Statistics: Examples and Exercises

20.109 Fall 2010

Module 1 Day 7

Your Data and Statistics

"Figures often beguile me," he wrote, "particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: 'There are three kinds of lies: lies, damned lies, and statistics.'"

Quote from Mark Twain, Chapters from My Autobiography, 1906

Why are stats important

- Sometimes two data sets look different, but aren't
- Other times, two data sets don't look that different, but are.

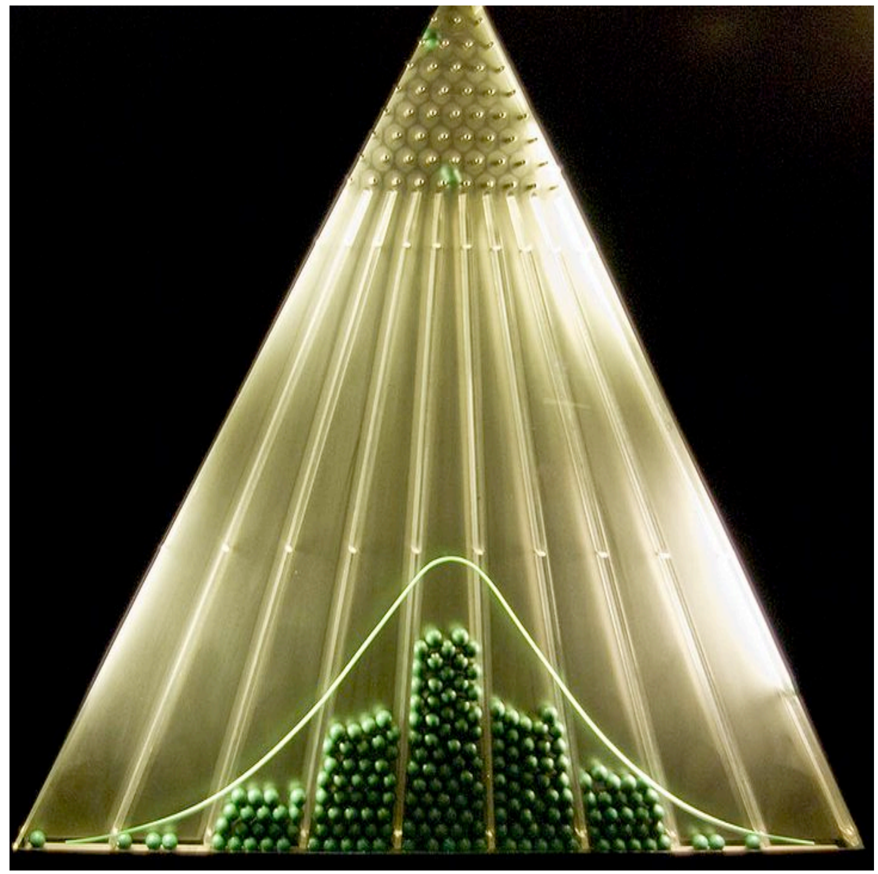
Why are stats important

- Informed experimental design is very powerful

- Save time, money, experimental subjects, patients, lab animals …….

Normal Distribution

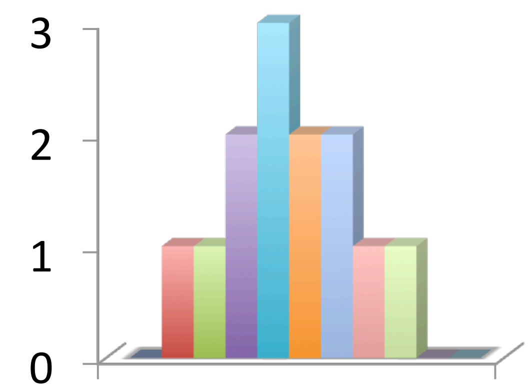
- The data are centered around the mean
- The data are distributed symmetrically around the mean



http://en.wikipedia.org/wiki/File:Planche_de_Galton.jpg

Mean μ vs $\bar{x}$

- The entire population mean is μ
- Sample population mean is $\bar{x}$
- As your sample population gets larger, $\bar{x} \rightarrow \mu$

