

BIOE 198MI Biomedical Data Analysis. Spring Semester 2019.  
 Lab 5: Introduction to Statistics

**A. Review: Ensemble and Sample Statistics**

The normal probability density function (pdf) from which random samples are drawn has two parameters: the ensemble mean  $\mu$  and ensemble standard deviation  $\sigma$  or variance  $\sigma^2$ . Generally, a normal pdf is

$$N(\mu, \sigma) = p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \text{ while the}$$

std normal equation is  $p(z; 0, 1) = p(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ .

The two are related using  $x = \sigma z + \mu$ . In Matlab, we illustrate the pdf using `normpdf(x, m, s)` or `pdf('norm', x, m, s)`. We don't directly measure ensemble statistics, i.e., population parameters  $\mu, \sigma^2$ . Instead, we intuit them from the physics of a problem. However, we can estimate the mean and variance from measurement data via sample statistics  $\bar{x}$  and  $s^2$ .

Sample statistics:

- Estimate  $\mu$ : the sample mean from  $N$  measurement samples is  $\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$ .
- Estimate  $\sigma^2$ : the sample variance from  $N$  samples is  $s^2 = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})^2$ , where  $N-1$  reflects the loss of one degree of freedom when computing  $\bar{x}$ .
- The sample standard deviation is the square root of the sample variance  $s = \sqrt{s^2}$ .

Note that  $\bar{x}, s$  are themselves random variables whereas  $\mu, \sigma$  are constant model parameters.

**B. Accuracy/Bias and Precision/Variance**

The mean-squared error (*MSE*) asks how well measurement  $X$  estimates  $\mu$  from random process  $p(x; \mu, \sigma) = \mathcal{N}(\mu, \sigma)$ . We note that  $MSE \geq 0$ , and that values near zero suggest  $X$  is an accurate and precise estimate of parameter  $\mu$ . Looking more closely,

$$\begin{aligned} MSE &\triangleq \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2 = \frac{1}{N} \sum_{n=1}^N x_n^2 - \frac{2}{N} \sum_{n=1}^N x_n \mu + \frac{1}{N} \sum_{n=1}^N \mu^2 && \text{(multiply to find three terms)} \\ &= \left[ \frac{1}{N} \sum_{n=1}^N x_n^2 \right] - \left[ \frac{2\mu}{N} \sum_{n=1}^N x_n \right] + \mu^2 \\ &= \left[ \frac{1}{N} \sum_{n=1}^N x_n^2 \right] - 2\mu\bar{x} + \mu^2 + \bar{x}^2 - \bar{x}^2 && \text{(add and subtract } \bar{x}^2 \text{)} \end{aligned}$$

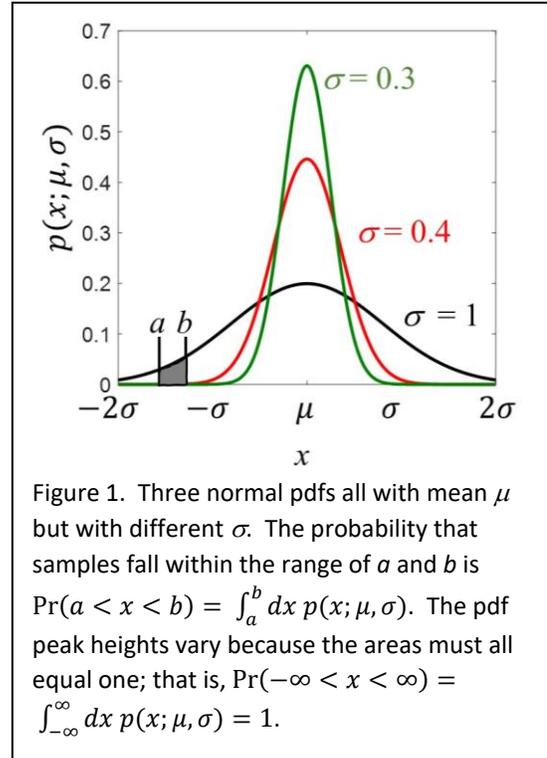


Figure 1. Three normal pdfs all with mean  $\mu$  but with different  $\sigma$ . The probability that samples fall within the range of  $a$  and  $b$  is  $\Pr(a < x < b) = \int_a^b dx p(x; \mu, \sigma)$ . The pdf peak heights vary because the areas must all equal one; that is,  $\Pr(-\infty < x < \infty) = \int_{-\infty}^{\infty} dx p(x; \mu, \sigma) = 1$ .

