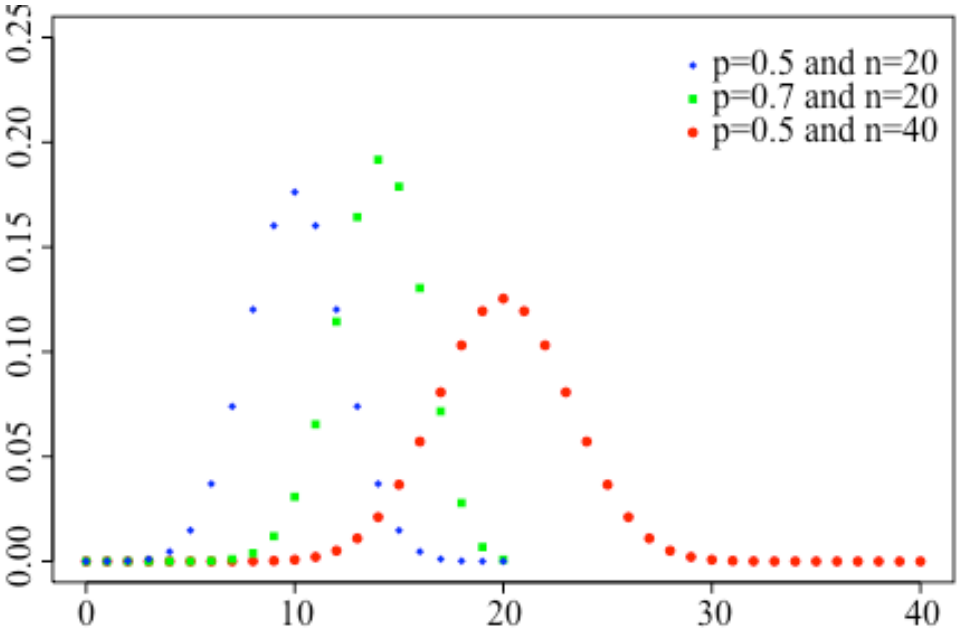
Statistical testing I

Overview: Inferential Statistics

- What is a sampling distribution?
- one sample test: z-score and z-test
  - empirical rule
- null-hypothesis testing
  - Type 1 and 2 errors
  - alpha and power
- one sample test: t-test (n<30)
- ANOVAs

Binomial Distribution



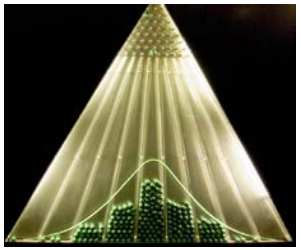
$$P(x) = \frac{N!}{x!(N-x)!}\pi^{x}(1-\pi)^{N-x}$$