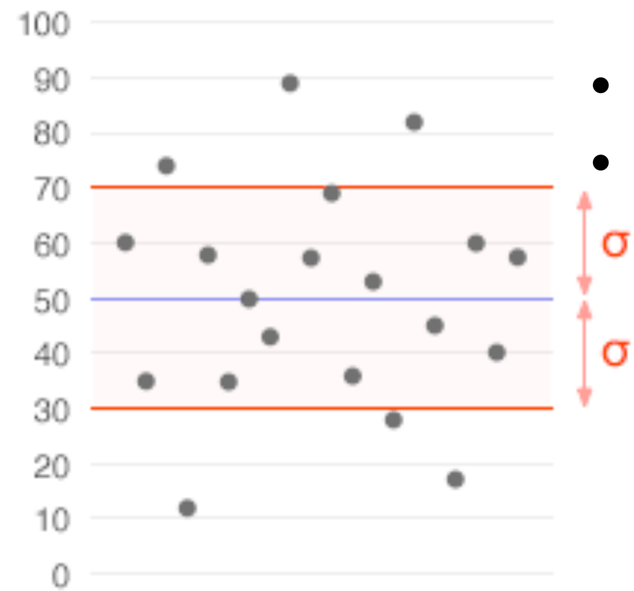


- Data Set
  - 2, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 9
- Mean

$$\bar{x} = \frac{2+3+4+4+5+5+6+6+7+7+8+9}{12}$$

Standard Deviation

- Describes how data are expected to vary from the mean



- $\mu = 50$
- $\sigma = 20$

- σ is s.d. of population
  s is s.d. of sample

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

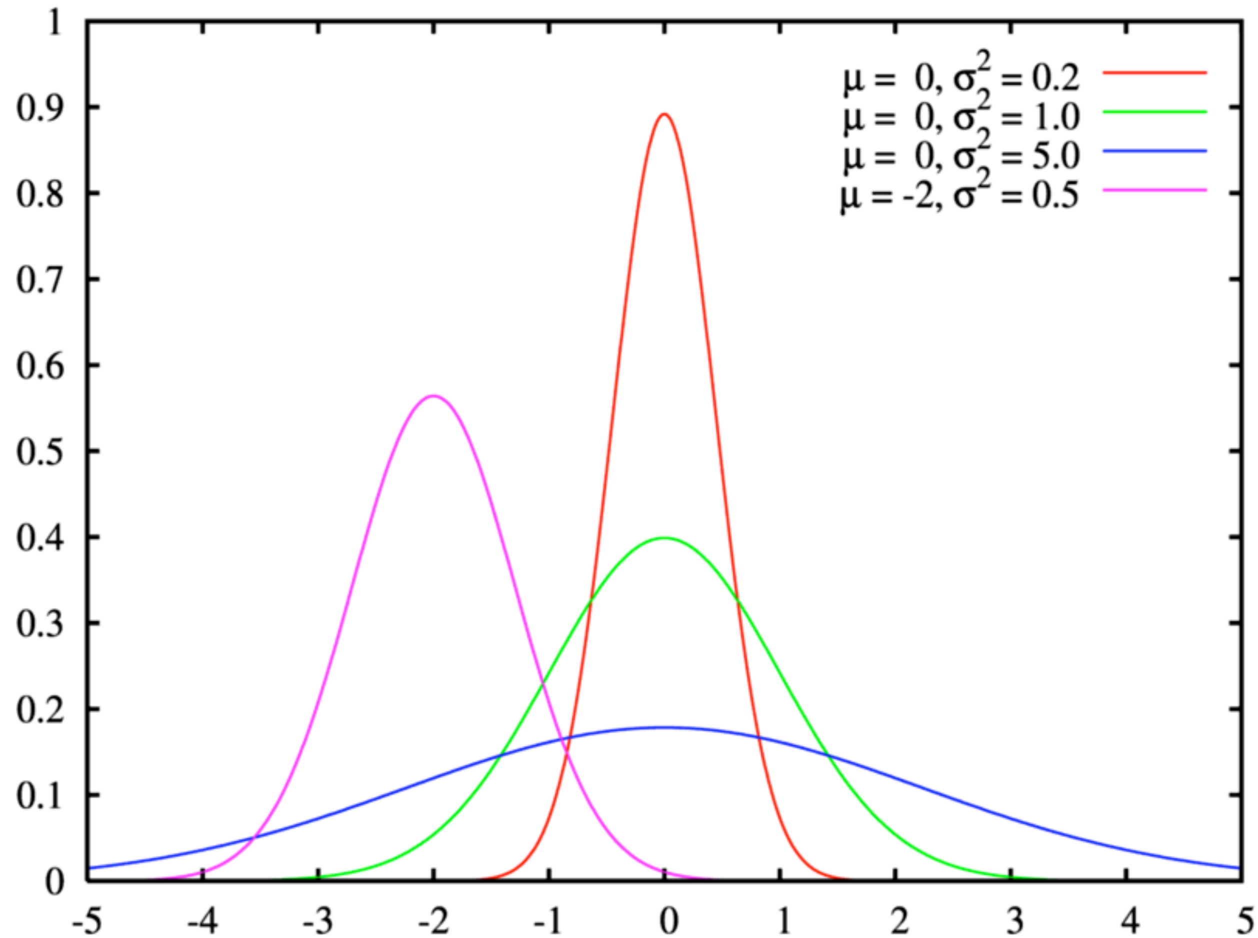
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