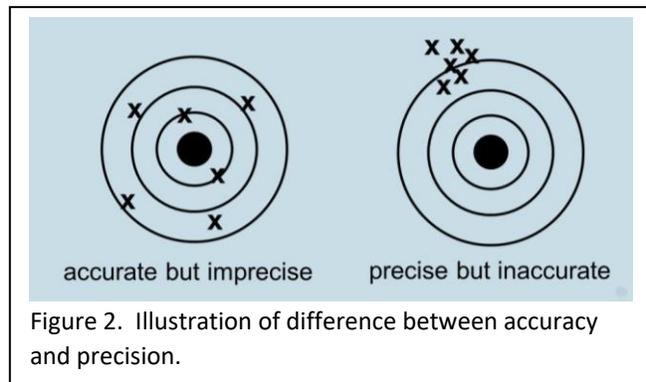


Figure 2. Illustration of difference between accuracy and precision.

$$\begin{aligned}
&= \frac{1}{N} [(\sum_{n=1}^N x_n^2) - N\bar{x}^2] + \bar{x}^2 - 2\bar{x}\mu + \mu^2 && \text{(rearrange \& complete square)} \\
&= \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2 + (\bar{x} - \mu)^2 \\
&= s^2 + b^2
\end{aligned}$$

The *MSE* statistic is the sum of sample variance  $s^2$  (except for  $1/N$  instead of  $1/(N-1)$ ) and squared bias  $b^2$ . If measurement bias is negligible, i.e.,  $(\bar{x} - \mu)^2 \cong 0$ , then  $MSE = s^2 \xrightarrow{N \rightarrow \infty} \sigma^2$ .

**Exercise B:** Generate in Matlab a 5000-point sequence of measurements using  $\mathcal{N}(5,3)$ . You are unsure of the distribution from which the data are formed and model it using  $\mathcal{N}(2,3)$ . Estimate *MSE* and  $s^2$  from the data and thereby show the bias equals 3.

### C. Ensemble Mean of Sample Statistics

Because sample statistics are random variables, we can ask about the ensemble mean of the sample mean  $\bar{X}$ , written as  $E(\bar{X})$ , and the ensemble variance the sample mean written as  $var(\bar{X})$  or  $\sigma_{\bar{x}}^2$ . Without showing details,

$$E(\bar{X}) = \mu, \quad var(\bar{X}) \triangleq \sigma_{\bar{x}}^2 = \frac{\sigma^2}{N}, \quad std(\bar{X}) \triangleq \sqrt{\sigma_{\bar{x}}^2} = \frac{\sigma}{\sqrt{N}}.$$

We say the sample mean is asymptotically unbiased; i.e., as  $N \rightarrow \infty$ ,  $\bar{x} \rightarrow \mu$ . However, the sample standard deviation is biased:  $\sigma_{\bar{x}} = \sigma/\sqrt{N}$ . (Fig 3)

You showed this was true in the assignment for Lab 4. Because the std of the mean is less than or equal to the std, we repeat experiments and average the results. *In summary, the average of many unbiased, statistically independent measurements is the best estimate of the population mean.*