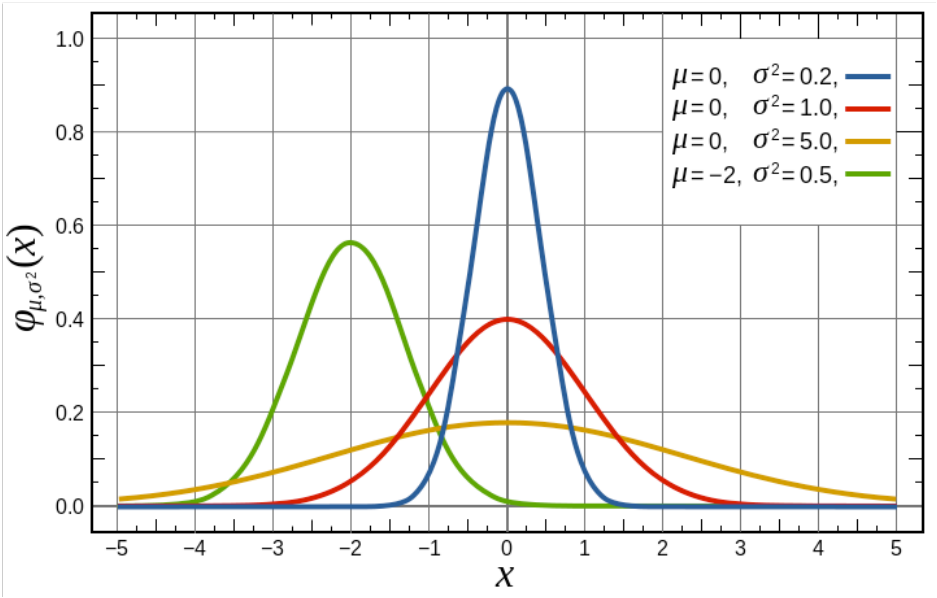
N: number of flips

x: number of desired outcome

π: probability of the desired outcome

Normal Distribution





$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

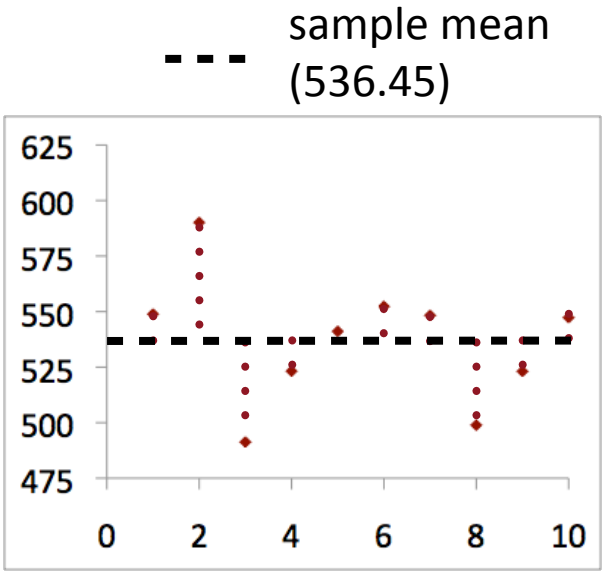
μ = true mean of the distribution

σ^2 = variance of the observations

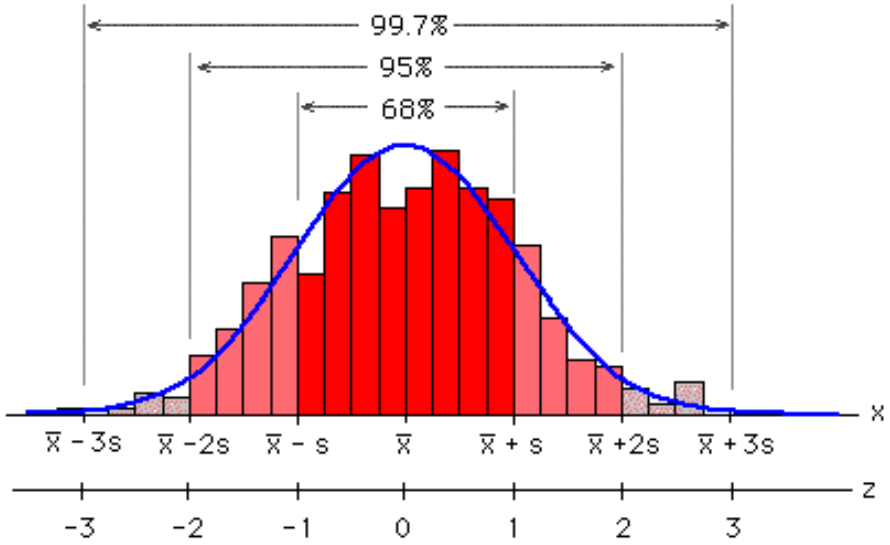
x = value of the observation

Terms

- mean, median, mode
- population vs. sample
- μ and σ^2 and σ
- $\bar{x}$ and s^2 and s
- z-scores



z scores



To look up the area under the curve, to the left of the calculated z-score (e.g., 1.65):

1. look down the row for 1.6
2. look across the columns for 0.05
3. the current z-score is larger/smaller than this proportion of the sampling distribution (e.g., 95.05%)

TABLE A: STANDARD NORMAL PROBABILITIES (CONTINUED)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Exercise: How much taller am I?

Data distribution of the height of German men:
μ: 180 cm
σ^2: 81
Lewis is 198 cm tall.
What is his z-score?

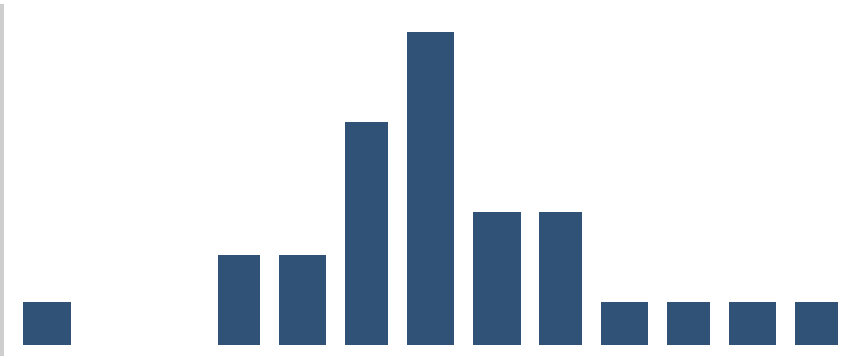
Solution:
$\sigma=9$
z-score = (198-180)/9
= 2.0
Lewis is taller than (95.0+2.5=)97.5% of German men. Or according to the table, 97.72%

TABLE A: STANDARD NORMAL PROBABILITIES (CONTINUED)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Sampling Distribution

- The sampling distribution is the distribution of a sampled statistic (e.g., $\bar{x}$) over many test samples.
- **Central limit theorem** suggests that this is a normal distribution.
- The standard error (of the mean) is the standard deviation of the sampling distribution (of the mean).



