## statistics for designers

dr. carman neustaedter

## how can I find out what % are red?

## what if there are too many to count? how can I find out the % red?

## count a sample and generalize

## how can we improve the sample?

## take more samples

**is the % red in one container the same as another?**

**variability:** the difference between multiple measurements from a population

**statistics** is about the relationship between samples and a population

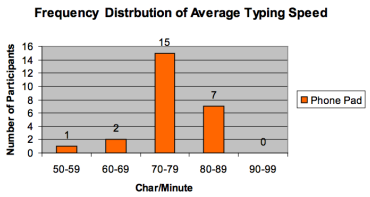
**two types of statistics**

**descriptive stats:** describe data
e.g., graphs, variability, central tendency

**inferential stats:** make inferences (conclusions) about the data
e.g., a sample generalizes to the population

**descriptive stat examples**

**frequency distribution:** plot of how frequently each value appears in the data for each level of the IV




**central tendency**

**mean:** add up all measurements, divide by total number of measurements

**median:** the measurement with half of the measurements above it and half below it

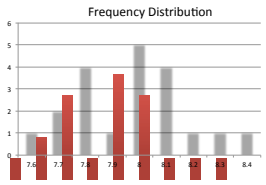
**mode:** the most frequent measurement

**mode?**
median?
mean?

| Value |
|---|
| 7.88 |
| 7.84 |
| 7.88 |
| 8.37 |
| 8.02 |
| 8.18 |
| 8.04 |
| 8.02 |
| 8.19 |
| 8.04 |
| 7.79 |
| 8.19 |
| 8.4 |
| 8.13 |
| 7.72 |
| 7.64 |
| 8.23 |
| 7.84 |
| 7.99 |
| 8.07 |

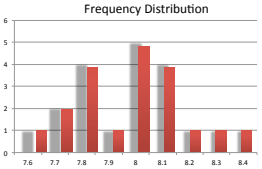
**mode?**
median?
mean?

| Value |
|---|
| 7.88 |
| 7.84 |
| 7.88 |
| 8.37 |
| 8.02 |
| 8.18 |
| 8.04 |
| 8.02 |
| 8.19 |
| 8.04 |
| 7.79 |
| 8.19 |
| 8.4 |
| 8.13 |
| 7.72 |
| 7.64 |
| 8.23 |
| 7.84 |
| 7.99 |
| 8.07 |




**mode?**
median?
mean?

| Value |
|---|
| 7.88 |
| 7.84 |
| 7.88 |
| 8.37 |
| 8.02 |
| 8.18 |
| 8.04 |
| 8.02 |
| 8.19 |
| 8.04 |
| 7.79 |
| 8.19 |
| 8.4 |
| 8.13 |
| 7.72 |
| 7.64 |
| 8.23 |
| 7.84 |
| 7.99 |
| 8.07 |




mode is the 8.0-8.9 group

mode?
**median?**
mean?

| Value |
|---|
| 7.88 |
| 7.84 |
| 7.88 |
| 8.37 |
| 8.02 |
| 8.18 |
| 8.04 |
| 8.02 |
| 8.19 |
| 8.04 |
| 7.79 |
| 8.19 |
| 8.4 |
| 8.13 |
| 7.72 |
| 7.64 |
| 8.23 |
| 7.84 |
| 7.99 |
| 8.07 |

mode?
**median?**
mean?

| Value | |
|---|---|
| 7.88 | 7.64 |
| 7.84 | 7.72 |
| 7.88 | 7.79 |
| 8.37 | 7.84 |
| 8.02 | 7.84 |
| 8.18 | 7.88 |
| 8.04 | 7.88 |
| 8.02 | 7.99 |
| 8.19 | 8.02 |
| 8.04 | 8.02 |
| 7.79 | 8.04 |
| 8.19 | 8.04 |
| 8.4 | 8.07 |
| 8.13 | 8.13 |
| 7.72 | 8.18 |
| 7.64 | 8.19 |
| 8.23 | 8.19 |
| 7.84 | 8.23 |
| 7.99 | 8.37 |
| 8.07 | 8.4 |

mode?
**median?**
mean?