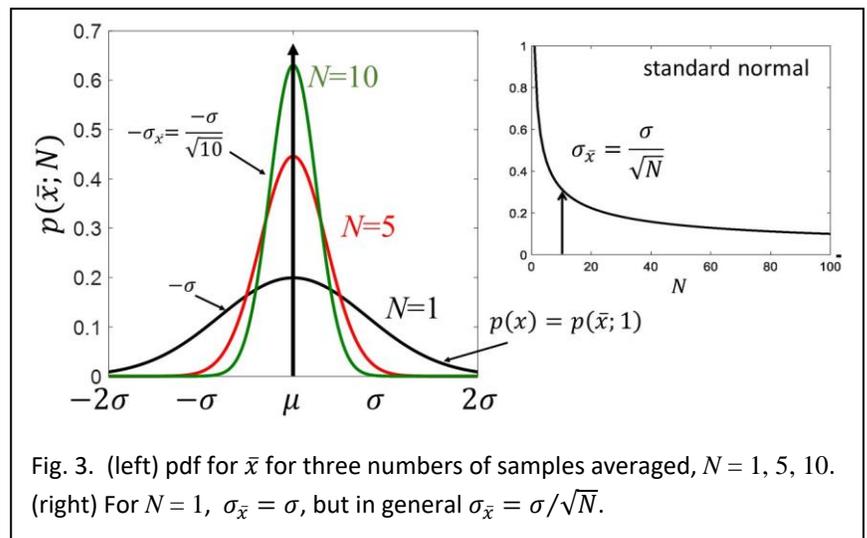


Fig. 3. (left) pdf for  $\bar{x}$  for three numbers of samples averaged,  $N = 1, 5, 10$ . (right) For  $N = 1$ ,  $\sigma_{\bar{x}} = \sigma$ , but in general  $\sigma_{\bar{x}} = \sigma/\sqrt{N}$ .

### D. Example of Statistical Prediction: Body Temperature Measurements

The mean body temperature in the healthy adult human population is  $98.2 \pm 1.5^\circ\text{F}$  ( $\mu \pm \sigma$ ).

Question (a): What is the probability that the next healthy adult you meet has a body temperature between  $97^\circ\text{F}$  and  $99^\circ\text{F}$ ? (Note, this  $\bar{X}$  range is for a group of one,  $N = 1$ .)

A. (a) Convert the temperature range to standard normal form:  $(97-98.2)/1.5 = -0.8$  and  $(99-98.2)/1.5 = 0.53$ . Then integrate the standard normal pdf via CDFs to find the probability requested (See Fig 4a):

$\Pr(97 \leq \bar{X} \leq 99)_{N=1} = CDF(0.53) - CDF(-0.80) = 0.702 - 0.212 = \boxed{0.49}$ . We find there is a 49% probability when  $N = 1$ . In Matlab: `Pr1=cdf('norm',0.53,0,1)-cdf('norm',-0.8,0,1)`.

Question (b): What is the probability of finding the average body temperature in the same range from an average of 10 patients?

$$A. (b) \Pr(97 \leq \bar{X} \leq 99)_{N=10} = CDF\left(\frac{99-98.2}{1.5/\sqrt{10}}\right) - CDF\left(\frac{97-98.2}{1.5/\sqrt{10}}\right)$$

$$CDF(1.69) - CDF(-2.53) = 0.954 - 0.0057 = \boxed{0.95}$$

The probability is 95% for  $N=10$ . (See Fig 4b).

In Matlab:

```
Pr10=cdf('norm',1.687,0,1)-cdf('norm',-2.529,0,1)
```

Conclusion: averaging more samples when computing sample means increases confidence in finding data within a specified range. **\*\*We assumed  $\mu, \sigma$  are known, which is not practical\*\***