http://en.wikipedia.org/wiki/File:Standard_deviation_illustration.gif

Meaning of Standard Deviation

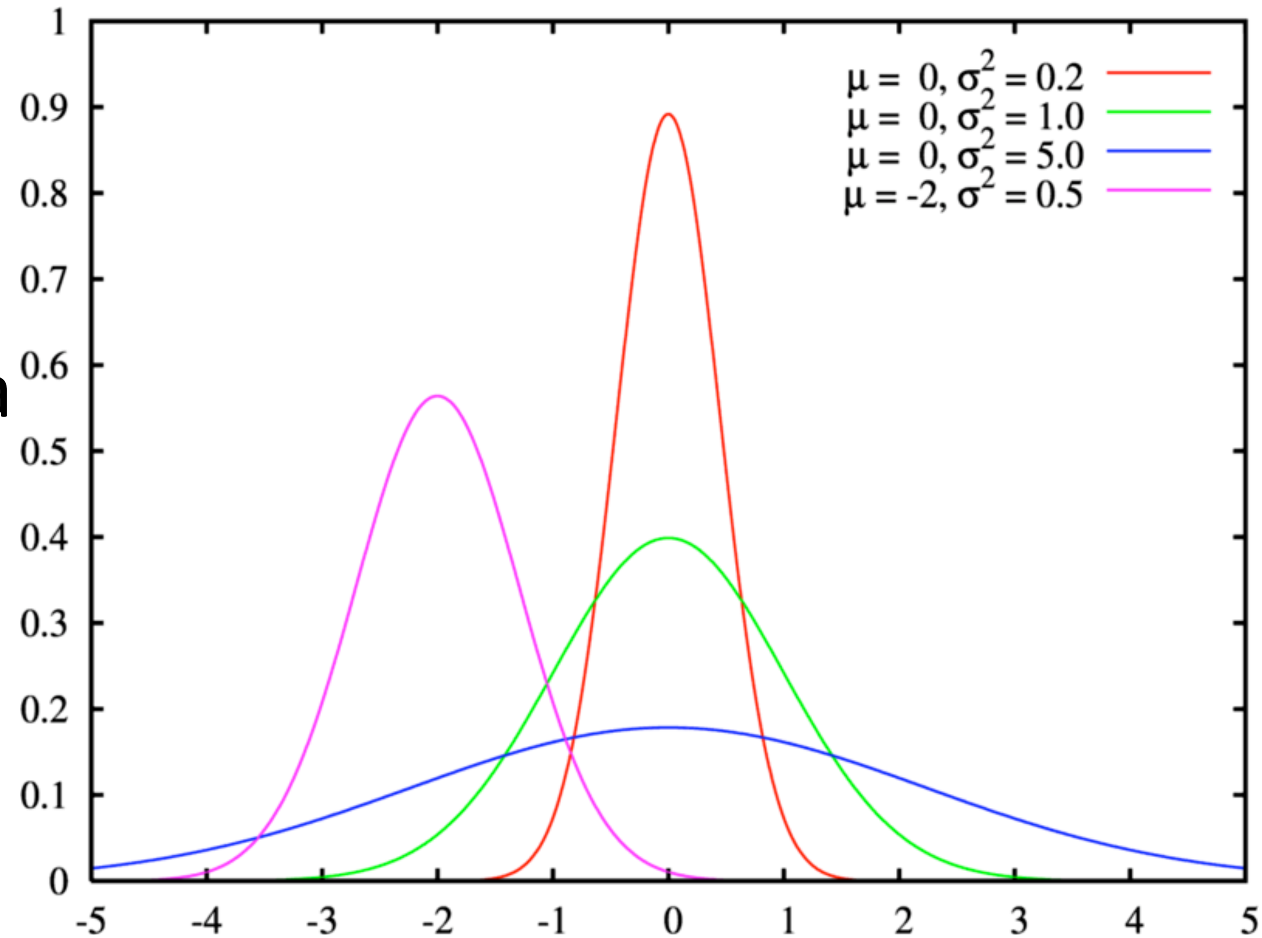
- Red, Green, Blue all same mean

- Different standard deviation



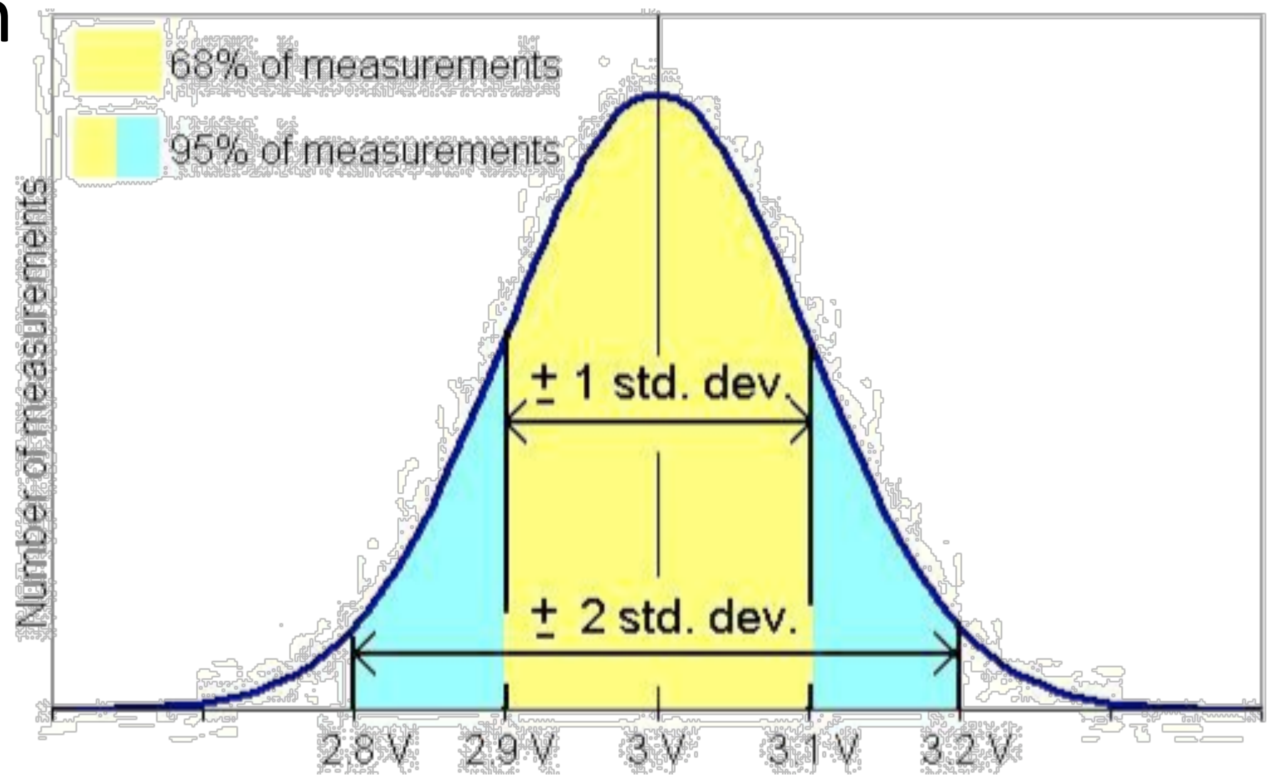
Meaning of Standard Deviation

- Data with a larger spread (blue and green) have a larger Standard Deviation



Standard Deviation

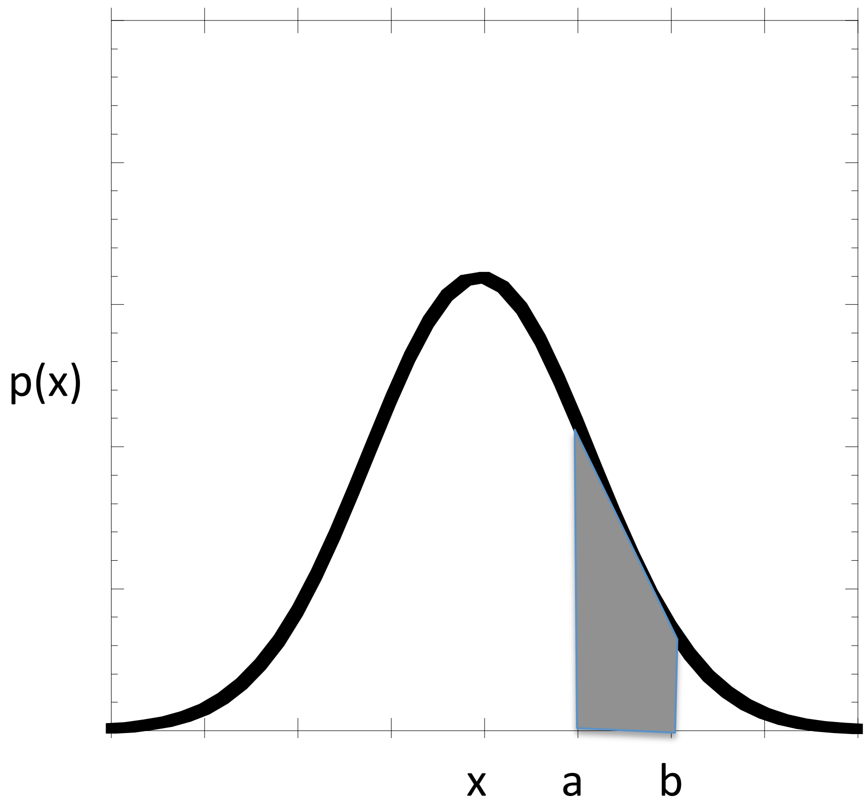
- 68% of values are within 1 standard deviation
- 95% of values are within 2 standard deviations of the mean



Statistical Significance

- How do we know that two data sets are truly different

Recap: Probability density function p(x)



x is a random number

Normalized

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

Probability that

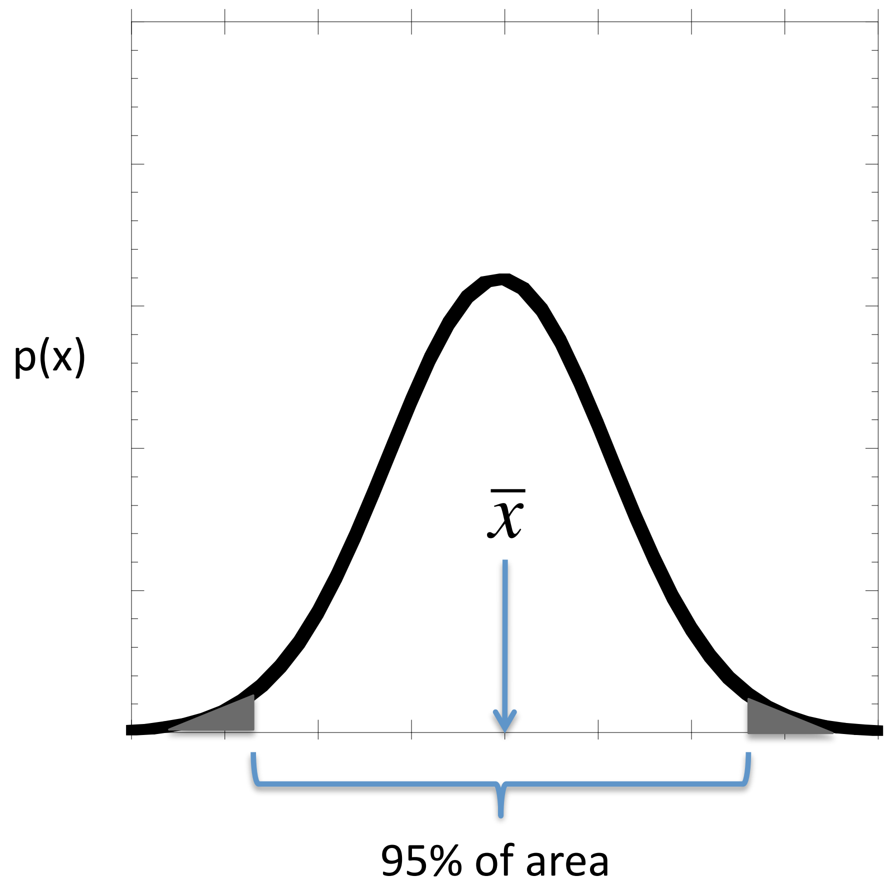
$$a < x < b$$

is

$$\int_{a}^{b} p(x)dx$$

95% confidence interval of an estimate

A range such that 95% of replicate estimates would be within it



95% Confidence interval for a normally distributed variable

$$\bar{x} - \frac{t_{0.025}s}{\sqrt{n}} < \mu < \bar{x} + \frac{t_{0.025}s}{\sqrt{n}}$$

# data points	$t_{0.025}$
2	12.706
3	4.303
4	3.182
5	2.776
10	2.262
20	2.093
30	2.045
50	2.010
100	1.984

Increasingly accurate estimate of σ

Note: Uncertainty decreases proportionally to $\frac{1}{\sqrt{n}}$

So take more data!

