Matlab exercise:

• Generate a sample with \( n = 1000 \) following standard normal distribution
• Calculate median, first, and third quartiles
• Calculate IQR and find ranges shown below
• Find and count left and right outliers
• Do not use built-in Matlab functions for this!
• Make box and whisker plot: use boxplot
• n=1000; % make divisible by 4
• r1=randn(n,1);
• %boxplot(r1);
• [a,b]=sort(r1);
• q1=(a(n./4)+a(n./4+1))./2
• q3=(a(3.*n./4)+a(3.*n./4+1))./2
• q2=(a(n./2)+a(n./2+1))./2
• IQR=q3-q1
• sum(r1<q1-1.5.*IQR)
• sum(r1>q3+1.5.*IQR)
How many right outliers one expects in a sample of \( n=1000 \) following normal distribution?

- \% find the third quartile of a standard distribution
  \[
  \text{norminv}(0.75) \quad \% \text{ans} = 0.6745
  \]

- \% Calculate IQR - Inter Quartile Range
  \[
  \text{IQR}=2.\times \text{norminv}(0.75) \quad \% 1.3490
  \]

- \% Calculate \( 0.5\times\text{IQR}+1.5\times\text{IQR} \) - the right whisker position
  \[
  \text{whisker}=0.5.\times\text{IQR}+1.5\times\text{IQR} \quad \% \text{ans} = 2.6980
  \]

- \% Find the probability to be above the right whisker
  \[
  1-\text{normcdf}(\text{whisker}) \quad \% \text{ans} = 0.00349
  \]

- \% Find number of right outliers in a sample of 1000 points
  \[
  1000.\times(1-\text{normcdf}(\text{whisker})) \quad \% \text{ans} = 3.49
  \]
Descriptive statistics:
Sample mean and variance
Some Definitions

• The random variables $X_1, X_2, \ldots, X_n$ are a random sample of size $n$ if:
  a) The $X_i$ are independent random variables.
  b) Every $X_i$ has the same probability distribution.

Such $X_1, X_2, \ldots, X_n$ are also called independent and identically distributed (or i. i. d.) random variables.

• A statistic is any function of the observations in a random sample.

• The probability distribution of a statistic is called a sampling distribution.
Statistic #1: Sample Mean

If the $n$ observations in a random sample are denoted by $x_1, x_2, ..., x_n$, the sample mean is

$$
\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (6-1)
$$
IMPORTANT.

Sample mean \( \bar{X} \) is drawn from a random variable
\[
\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}
\]

\[
E(\bar{X}) = \frac{n \cdot E(X_i)}{n} = \frac{n \cdot \mu}{n} = \mu
\]

\[
V(\bar{X}) = \frac{n \cdot V(X_i)}{n^2} = \frac{n \cdot \sigma^2}{n} = \frac{\sigma^2}{n}
\]

Standard dev. \((\bar{X}) = \frac{\sigma}{\sqrt{n}}\)
Central Limit Theorem

If \( X_1, X_2, \ldots, X_n \) is a random sample of size \( n \) is taken from a population (either finite or infinite) with mean \( \mu \) and finite variance \( \sigma^2 \), and if \( \bar{X} \) is the sample mean, then the limiting form of the distribution of

\[
Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad (7-1)
\]

for large \( n \), is the standard normal distribution. If \( X_1, X_2, \ldots, X_n \) are themselves normally distributed - for any \( n \)
Sampling Distributions of Sample Means

**Figure 7-1** Distributions of average scores from throwing dice.
Mean = (6+1)/2=3.5
Sigma^2 = [(6-1+1)^2-1]/12=2.92
Sigma=1.71

**Formulas**

\[ \mu = \frac{b + a}{2} \]

\[ \sigma^2_X = \frac{(b - a + 1)^2 - 1}{12} \]

\[ \sigma^2_{X^2} = \sigma^2_X / n \]