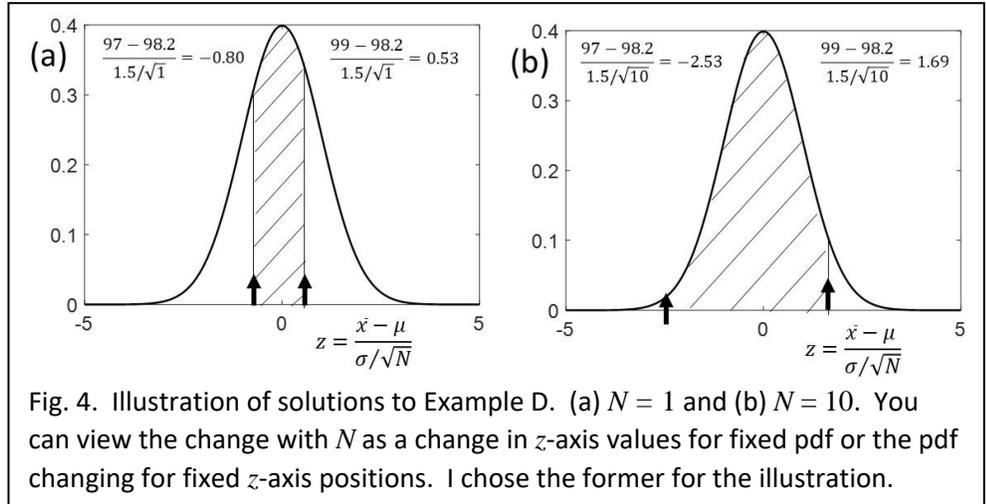


Fig. 4. Illustration of solutions to Example D. (a)  $N = 1$  and (b)  $N = 10$ . You can view the change with  $N$  as a change in  $z$ -axis values for fixed pdf or the pdf changing for fixed  $z$ -axis positions. I chose the former for the illustration.

### E. Hypothesis Testing and Error Thresholds:

Say we make two sets of  $N$  measurements with sample means  $\bar{x}_1 = 4.0$  and  $\bar{x}_2 = 7.6$ . Both could be drawn from population  $\mathcal{N}(\mu, \sigma)$  where  $\mu$  and  $\sigma$  are known to be 5 and 1, respectively. Converting from the original  $\bar{x}$  measurement axis to the standard-normal axis using  $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}}$ , we find the distribution of Fig 5. Our null hypothesis is that data for both sample means were drawn from  $\mathcal{N}(\mu, \sigma)$ . How do we decide whether to accept or reject the null hypothesis? This is statistical decision making!

First, we must select decision thresholds along the  $z$  axis based on the error we are willing to accept. If it is physically possible for  $\bar{x}$  to be both greater and less than  $\mu$ , then we will set two symmetric thresholds at  $z_\alpha$  and  $z_{1-\alpha}$ . The net probability of error is the areas under the pdf outside the two thresholds.

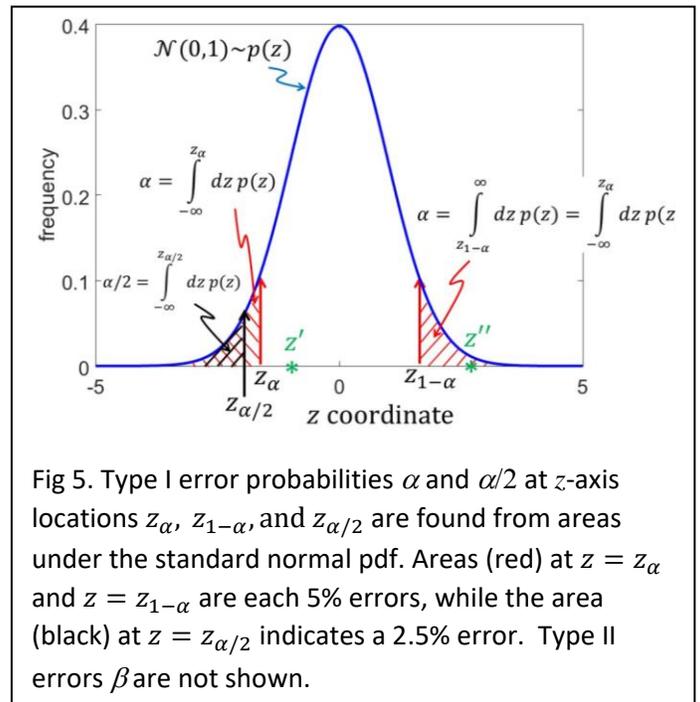


Fig 5. Type I error probabilities  $\alpha$  and  $\alpha/2$  at  $z$ -axis locations  $z_\alpha, z_{1-\alpha}$ , and  $z_{\alpha/2}$  are found from areas under the standard normal pdf. Areas (red) at  $z = z_\alpha$  and  $z = z_{1-\alpha}$  are each 5% errors, while the area (black) at  $z = z_{\alpha/2}$  indicates a 2.5% error. Type II errors  $\beta$  are not shown.