| Value | |
|---|---|
| 7.88 | 7.64 |
| 7.84 | 7.72 |
| 7.88 | 7.79 |
| 8.37 | 7.84 |
| 8.02 | 7.84 |
| 8.18 | 7.88 |
| 8.04 | 7.88 |
| 8.02 | 7.99 |
| 8.19 | 8.0 |
| 8.04 | 8.02 |
| 7.79 | 8.04 |
| 8.19 | 8.04 |
| 8.4 | 8.07 |
| 8.13 | 8.13 |
| 7.72 | 8.18 |
| 7.64 | 8.19 |
| 8.23 | 8.19 |
| 7.84 | 8.23 |
| 7.99 | 8.37 |
| 8.07 | 8.4 |

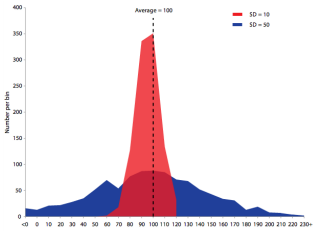
8.03 would give 10 above it and 10 below it

mode?
median?
**mean?**

| Value |
|---|
| 7.88 |
| 7.84 |
| 7.88 |
| 8.37 |
| 8.02 |
| 8.18 |
| 8.04 |
| 8.02 |
| 8.19 |
| 8.04 |
| 7.79 |
| 8.19 |
| 8.4 |
| 8.13 |
| 7.72 |
| 7.64 |
| 8.23 |
| 7.84 |
| 7.99 |
| 8.07 |

= 160.46 / 20 = 8.02

## is central tendency enough?




no, these data sets have the same mean but different variability

we also need to describe what the **variability** is

## measures of variability

**range:** the smallest score subtracted from the largest score

e.g., 8.40 – 7.64 = 0.76
"median of 8.03 with range of 0.76"

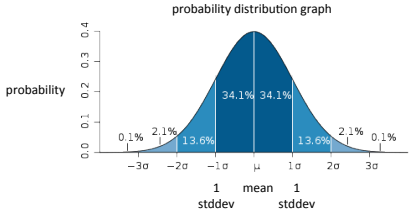
| Value |
|---|
| 7.88 |
| 7.84 |
| 7.88 |
| 8.37 |
| 8.02 |
| 8.18 |
| 8.04 |
| 8.02 |
| 8.19 |
| 8.04 |
| 7.79 |
| 8.19 |
| 8.4 |
| 8.13 |
| 7.72 |
| 7.64 |
| 8.23 |
| 7.84 |
| 7.99 |
| 8.07 |

## measures of variability

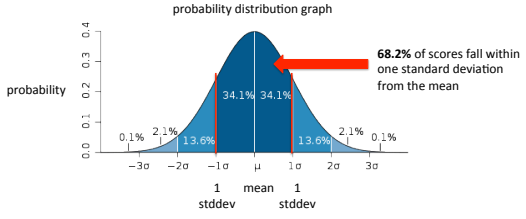
**variance:** average of the squared difference between the scores and the mean