show Matlab
Matlab demonstration

- Stats=100000; N=10;
- r_table=floor(6.*rand(Stats,N))+1;
- \%
- r1=r_table(:,1);
- step=1; [a,b1]=hist(r1,1:step:6);
- pdf_r1=a./sum(a)./step;
- figure; hold on; subplot(1,2,1); plot(b1,pdf_r1,'mo-'); hold on; axis([0 7 0 0.2]); subplot(1,2,2);
- semilogy(b1,pdf_r1,'mo-'); hold on; axis([0 7 1e-3 1]);
- \%
- r2=(r_table(:,1)+r_table(:,2))./2;
- step=0.5; [a,b2]=hist(r2,1:step:6); pdf_r2=a./sum(a)./step;
- subplot(1,2,1); plot(b2,pdf_r2,'rd-'); axis([0 7 0 0.4]); subplot(1,2,2); semilogy(b2,pdf_r2,'rd-');
- \%
- r3=(r_table(:,1)+r_table(:,2)+r_table(:,3))./3;
- step=1./3; [a,b3]=hist(r3,1:step:6); pdf_r3=a./sum(a)./step;
- subplot(1,2,1); plot(b3,pdf_r3,'gs-'); axis([0 7 0 0.4]); subplot(1,2,2); semilogy(b3,pdf_r3,'gs-');
- \%
- r5=sum(r_table(:,1:5),2)./5;
- step=1./5; [a,b5]=hist(r5,1:step:6); pdf_r5=a./sum(a)./step;
- subplot(1,2,1); plot(b5,pdf_r5,'b^-'); axis([0 7 0 0.6]); subplot(1,2,2); semilogy(b5,pdf_r5,'b^-'); axis([0 7 1e-4 1]);
- \%
- r10=sum(r_table(:,1:10),2)./10;
- step=1./10; [a,b10]=hist(r10,1:step:6); pdf_r10=a./sum(a)./step;
- subplot(1,2,1); plot(b10,pdf_r10,'kv-'); axis([0 7 0 0.8]); legend(num2str([1,2,3,5,10]));
- subplot(1,2,2); semilogy(b10,pdf_r10,'kv-'); legend(num2str([1,2,3,5,10]));
Matlab demonstration; part 2

- %%Now plot all of them normalized to 0 and std 1
- sigma=sqrt(35/12);
- mu=3.5;
- figure;
- sigma1=sigma;
- semilogy((b1-mu)./sigma1,pdf_r1.*sigma1,'mo-');
- axis([-4 4 1e-3 1]);
- hold on;
- sigma2=sigma./sqrt(2);
- semilogy((b2-mu)./sigma2,pdf_r2.*sigma2,'rd-');
- sigma3=sigma./sqrt(3);
- semilogy((b3-mu)./sigma3,pdf_r3.*sigma3,'gs-');
- sigma5=sigma./sqrt(5);
- semilogy((b5-mu)./sigma5,pdf_r5.*sigma5,'b^-');
- sigma10=sigma./sqrt(10);
- semilogy((b10-mu)./sigma10,pdf_r10.*sigma10,'kv-');
- axis([-4 4 1e-4 1]);
- %%
- %Let's see how well does the Gaussian fits it
- x=-4:0.1:4;
- semilogy(x,1./sqrt(2*pi)*exp(-x.^2./2),'y-');
Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of $n = 25$ resistors will have an average resistance of less than 95 ohms?
Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of \( n = 25 \) resistors will have an average resistance of less than 95 ohms?

\[
\mu = 100 \text{ ohms}, \quad \sigma = 10 \text{ ohms}, \quad n = 25
\]

\[
\mu_{\bar{x}} = \mu \cdot \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2 \text{ ohms}
\]

\[
Z_{\bar{x}} = \frac{95 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{95 - 100}{2} = -2.5
\]

\[
Prob(\bar{X} < 95) = \Phi(Z_{\bar{x}}) = \Phi(-2.5) = 0.0062
\]
Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of \( n = 25 \) resistors will have an average resistance of less than 95 ohms?

Answer:

\[
\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2.0
\]

\[
\Phi\left( \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \right) = \Phi\left( \frac{95 - 100}{2} \right) = \Phi(-2.5) = 0.0062
\]
Two Populations

We have two independent populations. What is the distribution of the difference of their sample means?

The sampling distribution of $\bar{X}_1 - \bar{X}_2$ has the following mean and variance:

$$
\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2
$$

$$
\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}
$$
Sampling Distribution of a Difference in Sample Means

• If we have two independent populations with means $\mu_1$ and $\mu_2$, and variances $\sigma_1^2$ and $\sigma_2^2$,

• And if $X_{-bar1}$ and $X_{-bar2}$ are the sample means of two independent random samples of sizes $n_1$ and $n_2$ from these populations:

• Then the sampling distribution of:

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

(7-4)

is approximately standard normal, if the conditions of the central limit theorem apply.