Imagine from Fig 5 that setting thresholds at  $z_\alpha = -5.0$  and  $z_{1-\alpha} = 5.0$  we will likely call virtually any measured sample mean as belonging to  $\mathcal{N}(\mu, \sigma)$ . The false-positive error probability (type I errors made by rejecting the null hypothesis when it is true) is very small, i.e.,  $\alpha \sim 0$ . However, the false-negative error probability (type II errors made by accepting the null hypothesis when it is false) is very large,  $\beta \sim 1$ .

Conversely, setting  $z_\alpha = -0.05$  and  $z_{1-\alpha} = 0.05$  will likely result in  $\alpha \sim 1$  and  $\beta \sim 0$ . Neither situation is desirable, so we may need to look closely at the cost of each error to decide how to set thresholds. What is certain is that errors are inevitable no matter how thresholds are set!

In one situation, we might be willing to accept a two-tailed, 10% net error probability, i.e.,  $\alpha = 0.05$  on either side of the mean in Fig 5. We then select symmetric decision thresholds,  $z_\alpha$ , and  $z_{1-\alpha}$  such that the probability of error is  $2\alpha = 0.10$ . In another situation, we might decide to accept a restrictive one-tailed, 2.5% error probability, and thus we will search for either  $z_{\alpha/2}$  or  $z_{1-\alpha/2}$  depending on whether the values are greater or less than  $\mu$ . Thresholds are found using the inverse CDF function in Matlab:

```
za=icdf('norm',0.05,0,1) = -1.645      %This finds  $z_\alpha$  for  $\alpha=0.05$ 
zao2=icdf('norm',0.025,0,1) = -1.960   %This finds  $z_{\alpha/2}$  for  $\alpha/2=0.025$ 
z1ma=icdf('norm',0.95,0,1) = 1.645     %This finds  $z_{1-\alpha}$  for  $1-\alpha=0.95$ 
```

Notice that  $z_{1-\alpha} = -z_\alpha$ . For the two-tailed 10% error condition, we accept the null hypothesis when  $-1.645 \leq z \leq 1.645$ . In terms of  $\bar{x}$ , the range is  $3.355 \leq \bar{x} \leq 6.645$ , since  $\bar{x} = (\sigma z / \sqrt{N}) + \mu$  and  $\mu = 5$ ,  $\sigma = 1$  and  $N = 1$ .

Returning to the example at the beginning of this section, where  $N = 1$  and  $\bar{x}_1 = 4.0$ , we find  $z_1 = \frac{\bar{x}_1 - 5}{1} = -1.0$ . This value falls into the range for accepting the null hypothesis  $-1.645 \leq z_1 \leq 1.645$ , and so we cannot reject the null hypothesis. That is, we decide that  $\bar{x}_1$  is part of the  $\mathcal{N}(\mu, \sigma)$  distribution.

In contrast, we find that for measurement  $\bar{x}_2 = 7.6$ , where  $z_2 = \frac{\bar{x}_2 - 5}{1} = 2.6$ , we must reject the null hypothesis with 90% confidence. That is, we say  $\bar{x}_2$  is different from  $\mathcal{N}(\mu, \sigma)$  at a 10% level of significance.

**Exercise E:** Test the null hypothesis that  $\bar{x}_2 = 7.6$  belongs to distribution  $\mathcal{N}(5,1)$  for a one-tailed error probability of  $\alpha = 0.01$ .

## F. Student's t-statistics

Assume the more realistic situation where the underlying population variance  $\sigma^2$  is unknown. Thus, we estimate variance using sample statistics, viz.,  $s^2$ . In this situation, the standard normal variable  $z$  changes to a t-variable with  $N-1$  degrees of freedom (dof), i.e.,  $= \frac{\bar{X} - \mu}{\sigma / \sqrt{N}} \rightarrow t = \frac{\bar{X} - \mu}{s / \sqrt{N}}$ . t-distributions are families of continuous pdfs that become relevant when estimating the mean of a normally distributed population with small sample size  $N$  and unknown  $\sigma$ . Substituting  $s$  for  $\sigma$  in  $z$  changes the distribution of the test statistic to another symmetric form also with zero mean but with more area under the tails, as shown in Fig 6. See `tcdf`, `tinv` and `tpdf`.