Example

3 measurements of absorbance at 600 nm: 0.110, 0.115, 0.113

95% confidence limit?

Soln: $\bar{x} = 0.113, s = 0.0025$

$$\bar{x} - \frac{t_{0.025} s}{\sqrt{n}} < \mu < \bar{x} + \frac{t_{0.025} s}{\sqrt{n}}$$

$$0.113 - \frac{4.303(0.0025)}{\sqrt{3}} < \mu < .113 + \frac{4.303(0.0025)}{\sqrt{3}}$$

$$0.107 < \mu < 0.119$$

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Confidence Intervals

- Use t to find interval containing μ if $\bar{x}$ is known

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}}$$

- Example:

t_{95} = 2.6

$$\mu_1 = 7.5 \pm \frac{2.6 \times 1.0}{\sqrt{6}}$$

Hawks	Cyclones
9	4
8	6
7	5
6	2
7	4
8	5
X_1 7.5	X_2 4.3
s_1 1.0	s_2 1.4

$$6.4 < \mu < 8.6$$

I am 95% confident that the population mean lies between 6.4 and 8.6

T-tests

- Compare confidence intervals to see if data sets are significantly different
- Assumptions
  - Data are normally distributed
  - The mean is independent of the standard deviation
    - $\mu \neq f(\sigma)$
- Various types
  - One sample t-test
    - Are these data different than the entire population?
  - Two sample t-test
    - Do these two data sets come from different populations?
  - Paired t-test
    - Do individual changes show an overall change?

Use t-test to compare means

- We have $\bar{x}_1$ and $\bar{x}_2$
  - Do they come from different populations?
    - Are μ_1 and μ_2 different?
- Null Hypothesis H_o:
  - $\bar{x}_1 = \bar{x}_2$
- Alternative Hypothesis H_a:
  - $\bar{x}_1 > \bar{x}_2$
- t statistic tests H_o. If $t < 0.05$, then reject H_o and accept H_a

T-test Illustration

- Two populations that are significantly different, with X_2 larger than X_1

T-test Illustration

- Two populations that are not significantly different, but X_2 is still larger than X_1

$$H_o : \bar{x}_1 = \bar{x}_2$$

$$H_A : \bar{x}_1 \rangle \bar{x}_2$$

Exercise: Find 99% Confidence

- $$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

  t=?

	MIT		Harvard
	100		46
	87		54
	56		76
	87		92
	98		87
	90		60
X_1	86.3	X_2	69.2
s_1	15.9	s_2	18.6

$$s = \sqrt{\frac{\sum_{set1}(x_i - \bar{x}_1)^2 + \sum_{set2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

$s = ?$

$t_{calc} = 1.79$

$t_{99} = ?$

t_{calc} ? T_{99}

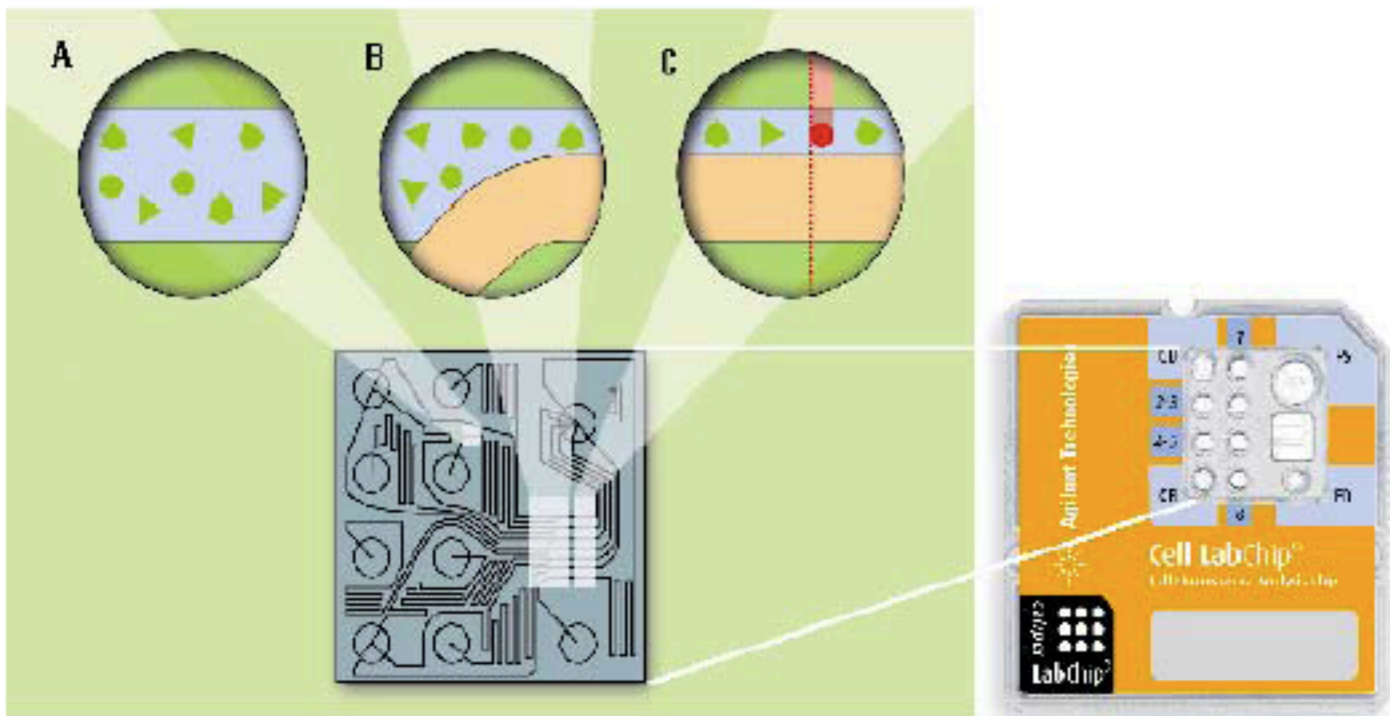
Go to table in notes to find t_{99} with 11 degrees of freedom

