of this fact. To which M&Ms corrected themselves with one last email:

In response to your email regarding M&M'S CHOCOLATE CANDIES

Thank you for your email.
On average, our new mix of colors for M&M'S® Chocolate Candies is:

- **M&M'S® Milk Chocolate**: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown.
- **M&M'S® Peanut**: 23% blue, 23% orange, 15% green, 15% yellow, 12% red, 12% brown.
- **M&M'S® Kids MINIS®**: 25% blue, 25% orange, 12% green, 13% yellow, 12% red, 13% brown.
- **M&M'S® Crispy**: 17% blue, 16% orange, 16% green, 17% yellow, 17% red, 17% brown.
- **M&M'S® Peanut Butter and Almond**: 20% blue, 20% orange, 20% green, 20% yellow, 10% red, 10% brown.

Have a great day!

Your Friends at Masterfoods USA
A Division of Mars, Incorporated

How to estimate these probabilities from a finite sample and how to set confidence interval on these estimates?
Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

How large is a sample needed for 95% CI on the percentage of blue M&Ms to be less than +/- 4%? Same question for red M&Ms?
Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: **24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown**?

How large is a sample needed for 95% CI on the percentage of blue M&Ms to be less than +/- 4%?
Same question for red M&Ms?

For blue M&Ms \( p = 0.24 \)

\[
1.96 \sqrt{\frac{0.24(1-0.24)}{n}} < 0.04
\]

\[
n > \left( \frac{1.96}{0.04} \right)^2 \times 0.24 \times (1-0.24) \approx 438 \text{ M&Ms or } \sim 2 \times 70 \text{ bags with 210 candies each}
\]

For red M&Ms \( p = 0.13 \)

\[
n > \left( \frac{1.96}{0.04} \right)^2 \times 0.13 \times (1-0.13) \approx 271 \text{ M&Ms or } \sim 1 \times 70 \text{ bag}
\]
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 $\bar{x} = 48$ mRNAs/cell (or even smaller) and $\sigma = 10$ mRNAs/cell in a healthy population with $\mu_h = 50$ 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 is 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.
\( \mu_h = 50 \)

\( H_0: \mu_c = \mu_h \)

\( n = 100, \bar{X} = 48, \sigma = 10 \)

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

\[ z_0 = 1.64 \]

Area = 0.05

\( \bar{X} = 48 \)

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

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

\[ H_0: \mu_c = \mu_h \]

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

\[ H_1: \mu_c < \mu_h \]

\[ \bar{x} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1 \]

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

\[ \frac{m_h - \bar{x}}{\sigma/\sqrt{n}} = \frac{50 - 48.36}{1.64} = 1.64 \]

Type II error

\[ \beta = P(\text{Accept } H_0 | H_1 \text{ is true}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left( - \frac{(x-48.36)^2}{2} \right) dx = 48.36 = 1 - \Phi(1.36) = 8.8\% \]
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 $\mu_0 +/- z_{\alpha/2} * S$ 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 $\mu_0 +/- t_{\alpha/2,n-1} T$ 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 $\beta$?

$$n \approx \frac{(z_{\alpha/2} + z_{\beta})^2 \sigma^2}{\delta^2}$$

where $\delta = \mu - \mu_0$ (9-22)
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 $\alpha = .05$ and $\alpha = .01$ 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
A peculiar prevalence of $p$ values just below .05

E. J. Masicampo$^1$, and Daniel R. Lalande$^2$

$^1$Department of Psychology, Wake Forest University, Winston-Salem, NC, USA
$^2$Department of Health Sciences, Université du Québec à Chicoutimi, Chicoutimi, QC, Canada
Hypothesis testing: two samples
10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

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

Sample 1: \(x_{11}, x_{12}, \ldots, x_{1n_1}\)

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

$$T^*_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

(10-15)

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

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

or more generally
Multiple null hypotheses: Bonferroni correction

• What if you have \( m \) independent null hypotheses? Say you have \( m=25,000 \) genes in a genome?
• What is the probability that at least one of the null-hypotheses will be shown to be false at significance threshold \( \alpha_1 \)?
• Answer: Family-Wise Error Rate or \( \text{FWER}=1-(1-\alpha_1)^m \approx m\alpha_1 \)
• If \( m=20 \) and \( \alpha_1=0.05 \), \( \text{FWER}= 0.6415 \)
• If you want to get \( \text{FWER}<\alpha \), use \( \alpha_1=\frac{\alpha}{m} \)

Carlo Emilio Bonferroni (1892 –1960)
Italian mathematician worked on probability theory.
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 mechanisms. 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 twice with n-hexane before extracting them with a mixture of water, acetone and acetic acid (70:29:8.0:2 by volume). We measured their in vitro total antioxidant capacities using the ferric-reducing antioxidant potential (FRAP) assay; 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 could be due to the formation of secondary bonds between chocolate flavonoids and milk proteins, which would reduce the biological accessibility of the flavonoids and therefore the chocolate’s potential antioxidant properties in vivo.

Our findings highlight the possibility...
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 milk = dark against $H_1$ that dark > 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