**standard deviation:** the square root of the variance
e.g., or the average distance of each data point from the mean

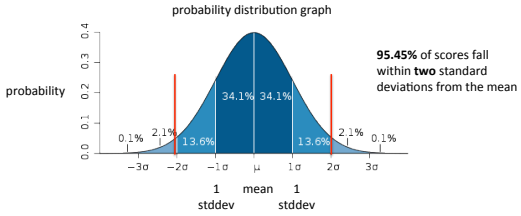
## standard deviation




## standard deviation

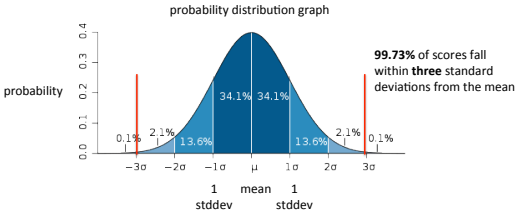


## standard deviation




**95.45%** of scores fall within **two** standard deviations from the mean

## standard deviation



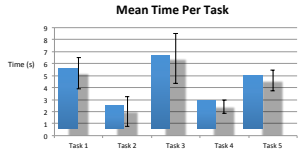

**99.73%** of scores fall within **three** standard deviations from the mean

## graphing mean and standard deviation

| | | | | | | average | stdev | plus stdev | minus stdev |
|---|---|---|---|---|---|---|---|---|---|
| Task 1 | 4 | 5 | 6 | 7 | 4 | 5.2 | 1.30 | 6.50 | 3.90 |
| Task 2 | 1 | 2 | 4 | 2 | 1 | 2 | 1.22 | 3.22 | 0.78 |
| Task 3 | 8 | 9 | 4 | 6 | 5 | 6.4 | 2.07 | 8.47 | 4.33 |
| Task 4 | 2 | 2 | 2 | 3 | 3 | 2.4 | 0.55 | 2.95 | 1.85 |
| Task 5 | 4 | 5 | 4 | 6 | 4 | 4.6 | 0.89 | 5.49 | 3.71 |
| | P1 | P2 | P3 | P4 | P5 | | | | |
| Time for each task | | | | | | | | | |

## graphing mean and standard deviation

| | | | | | | average | stdev | plus stdev | minus stdev |
|---|---|---|---|---|---|---|---|---|---|
| Task 1 | 4 | 5 | 6 | 7 | 4 | 5.2 | 1.30 | 6.50 | 3.90 |
| Task 2 | 1 | 2 | 4 | 2 | 1 | 2 | 1.22 | 3.22 | 0.78 |
| Task 3 | 8 | 9 | 4 | 6 | 5 | 6.4 | 2.07 | 8.47 | 4.33 |
| Task 4 | 2 | 2 | 2 | 3 | 3 | 2.4 | 0.55 | 2.95 | 1.85 |
| Task 5 | 4 | 5 | 4 | 6 | 4 | 4.6 | 0.89 | 5.49 | 3.71 |
| | P1 | P2 | P3 | P4 | P5 | | | | |
| Time for each task | | | | | | | | | |




**68.2%** of scores fall within one standard deviation from the mean

## differences between means

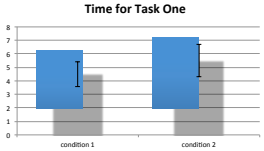
**condition one:**
3, 4, 4, 4, 5, 5, 5, 6

**condition two:**
4, 4, 5, 5, 6, 6, 7, 7

## differences between means

**condition one:**
3, 4, 4, 4, 5, 5, 5, 6

**condition two:**
4, 4, 5, 5, 6, 6, 7, 7




mean = 4.5 +/ 0.9  mean = 5.5 +/ 1.2

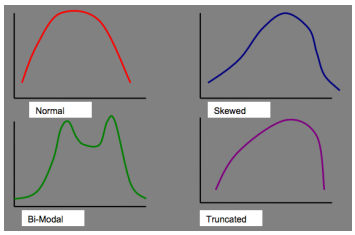
is there a **statistically significant** difference between the means?

**t-test**

**a simple statistical test:** allows one to say something about differences between means at a certain confidence level

**different types of distributions**

t-tests work on normal distributions




**t-test**

**null hypothesis of the t-test:**
no difference exists between the means of two sets of collected data

possible results:

1) I am 95% sure that the null hypothesis is rejected
e.g., there is probably a true difference between the means

2) I cannot reject the null hypothesis
e.g., the means are likely the same

**levels of significance**

how certain do we want to be?