-

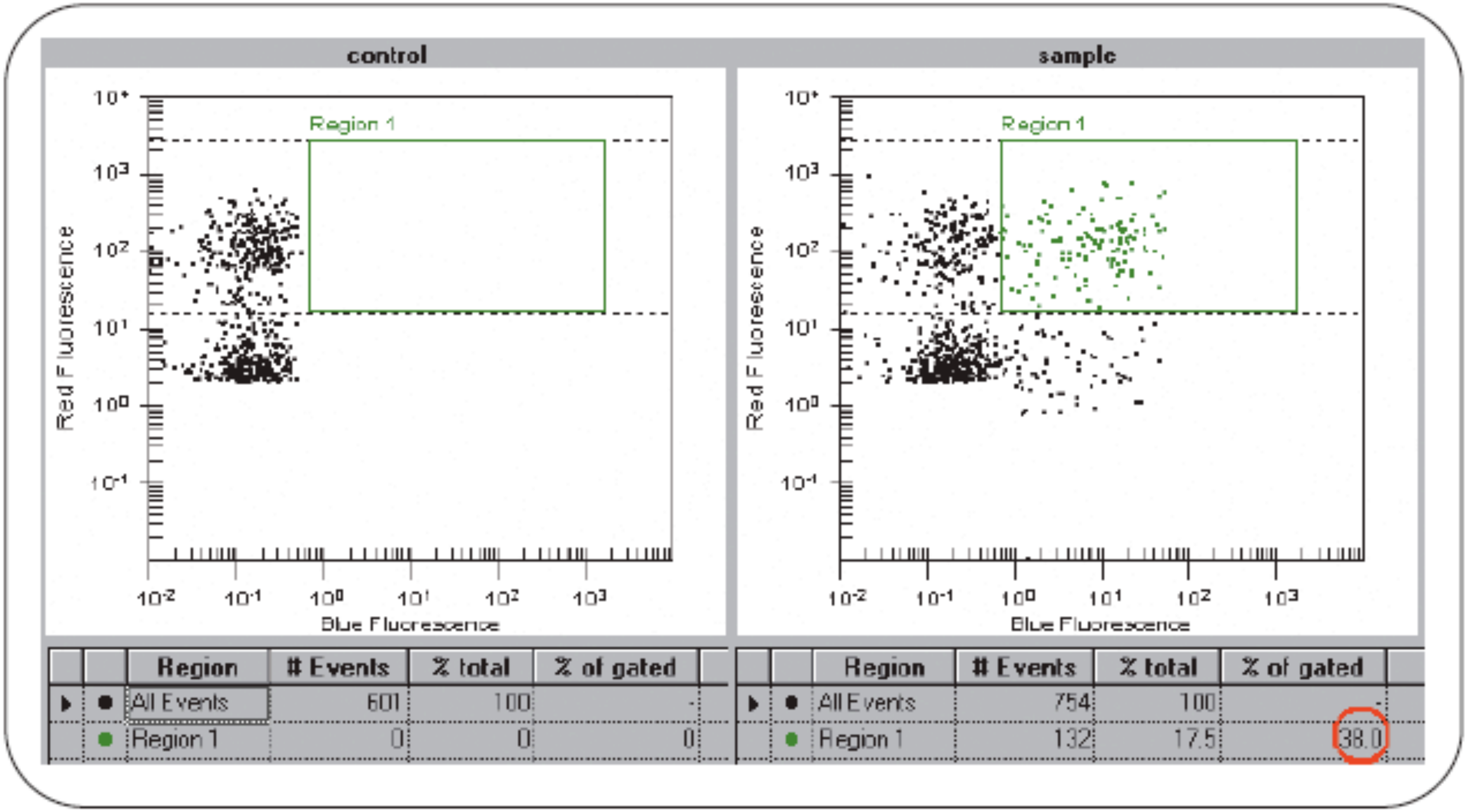
Today and Thursday's Experiments

- Transfections today
- Measure fluorescence via Bioanalyzer on Thursday

Thursday's Experiments: Bioanalyzer

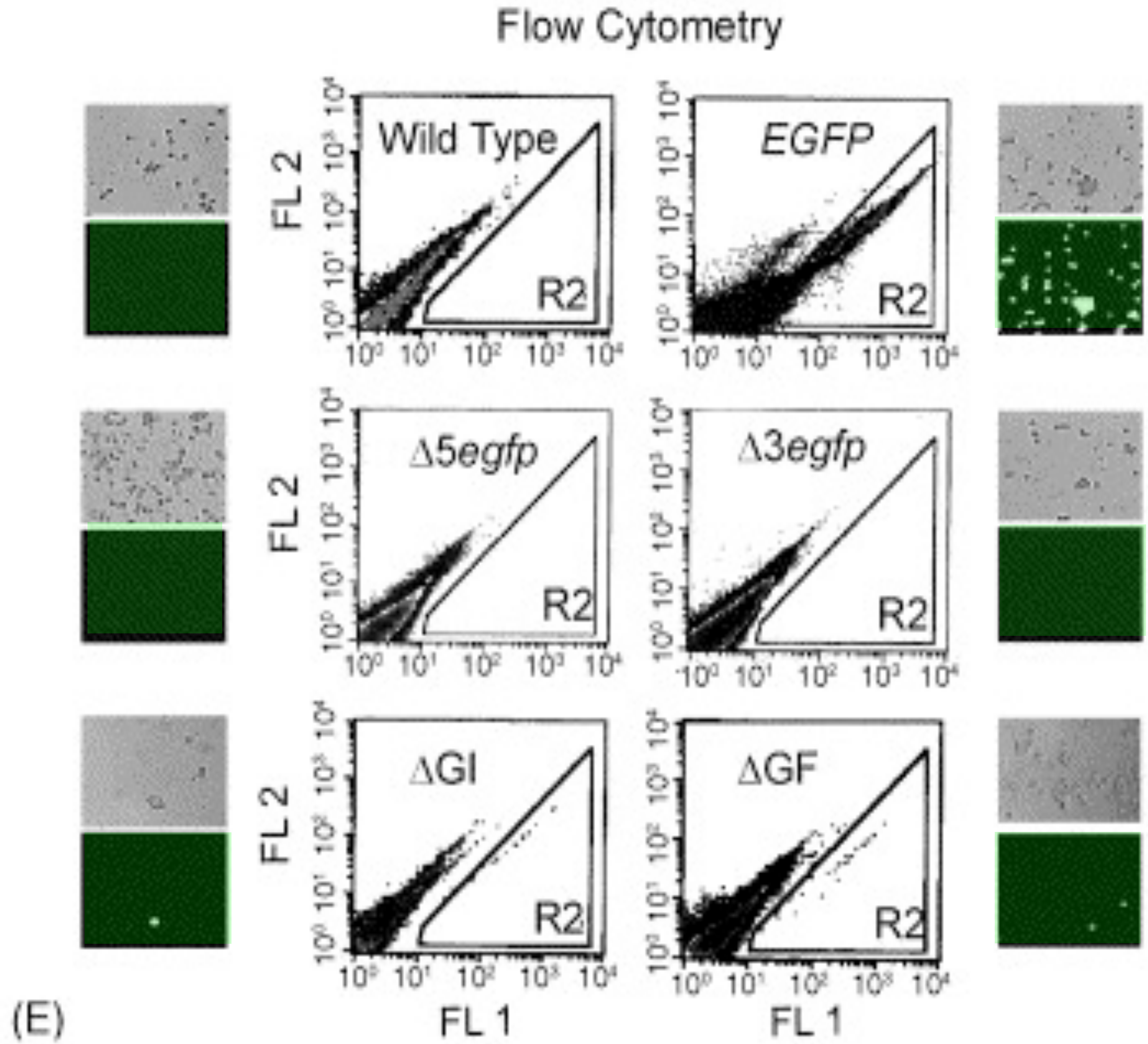


Bioanalyzer Output



FACS Data

Targeted cells showed green fluorescence via flow cytometry at expected frequency.



