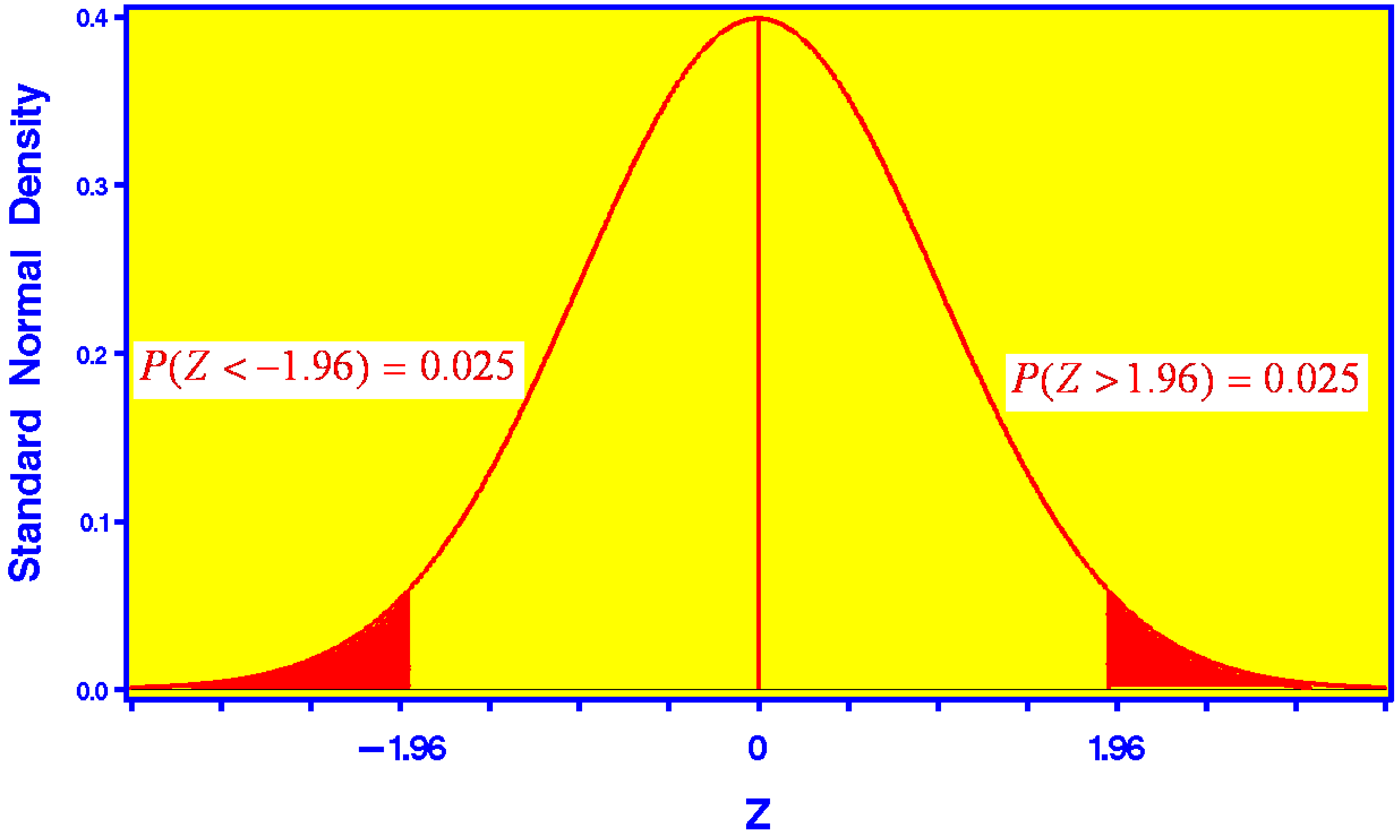
Part 1
Lecture 2b The Central Limit Theorem