We set thresholds for t-statistics similar to those for hypothesis testing with known  $\sigma^2$ . In Fig 6, the  $t$  coordinate for a 5% right-tailed error,  $1-\alpha = 0.95$ , 2 dof =  $N - 1$  is found in Matlab using

```
t1ma2=tinv(0.95,2) = 2.920.
```

Similarly, the 5% left-tailed error is  $\alpha = 0.05$ , 2 dof

$t_{1-\alpha,2} = \text{tinv}(0.05, 2) = -2.920$  ,

where again  $t_{1-\alpha} = -t_{\alpha}$  .

Compared with the standard normal pdf, the t-distribution has more area in the distribution tails particularly when the #dof =  $N - 1$  is small. To see this, we fix  $\alpha$  at 0.05 and vary #dof from 1:100 to compare the results with that of a standard normal pdf:

```
q=[1 2 5 10 100];
for j=1:5
    t(j)=tinv(0.05,q(j))
end
```

$d$	$t_{0.05,d}$
1	-6.3138
2	-2.9200
5	-2.0150
10	-1.8125
100	-1.6602

The code above results in threshold values along the t axis shown in the table. Comparing these values to  $z_{0.05} = -1.645$  from the standard normal variable, we see they are nearly equal for  $N = 101$ .

Table values reveal that if you wish to restrict one-tailed decision errors to 5% and you do not know the population parameters, you need to extend the decision threshold past 6 standard deviations of the mean if you only have one degree of freedom (average  $N = 2$  measurements). However, averaging just three measurements reduces the threshold more than a factor of two to about three standard deviations of the mean. Averaging more than 100 measurements returns you to thresholds approximately given by the standard normal pdf where population parameters are known.