Jonnalagadda, *et al*. 2005 *DNA Repair*. (4) 594-605.

FACS vs. Bioanalyzer

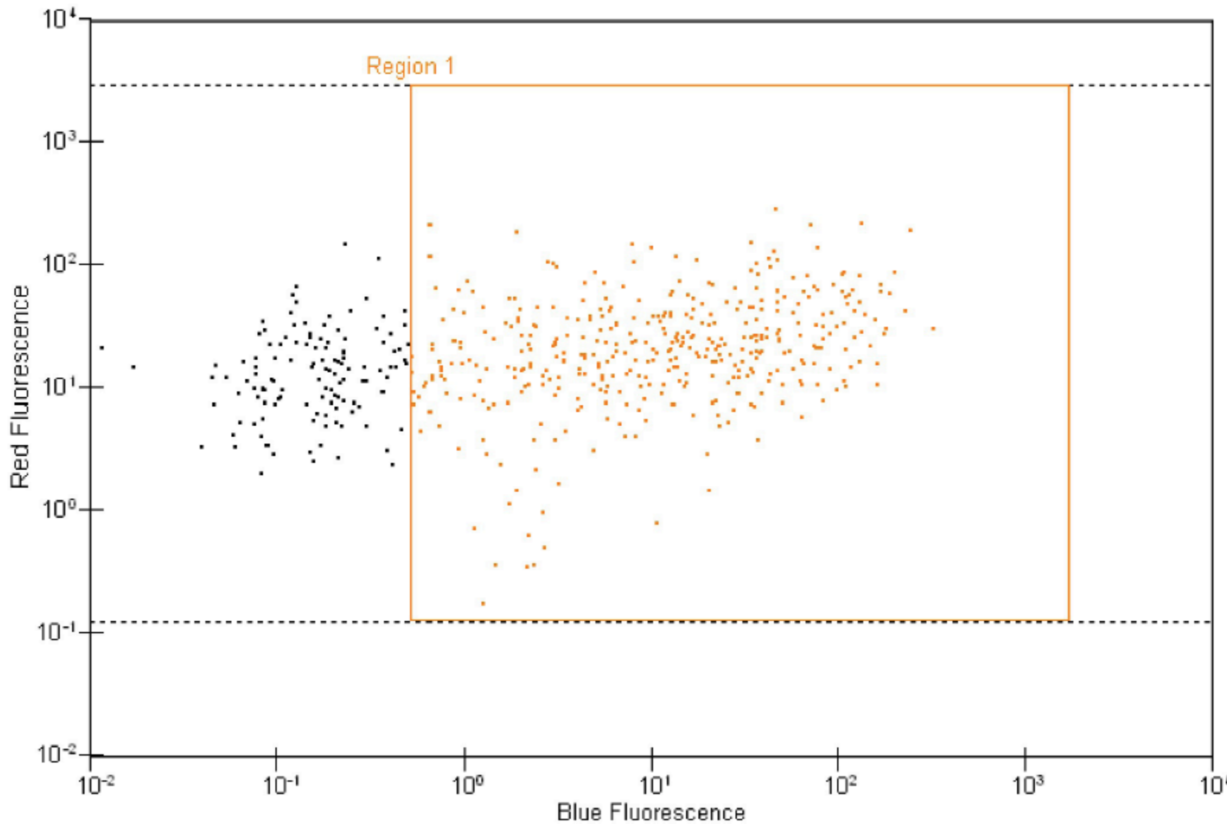
- Ultimate readout will be fluorescence intensity in red and green channels for each cell

- FACS measures thousands of events, while the Bioanalyzer measures hundreds

- What can this mean for your statistics???