• If the two populations are normal, then the sampling distribution is exactly standard normal.
Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a random variable with \( \mu_{\text{old}} = 5000 \text{ hours} \) and \( \sigma_{\text{old}} = 40 \text{ hours (old)} \). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters to \( \mu_{\text{new}} = 5050 \text{ hours} \) and \( \sigma_{\text{new}} = 30 \text{ hours (new)} \).

Random samples of 16 components manufactured using “old” process and 25 components using “new” process are chosen.

What is the probability new sample mean is at least 25 hours longer than old?
Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a random variable with $\mu_{\text{old}}=5000$ hours and $\sigma_{\text{old}}=40$ hours (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters to $\mu_{\text{new}}=5050$ hours and $\sigma_{\text{new}}=30$ hours (new).

Random samples of 16 components manufactured using “old” process and 25 components using “new” process are chosen.

What is the probability new sample mean is at least 25 hours longer than old?

$$\sigma_{\bar{X}_{\text{old}}} = \frac{\sigma_{\text{old}}}{\sqrt{16}} = 10 \text{ hrs}$$

$$\bar{X}_{\text{new}} = \frac{\sigma_{\text{new}}}{\sqrt{25}} = 6 \text{ hrs}$$

$$\sigma_{\text{tot}} = \sqrt{\sigma_{\bar{X}_{\text{old}}}^2 + \sigma_{\bar{X}_{\text{new}}}^2} = \sqrt{100 + 36} \approx 11.7 \text{ hrs}$$

$$\mu_{\text{new}} - \mu_{\text{old}} = 50 \text{ hrs}$$

$$Z = \frac{25 - (50)}{11.7} = -2.14$$

$$\text{Prob}(Z \geq -2.14) = 0.9840$$
Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a normal-distributed random variable with parameters shown (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters mu and sigma as shown (new). Random samples are selected from the “old” process and “new” process as shown.

What is the probability new sample mean is at least 25 hours longer than old?

<table>
<thead>
<tr>
<th>Process</th>
<th>Old (1)</th>
<th>New (2)</th>
<th>Diff (2-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mu</td>
<td>5,000</td>
<td>5,050</td>
<td>50</td>
</tr>
<tr>
<td>sigma</td>
<td>40</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>n</td>
<td>16</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Calculations</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s / \sqrt{n} = 10 )</td>
</tr>
<tr>
<td>( z = \frac{\bar{x}_2 - \bar{x}_1}{s / \sqrt{n}} = -2.14 )</td>
</tr>
<tr>
<td>( P(\bar{x}_2 - \bar{x}_1 &gt; 25) = P(Z &gt; z) = 0.9840 )</td>
</tr>
<tr>
<td>Questions Found in Google Autocomplete</td>
</tr>
<tr>
<td>----------------------------------------</td>
</tr>
<tr>
<td>Why Do Whales Jump?</td>
</tr>
<tr>
<td>Why Are Witches Green?</td>
</tr>
<tr>
<td>Why Are There Mirrors Above Beds?</td>
</tr>
<tr>
<td>Why Do I Say Uh?</td>
</tr>
<tr>
<td>Why Is Sea Salt Better?</td>
</tr>
<tr>
<td>Why Are There Tres In The Middle Of Fields?</td>
</tr>
<tr>
<td>Why Is There Not A Pokemon MMO?</td>
</tr>
<tr>
<td>Why Is There Laughing In TV Shows?</td>
</tr>
<tr>
<td>Why Are There Doors On The Freeway?</td>
</tr>
<tr>
<td>Why Are There So Many Swooshes Running?</td>
</tr>
<tr>
<td>Why Aren't There Any Countries In Antarctica?</td>
</tr>
<tr>
<td>Why Are There Scary Sounds In Minecraft?</td>
</tr>
<tr>
<td>Why Is There Kicking In My Stomach?</td>
</tr>
<tr>
<td>Why Are There Two Slashes After HTTP?</td>
</tr>
<tr>
<td>Why Are There Celebrities?</td>
</tr>
<tr>
<td>Why Do Snakes Exist?</td>
</tr>
<tr>
<td>Why Do Oysters Have Pearls?</td>
</tr>
<tr>
<td>Why Are Ducks Called Ducks?</td>
</tr>
<tr>
<td>Why Do They Call It The Clap?</td>
</tr>
<tr>
<td>Why Are Kyle And Cartman Friends?</td>
</tr>
<tr>
<td>Why Is There An Arrow On Aang's Head?</td>
</tr>
<tr>
<td>Why Are Text Messages Blue?</td>
</tr>
<tr>
<td>Why Are There Mustaches On Clothes?</td>
</tr>
<tr>
<td>Why Are There Mustaches On Cars?</td>
</tr>
<tr>
<td>Why Are There Mustaches Everywhere?</td>
</tr>
<tr>
<td>Why Are There So Many Birds In Ohio?</td>
</tr>
<tr>
<td>Why Is There So Much Rain In Ohio?</td>
</tr>
<tr>
<td>Why Is Ohio Weather So Weird?</td>
</tr>
<tr>
<td>Why Are There Male And Female Bikes?</td>
</tr>
<tr>
<td>Why Are There Tiny Spiders In My House?</td>
</tr>
<tr>
<td>Why Do Spiders Come Inside?</td>
</tr>
<tr>
<td>Why Are There Huge Spiders In My House?</td>
</tr>
<tr>
<td>Why Are There Lots Of Spiders In My House?</td>
</tr>
<tr>
<td>Why Are There Spiders In My Room?</td>
</tr>
<tr>
<td>Why Are There So Many Spiders In My Room?</td>
</tr>
<tr>
<td>Why Do Spider Bites itch</td>
</tr>
<tr>
<td>Why Is Dying So Scary?</td>
</tr>
<tr>
<td>Why Is There A Ghost?</td>
</tr>
<tr>
<td>Why Are There Female Ghosts?</td>
</tr>
<tr>
<td>Why Is Mt. Vesuvius There?</td>
</tr>
<tr>
<td>Why Do They Say T Minus</td>
</tr>
<tr>
<td>Why Are There Obelisks?</td>
</tr>
<tr>
<td>Why Are Wrestlers Always Wet?</td>
</tr>
<tr>
<td>Why Are Oceans Becoming More Acidic?</td>
</tr>
<tr>
<td>Why Is Arwen Dying?</td>
</tr>
<tr>
<td>Why Aren't My Qaul Laying Eggs?</td>
</tr>
<tr>
<td>Why Aren't My Qaul Eggs Hatching?</td>
</tr>
<tr>
<td>Why Aren't There Any Foreign Military Bases In America?</td>
</tr>
</tbody>
</table>
Descriptive statistics:
Point estimation:
Sample mean and variance
Point Estimation