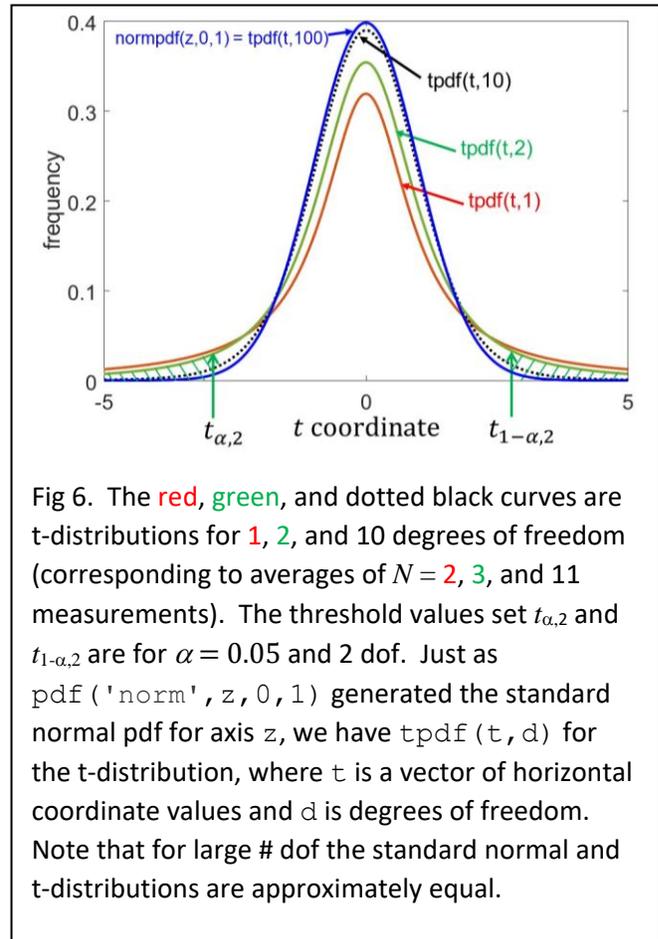


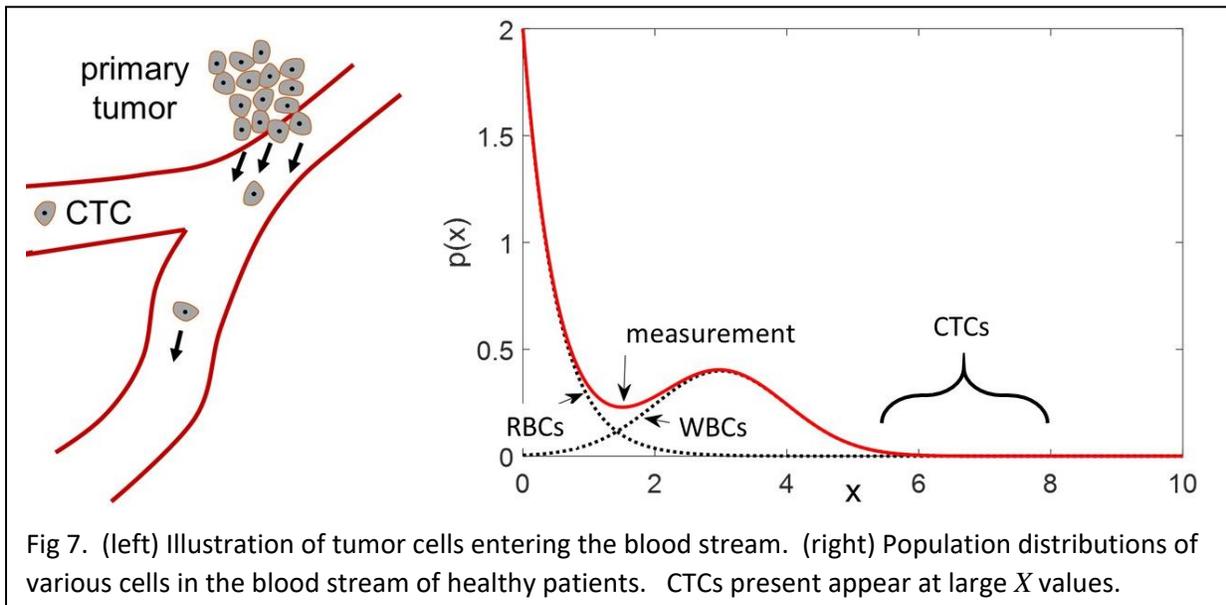
Fig 6. The red, green, and dotted black curves are t-distributions for 1, 2, and 10 degrees of freedom (corresponding to averages of  $N = 2, 3,$  and 11 measurements). The threshold values set  $t_{\alpha,2}$  and  $t_{1-\alpha,2}$  are for  $\alpha = 0.05$  and 2 dof. Just as  $\text{pdf}('norm', z, 0, 1)$  generated the standard normal pdf for axis  $z$ , we have  $\text{tpdf}(t, d)$  for the t-distribution, where  $t$  is a vector of horizontal coordinate values and  $d$  is degrees of freedom. Note that for large # dof the standard normal and t-distributions are approximately equal.

### Assignment 1:

Most cancer deaths result from complications associated with metastatic disease. Metastases result when circulating tumor cells (CTCs) from epithelial cancers, e.g., breast, prostate, lung, and colon, travel through the vasculature to implant in remote regions of the body and grow into tumors; see Fig. 7 (left). There is much interest in developing simple and low-cost techniques for detecting CTCs in patients so those at risk for metastatic disease can be aggressively treated early.

To understand the measurement, first note that each milliliter of whole blood contains about a billion red blood cells (RBCs), a few million white blood cells (WBCs) of various types, and less than 10 CTCs if they are present. CTCs are very sparse and thus difficult to detect even when they are specifically labeled for detection. CTCs are found by labeling receptor sites on the cells

that serve as biomarkers, and then painstakingly examining a great many cells often by eye. The process can be sensitive and specific but it is time consuming and expensive, and therefore it not used as much as it could be.