95% = $p < 0.05$ (there is a 5% chance we are wrong)
5 out of 100 times, this result would occur by chance

99% = $p < 0.01$ (there is a 1% chance we are wrong)
1 out of 100 times, this result would occur by chance

in HCI, we typically pick $p < 0.05$ or $p < 0.01$
decide on the p level before testing

**calculating the t-test**

compute a t-test using a stats package
e.g., Excel, SPSS, JMP

1. select a p value you want to test against, e.g., 0.05

2. compare each condition's column of measurements in the test

3. t-test gives you a p-value

| Number of Subjects n | Typing Speed Ave Char/min Phonepad | Typing Speed Ave Char/min Alphabetic |
|---|---|---|
| 1 | 80 | 69 |
| 2 | 77 | 73 |
| 3 | 78 | 70 |
| 4 | 79 | 71 |
| 5 | 81 | 74 |
| 6 | 65 | 73 |
| 7 | 66 | 70 |
| 8 | 67 | 68 |
| 9 | 64 | 72 |
| 10 | 66 | 71 |
| 11 | 66 | 72 |
| 12 | 56 | 77 |
| 13 | 67 | 69 |
| 14 | 82 | 69 |
| 15 | 82 | 70 |
| 16 | 80 | 71 |
| 17 | 66 | 72 |
| 18 | 67 | 72 |
| 19 | 69 | 70 |
| 20 | 56 | 72 |

**calculating the t-test**

**significance:** if $p < 0.05$, there is a statistically significant difference between the two samples.

if your null-hypothesis said there was no difference (no effect), then **reject it**

| Number of Subjects n | Typing Speed Ave Char/min Phonepad | Typing Speed Ave Char/min Alphabetic |
|---|---|---|
| 1 | 80 | 69 |
| 2 | 77 | 73 |
| 3 | 78 | 70 |
| 4 | 79 | 71 |
| 5 | 81 | 74 |
| 6 | 65 | 73 |
| 7 | 66 | 70 |
| 8 | 67 | 68 |
| 9 | 64 | 72 |
| 10 | 66 | 71 |
| 11 | 66 | 72 |
| 12 | 56 | 77 |
| 13 | 67 | 69 |
| 14 | 82 | 69 |
| 15 | 82 | 70 |
| 16 | 80 | 71 |
| 17 | 66 | 72 |
| 18 | 67 | 72 |
| 19 | 69 | 70 |
| 20 | 56 | 72 |

## calculating the t-test

**significance:** if $p < 0.05$, there is a statistically significant difference between the two samples.

if your null-hypothesis said there was no difference (no effect), then **reject it**

**no significance:** if $p >= 0.05$, there is **not** a statistically significant difference between the two samples. But we **cannot** say they are the same.

if your null-hypothesis said there was no difference (no effect), then you cannot **reject it.** you also cannot accept it.

| Number of Subjects n | Typing Speed Ave Char/min Phonepad | Typing Speed Ave Char/min Alphabetic |
|---|---|---|
| 1 | 80 | 69 |
| 2 | 77 | 73 |
| 3 | 78 | 70 |
| 4 | 79 | 71 |
| 5 | 81 | 74 |
| 6 | 65 | 73 |
| 7 | 66 | 70 |
| 8 | 67 | 68 |
| 9 | 64 | 72 |
| 10 | 66 | 71 |
| 11 | 66 | 72 |
| 12 | 56 | 77 |
| 13 | 67 | 69 |
| 14 | 82 | 59 |
| 15 | 82 | 70 |
| 16 | 80 | 71 |
| 17 | 66 | 72 |
| 18 | 67 | 72 |
| 19 | 69 | 70 |
| 20 | 58 | 72 |

## demo in Excel

## different types of t-tests

**unpaired:** comparing two sets of independent observations
e.g., different subjects in each group (between subjects)

**paired:** comparing two sets of dependent observations
e.g., same subjects in each group (within subjects)

## different types of t-tests

**directional (one-tailed)**: tests if data set A is greater than data set B, but not the other way

**non-directional (two tailed):** tests both directions

most commonly we use a two-tailed test because our distributions are symmetric around the mean

## types of errors

**type 1 – false positive:** we see a difference but there isn't really one

e.g., our samples happen to be different because of random sampling. if we selected different samples, we'd see a difference

low confidence level (e.g., $p < 0.1$): greater chance of Type 1 errors

## types of errors

**type 2 – false negative:** we don't see a difference but there really is one

e.g., our samples happen to be the same because of random sampling. if we selected different samples, the samples would be different

high confidence level (e.g., $p < 0.0001$): greater chance of Type 2 errors

what if we want to compare more than two sets of data?