• A sample was collected: $X_1, X_2, \ldots, X_n$
• We suspect that sample was drawn from a random variable distribution $f(x)$
• $f(x)$ has k parameters that we do not know
• Point estimates are estimates of the parameters of the $f(x)$ describing the population based on the sample
  — For exponential PDF: $f(x) = \lambda \exp(-\lambda x)$ one wants to estimate $\lambda$
  — For Bernoulli PDF: $p^x(1-p)^{1-x}$ one wants to estimate $p$
  — For normal PDF one wants to estimates both $\mu$ and $\sigma$
• Point estimates are uncertain: therefore we can talk of averages and standard deviations of point estimates
Point Estimator

A point estimate of some parameter $\theta$ describing population random variable is a single numerical value $\hat{\theta}$ depending on all values $x_1, x_2, \ldots, x_n$ in the sample.

The sample statistic (whichever a random variable $\hat{\Theta}$ defined by a function $\hat{\Theta}(X_1, X_2, \ldots, X_n)$) is called the point estimator.

- There could be multiple choices for the point estimator of a parameter.
- To estimate the mean of a population, we could choose the:
  - Sample mean
  - Sample median
  - Peak of the histogram
  - $\frac{1}{2}$ of (largest + smallest) observations of the sample.
- We need to develop criteria to compare estimates using statistical properties.
Unbiased Estimators Defined

The point estimator $\hat{\Theta}$ is an unbiased estimator for the parameter $\theta$ if:

$$E\left(\hat{\Theta}\right) = \theta \quad \text{(7-5)}$$

If the estimator is not unbiased, then the difference:

$$E\left(\hat{\Theta}\right) - \theta \quad \text{(7-6)}$$

is called the bias of the estimator $\hat{\Theta}$. 
Example 7-4: Sample Mean is Unbiased

- $X$ is a random variable with mean $\mu$ and variance $\sigma^2$. Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$.
- Show that the sample mean ($\bar{X}$) is an unbiased estimator of $\mu$.

\[
E(\bar{X}) = E\left( \frac{X_1 + X_2 + \ldots + X_n}{n} \right) = \frac{1}{n} \left[ E(X_1) + E(X_2) + \ldots + E(X_n) \right] = \frac{1}{n} \left[ \mu + \mu + \ldots + \mu \right] = \frac{n\mu}{n} = \mu
\]