Example Bioanalyzer Data

- Live cells will be labeled red, HR cells will also be green
- Positive Control

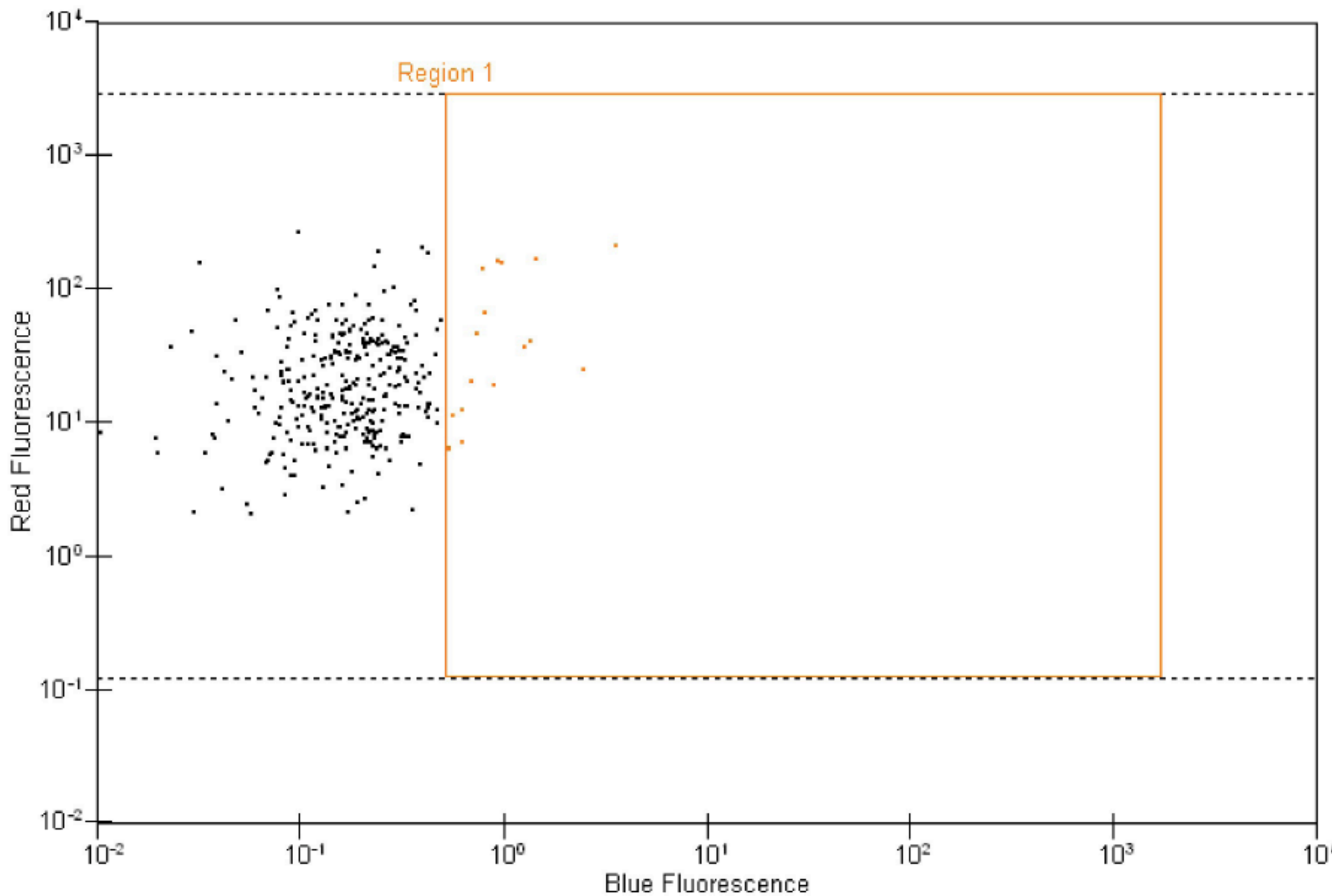


Dot plot statistics for sample 6 : **Sample 6**

Region	XMean	YMean	#Events	%Total	% of gated	StdDevX	StdDevY	CV%X	CV%Y	X GMean	Y GMean
All Events	22.56	27.51	546	100.00	N/A	40.10	31.49	177.80	114.49	3.58	17.38
Region 1	30.07	31.02	408	74.70	74.70	43.86	34.19	145.88	110.22	10.94	19.52

Example Bioanalyzer Data

- Live cells will be labeled red, HR cells will also be green
- Possible Experimental Sample Output



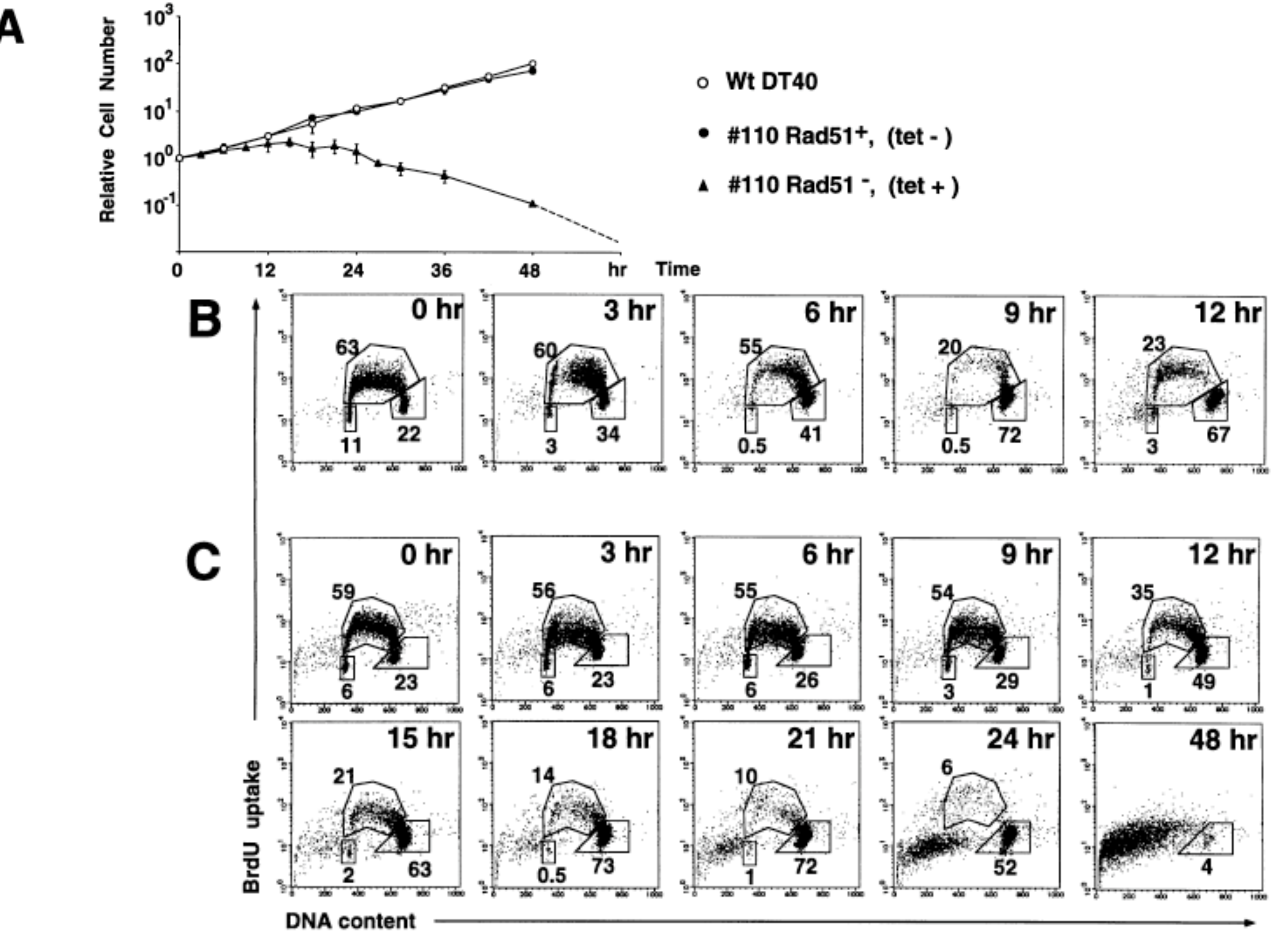
Excel Example: Day 8 Results

	Fluorescence Intensity: EGFP	
Cell	Δ3	Δ3 + Δ5
1	25	22
2	22	25
3	27	87
4	38	105
5	32	200
6	21	22
7	48	23
8	15	48
9	26	320
10	22	29
.	.	.
.	.	.
.	.	.

Conclusion

- Due to the nature of the data
  - Look at gating for individual cell data
  - Consider a Gaussian distribution for significance when comparing across conditions and groups
- Think about how much data you have within each population and use different distributions to think about certainty in your data

Extra Slides



A
Relative Cell Number
Time
hr
Wt DT40
#110 Rad51+, (tet -)
#110 Rad51 -, (tet +)
B
0 hr
3 hr
6 hr
9 hr
12 hr
C
15 hr
18 hr
21 hr
24 hr
48 hr
BrdU uptake
DNA content