1

Medical Imaging
UNIVERSITY OF TORONTO

MiDATA

STANDARD GAUSSIAN PROBABILITY DENSITY



TESTING THE NULL HYPOTHESIS

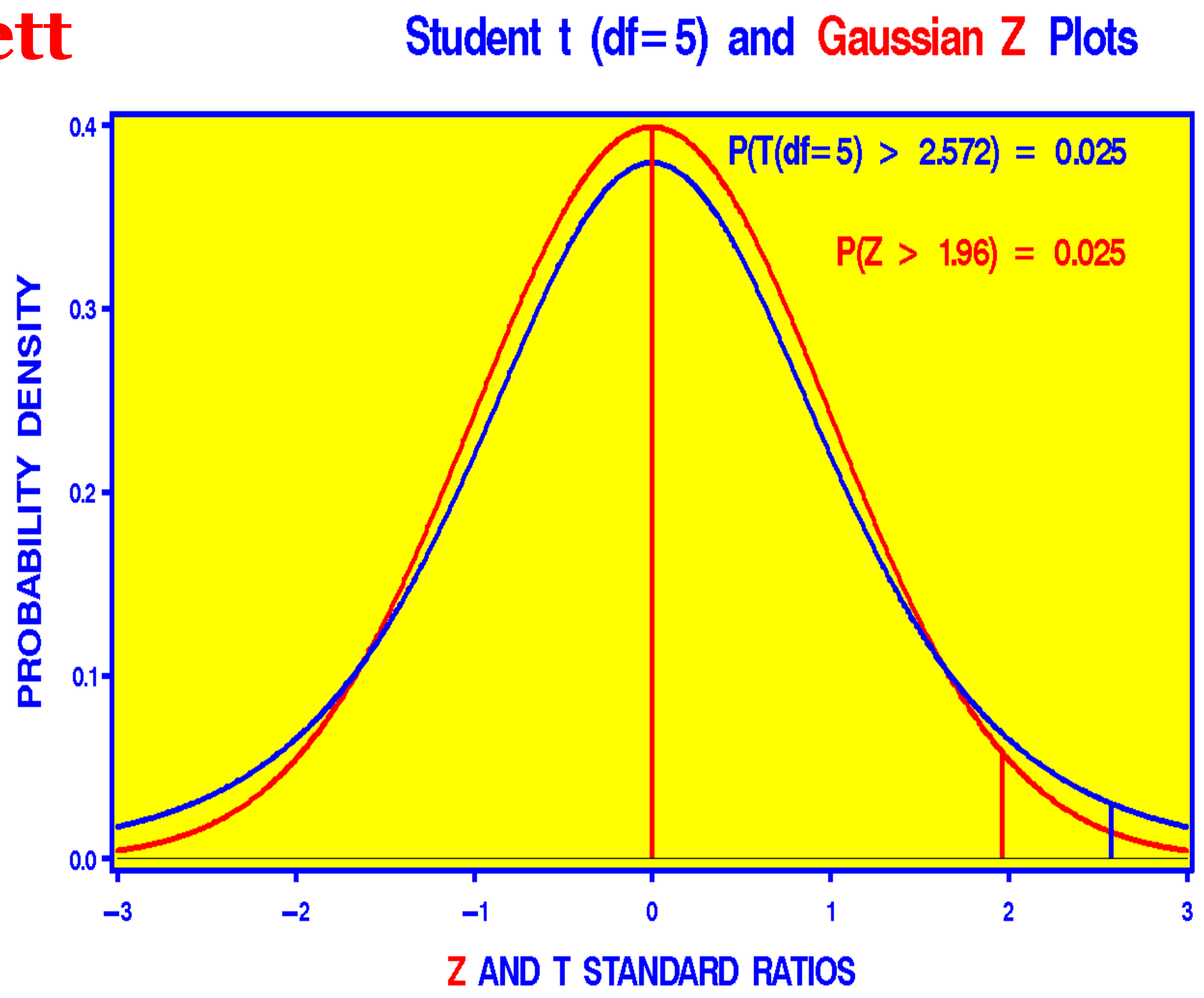
$$Z = \frac{\overline{DIFF} - (\mu_1 - \mu_2)}{\sqrt{2 \times \frac{\sigma^2}{n}}} \quad ---\gg \quad Z = \frac{\overline{DIFF} - 0}{\sqrt{2 \times \frac{s^2}{n}}}$$

The Null or Chance Hypothesis is $\Delta = \mu_1 - \mu_2 = 0$.

The population or theoretical variance σ^2 is replaced by sample variance s^2. It is not appropriate if sample size n is small.
Recall the three questions:

- How large is the mean difference between groups?
- How many patients were in each group ?
- How large is the *Variation* in response among patients?