You have an idea that could make detection of CTCs faster and cheaper if it works. The idea is to implement high-throughput measurements of four biomarkers from each blood cell using flow cytometry. We multiply four measurement values to form one normal random variable  $X$  that we subject to hypothesis testing:

$$X = S \times n \times CK \times \frac{1}{CD45} .$$

$S$  is cell size in  $\mu\text{m}$ , where CTCs are generally larger than other blood elements.  $n$  is a nuclear factor that is zero when no nucleus is detected, as in RBCs, and  $n = 1$  when a nucleus is detected, as in WBCs and CTCs.  $CK$  is the optical intensity of cytokeratin fluorescent marker; large  $CK$  values indicate a high probability of CTCs.  $CD45$  is a receptor-linked protein tyrosine phosphatase expressed on leucocytes. The marker for  $CD45$  is weak when CTCs are present and significant otherwise. To decide if  $X$  measured for each cell indicates a CTC, we use hypothesis testing and aggressive thresholds (small values of  $\alpha$ ) to minimize type I errors. We need to be sure before calling a patient positive for CTCs because the cost of type I errors is high.

The distribution of  $X$  measured from blood samples from many healthy subjects is bimodal (Fig. 7 right). If CTCs are present, they are found at large  $X$  values. The large narrow peak near  $x = 0$  is from RBCs and the smaller peak near  $x = 3$  is from WBCs. Assume you have data from 7 volunteer blood samples to estimate sample mean and variance,  $\bar{x}$  and  $s^2$ . Design detection thresholds to conduct hypothesis testing for the error probabilities  $\alpha = 0.001, 0.01, 0.05$ .

(a) Plot the appropriate t-pdf and standard normal pdf for this experiment. Indicate in a table and on the pdf plot where the thresholds are located. You should guess at an appropriate value for  $s$ .

(b) What is the threshold for the standard normal pdf at  $\alpha = 0.001$ . What is the equivalent t-statistic error for this standard normal result?

(c) Convert the t-pdf thresholds to now be thresholds along the  $X$  axis.

(d) You run a set of annotated patient data with at least one patient having metastatic disease. At  $\alpha = 0.001$ , you measure no positive cases. What do you do? What questions should be asked?

### Assignment 2:

The birth weights of 10,000 full-term human fetuses were measured during one year in a cluster of Midwestern US cities. The distribution was normally distributed with mean  $\mu_w = 3$  kg but the variance was not well determined and so is considered unknown.

Two smaller studies in Chicago hospitals (labeled A and B) involving fewer babies each gave sample mean birth weight of  $\bar{w} \pm s = 2650 \text{ g} \pm 550 \text{ g}$  and  $N = 25$  (study A) and  $2725 \text{ g} \pm 691 \text{ g}$  and  $N = 45$  (study B). Since the mean values are somewhat lower than the larger study, and the associated neighborhoods are high in economically disadvantaged families, there is concern. It is known that low birth weight can be an indicator of unhealthy adolescents. Should subpopulations A or B be considered at risk? Argue for or against the significance of the lower sample means for studies A and B. Explore  $\alpha = 0.001$ ,  $0.01$ , and  $0.1$  but note the value for  $\alpha$  most relevant for making decisions is one that depends on many factors outside of the data given in this problem. So all you can do at this time is report on null hypothesis test results for the three values of  $\alpha$  above and then discuss them with study investigators.

There is no need to generate plots for this problem. It could help to organize results using a table.

### Rubric:

- Introduction for both projects, where you explain you are using hypothesis testing on two different problems. 2 pts.
- Methods: Describe the approach to computing these thresholds (either  $z$  or  $t$  values) in both problems. One point each for methods sections for Assignments 1 and 2.
- Results: Assignment 1 requires plots and Assignment 2 does not (although you can use plot or table to tell the story). One point each for the results in the two assignments.
- Discussion and Conclusion: Please discuss your thinking about how you can (1) adjust your threshold in the first assignment to obtain the best diagnostic performance. (2) Give reasons why you decided if the subpopulations had lower birth weights. 3 pts.
- One point for appearance of the report including visual displays of results, clarity and logic of the flow of ideas and a clear answer.