Application

$$H_o : \bar{x}_1 = \bar{x}_2$$
$$H_A : \bar{x}_1 \rangle \bar{x}_2$$

Hawks	Cyclones
9	4
8	6
7	5
6	2
7	4
8	5
X_1 7.5	X_2 4.3
s_1 1.0	s_2 1.4

- $$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$t = \frac{7.5 - 4.3}{1.2}\sqrt{\frac{6 \times 6}{6 + 6}}$$

$$t = 4.6$$

- $$s = \sqrt{\frac{\sum_{set1}(x_i - \bar{x}_1)^2 + \sum_{set2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

$$s = 1.2$$

$t_{calc} = 4.6$

$t_{95} = 2.2$

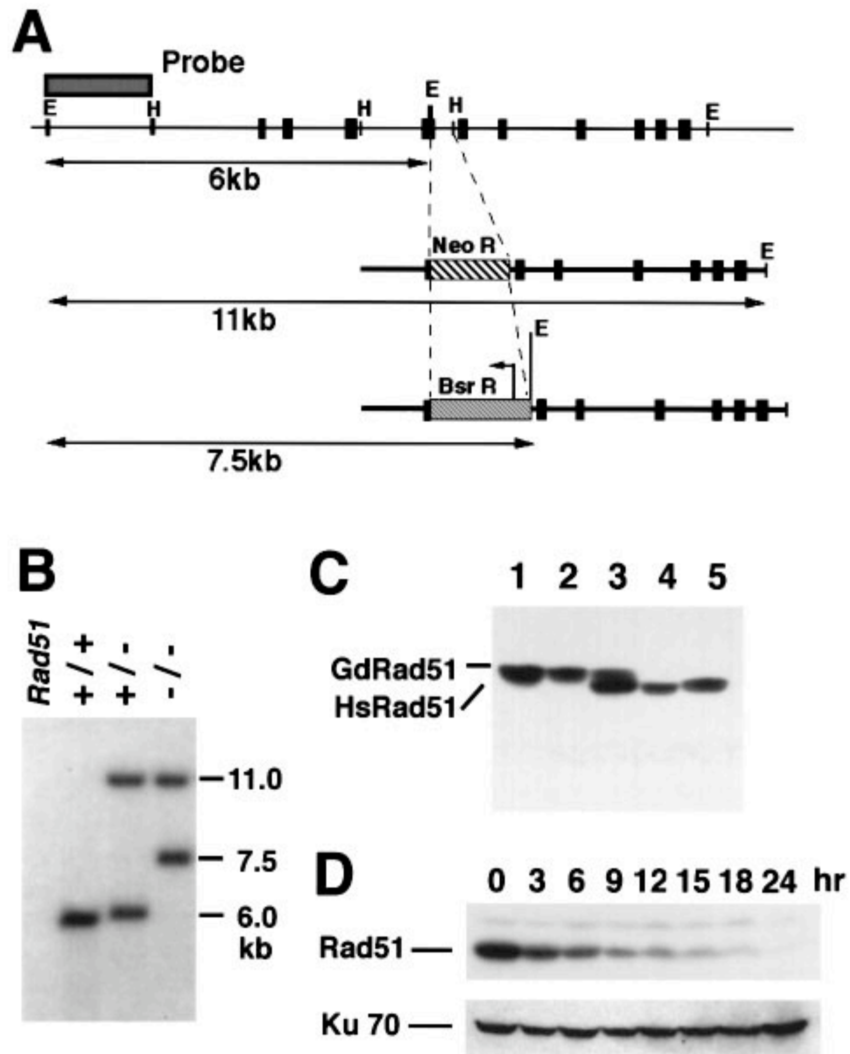
$t_{calc} > t_{95}$

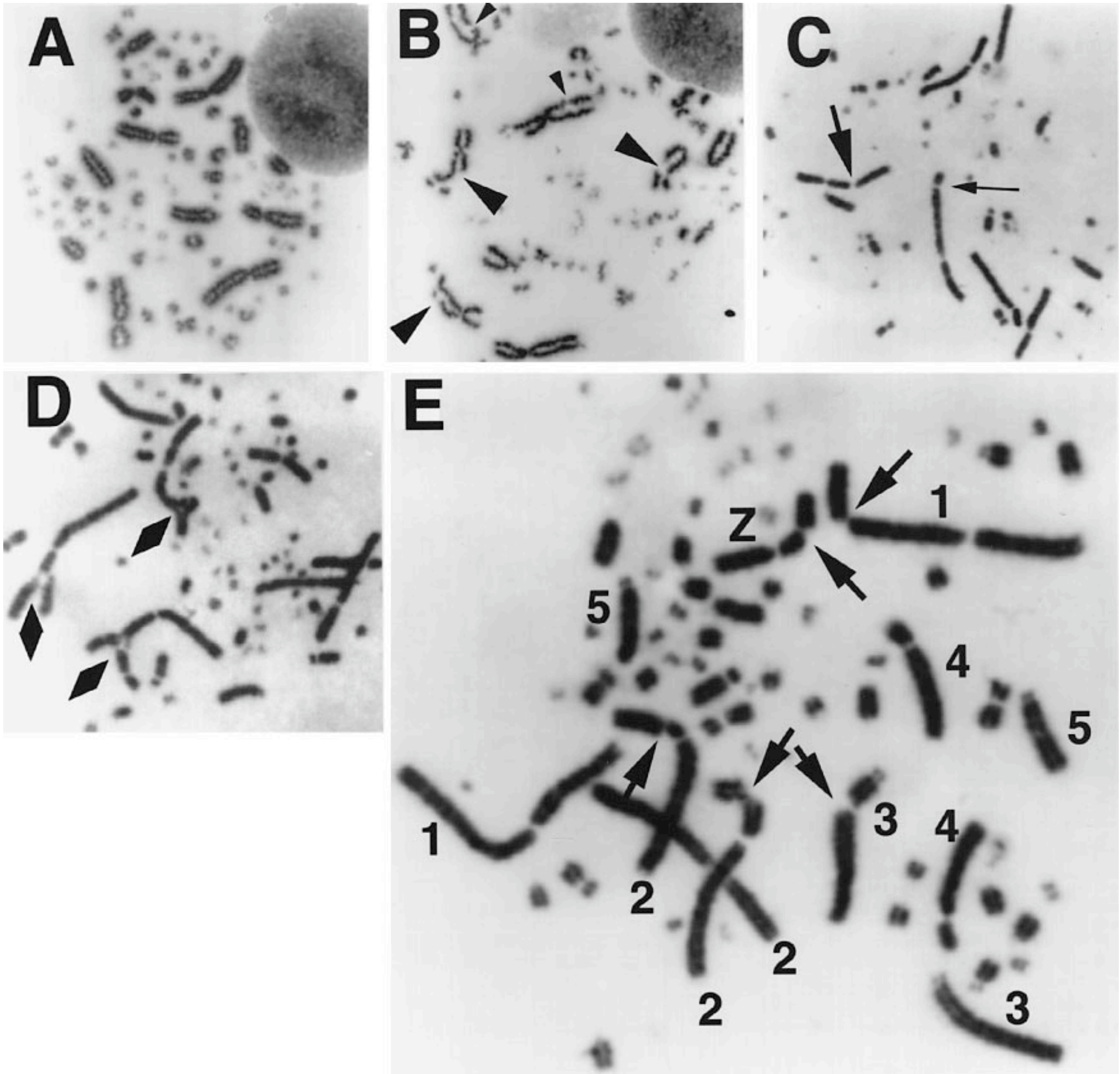
Go to table in notes to find t_{95} with 11 degrees of freedom (12-1)

(The excel sheet does a different comparison)

HAWKS WIN!

Figure 2


A
Probe
E
H
H
E
H
E
6kb
Neo R
E
11kb
E
Bsr R
7.5kb
B
Rad51
+/+
+/-
-/-
11.0
7.5
6.0
kb
C
1 2 3 4 5
GdRad51
HsRad51
D
0 3 6 9 12 15 18 24 hr
Rad51
Ku 70


A
B
C
D
E
1
Z
5
4
5
1
3
4
2
2
2
3