William Sealy Gossett
1876 – 1937



STUDENT TTEST TO COMPARE UNPAIRED MEANS

$$\text{Student T ratio} = \frac{\overline{DBP_2} - \overline{DBP_1} - \Delta}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s^2_{Pooled}}} = \frac{\overline{DBP_2} - \overline{DBP_1} - \Delta}{\sqrt{\frac{2 \times s^2_{Pooled}}{n}}}$$

$$\text{Pooled Variance } s^2_{Pooled} = \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{s_1^2 + s_2^2}{2}$$

$if\ n_1 = n_2 = n$

$\Delta = \mu_2 - \mu_1$

$DF = Degrees\ of\ Freedom = n_1 - 1 + n_2 - 1 = 2 \times (n - 1)$

COMPARING STUDENT T PROBABILITY (PROBT)WITH THE GAUSSIAN PROBABILITIES FOR EXCEEDING 1.96

$P(|Z| > 1.96) = P(Z > 1.96 \text{ OR } Z < -1.96) = 0.05$

n1	n2	PROBT	n1	n2	PROBT
3	3	0.1216	61	61	0.0523
5	5	0.0857	101	101	0.0514
7	7	0.0736	181	181	0.0508

NOTE: n1 and n2 are the number of subjects in each group.

RECALL THE LAST FOUR QUESTIONS

4. How large is the mean DBP difference between groups?

5. How many patients were in each group ?

6. How large is the variation in response among patients?

7. Are the two groups comparable ?

NOTE: DBP = Decrease in blood pressure.

BEFORE CONSIDERING THESE FOUR IMPORTANT QUESTIONS

LET US CONSIDER A FEW IMPORTANT STATISTICAL SUMMARIES

MEASURES OF LOCATION

$Data$: $Y_1 = 4 \quad Y_2 = 22 \quad Y_3 = 18 \quad Y_4 = 9 \quad Y_5 = 7$

Ordered Data:	$Y_{(1)}$	$Y_{(5)}$	$Y_{(4)}$	$Y_{(3)}$	$Y_{(2)}$
	4	7	9	18	22
Rank:	1	2	3	4	5

$$Median = Y_{(4)} = 9$$

$$Sample\ Mean = \overline{Y} = \frac{\sum_{j=1}^{j=5} Y_j}{n} = \frac{4 + 22 + 18 + 9 + 7}{5} = \frac{60}{5} = 12$$

TWO SUMMARY STATISTICS OF VARIATION

Ordered Data:

$Y_{(1)}$	$Y_{(2)}$	$Y_{(3)}$	$Y_{(4)}$	$Y_{(5)}$
4	7	9	18	22

$Minimum = 4 \quad Maximum = 22 \quad Range = 18$

$First\ Quartile = 7 \quad Third\ Quartile = 18$

$Interquartile\ Range = 18 - 7 = 11$

CALCULATING A MEASURE OF VARIATION USING ALL OF THE DATA

MEAN OF DEVIATIONS ABOUT SAMPLE MEAN

$$= \frac{(8-20)+(14-20)+(22-20)+(26-20)+(30-20)}{5}$$

$$= \frac{-12-6+2+6+10}{5} = \frac{0}{5} = 0$$

$Always\ equal\ to\ zero\ so\ not\ a\ useful\ measure\ of\ anything.$

CALCULATING A MEASURE OF VARIATION USING ALL OF THE DATA

$MEAN\ OF\ ABSOLUTE\ DEVIATIONS\ ABOUT\ SAMPLE\ MEAN$

$$= \frac{|8-20| + |14-20| + |22-20| + |26-20| + |30-20|}{5}$$

$$= \frac{12+6+2+6+10}{5} = \frac{36}{5} = 7.2$$

$Nice\ try\ but\ no\ cigar$!!

THE SAMPLE VARIANCE S^2 IS THE MEAN OF THE SQUARED DEVIATIONS ABOUT THE SAMPLE MEAN

$$S^2 = \frac{(8-20)^2 + (14-20)^2 + (22-20)^2 + (26-20)^2 + (30-20)^2}{5-1}$$

$$= \frac{144 + 36 + 4 + 36 + 100}{5-1} = \frac{320}{4} = 80$$

S^2 *is an* ***UNBIASED*** *estimate of the population variance* σ^2 *which we saw was in the formula of the famous Gaussian probability function.*

With $n = 4$ *in the denominator instead of* 5 *this estimate of* σ^2 *becomes* ***UNBIASED***.

SUM of SQUARED DEVIATIONS ABOUT 22

$$S^2 = \frac{(8-22)^2 + (14-22)^2 + (22-22)^2 + (26-22)^2 + (30-22)^2}{5-1}$$

$$= \frac{196 + 64 + 0 + 16 + 64}{5-1} = \frac{340}{4} = 85 \quad > \quad \textit{sample variance of } 80$$

The sum of squares deviations is minimized when the squared deviations are about the sample mean. That is why the sample mean is called the Least Squares Estimator of the true mean.

End of Lecture 2

Next up in Part 1 Lecture 3: Association