**single factor analysis of variance (ANOVA)**

compare three or more means
e.g., comparing mouse-typing on three keyboards

| qwerty | alphabetic | dvorak |
|---|---|---|
| S1-10 | S11-20 | S21-30 |

**possible results:**
mouse typing is fastest on qwerty keyboard
the same on alphabetic & dvorak

**analysis of variance (ANOVA)**

compares relationships between many factors

| | qwerty | alphabetic | dvorak |
|---|---|---|---|
| cannot touch type | S1-10 | S11-20 | S21-30 |
| can touch type | S31-40 | S41-50 | S51-60 |

when can we use different **descriptive** and **inferential statistics**?

what **type of data** do we need?

**scales of measurements**

**nominal scale:** numbers are used as a name or identifier; no real quantitative properties

can only count the frequencies, no allowable stats

| Jersey Number 22 | First Worn | H/A | Vs. | Comments | Last Worn |
|---|---|---|---|---|---|
| Garth Rizzuto ..... | 1970-71 | | | (?) played 70-71 | 1971/07/01 |
| Jocelyn Guevremont . | 1971/10/08 | H | TOR | First NHL Game (#7) | 1971/12/25 |
| Gregg Boddy ..... | 1971/12/26 | H | CAL | (?) First NHL Game | 1976/07/01 |
| Andy Spruce ..... | 1976-77 | | | (?) played 76-77 | 1977/07/01 |
| Bob Manno ..... | 1977-78 | | | (?) played 76-81 | 1981/07/03 |
| Tiger Williams ..... | 1981-82 | | | (?) played 79-84 | 1984/07/01 |
| Dave Lowry ..... | 1985/10/10 | A | LA | First NHL Game | 1988/07/01 |
| Dan Hodgson ..... | 1988/10/06 | H | WIN | New Number (was 16) | 1989/01/31 |
| Greg C. Adams ..... | 1989/03/08 | H | WIN | First Game as Canuck | 1989/04/15 |
| Craig Coxe ..... | 1989/10/15 | H | BOS | Returns to Canucks | 1991/01/31 |
| Robert Dirk ..... | 1991/03/05 | A | PIT | First Game as Canuck | 1994/03/20 |
| Jeff Brown ..... | 1994/03/25 | H | NYR | First Game as Canuck | 1995/12/18 |
| Markus Naslund ..... | 1996/03/22 | H | DAL | First Game as Canuck | 1996/04/27 |
| Larry Courville .... | 1997/02/11 | H | WAS | New Number (was 14) | 1997/12/31 |
| Peter Zezel ..... | 1998/02/06 | H | EDM | First Game as Canuck | 1999/03/23 |
| Daniel Sedin ..... | 2000/10/05 | A | PHI | First NHL Game | (current) |


**scales of measurements**

**ordinal scale:** numbers can be ordered or ranked, but we don't know the difference between ranks

**allowable stats:**
median & percentiles

e.g., first place runner finished ahead of second place finisher but we don't know by how much they won

e.g., children rate their preference of a new mouse compared to their old one:
1 – worst
2 – not as good
3 – neutral
4 – better
5 – best

**scales of measurements**

**interval scale:** numbers can be ordered and we know the difference between ranks; zero is by convention

**allowable stats:**
mean, standard deviation, variance, range, t-test, ANOVA

e.g., temperature in degrees C or F

e.g., Likert scale (where there is no real "zero" value)
1 – strongly disagree
2 – disagree
3 – neutral
4 – agree
5 – strongly agree


**scales of measurements**

**ratio scale:** interval scale with an absolute, non-arbitrary zero

**allowable stats:**
all those for interval, coefficient of variation

e.g., temperature in degrees K, length, weight, time periods
e.g., how fast did people perform on interface A vs. interface B
e.g., accuracy on interface A vs. interface B

**wrapping up**