Central Limit Theorem

- mean of $\bar{x}$ = μ
- $s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$
- if the population distribution is normal, so will be the distribution of $\bar{x}$ for n.
- For large n, the distribution of $\bar{x}$ is approximately normal regardless of the population distribution

Rule of thumb:
n>30

Estimating the Z-Statistic

- A sample of coffee/day is taken 1 week before the exam
  - N = 49
  - $\bar{x}_{new}$ = 5.5
  - s_{new}=3.5
  - sem = 3.5/√49

- Where is this data sample in the sampling distribution of “normal coffee intake”?

Estimating the Z-Statistic

- Where is this data sample in the sampling distribution of "normal coffee intake"?

- A sample of coffee/day is taken 1 week before the exam.
  - N=49
  - $\bar{x}_{new}$= 5.5; $\bar{x}_{old}$=4.63
  - s_{new}=3.5; σ = 0.5
  - z_{new}= (5.5-4.63)/0.5 = 1.74

$\bar{x}$

The Empirical Rule

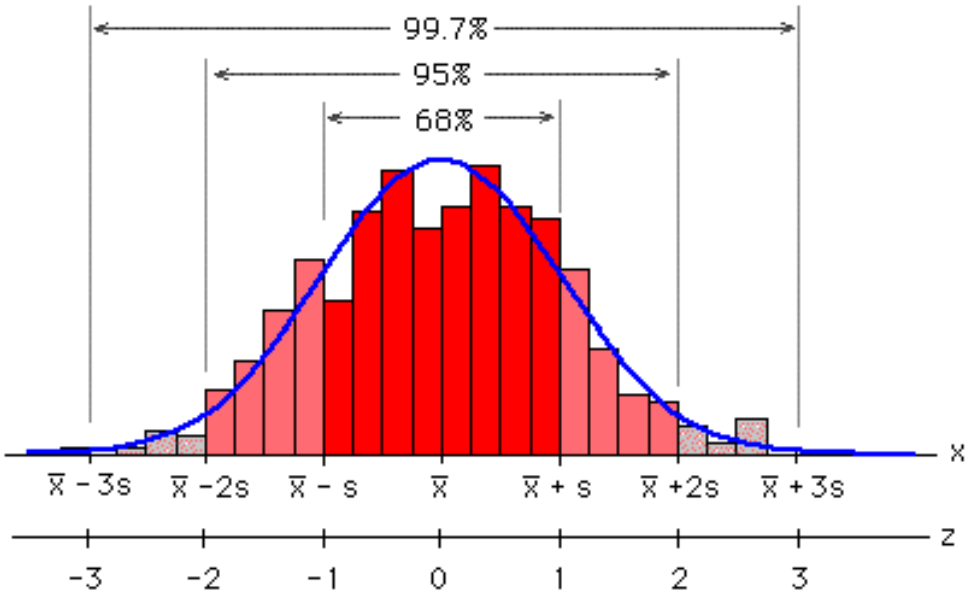


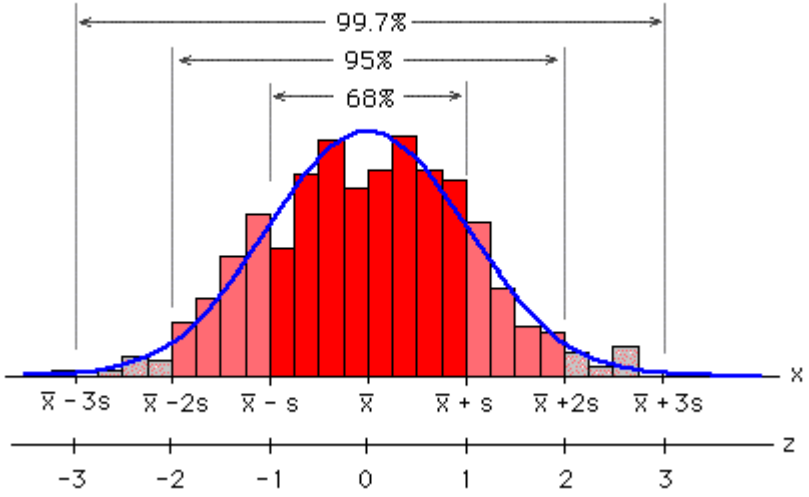
TABLE A: STANDARD NORMAL PROBABILITIES (CONTINUED)

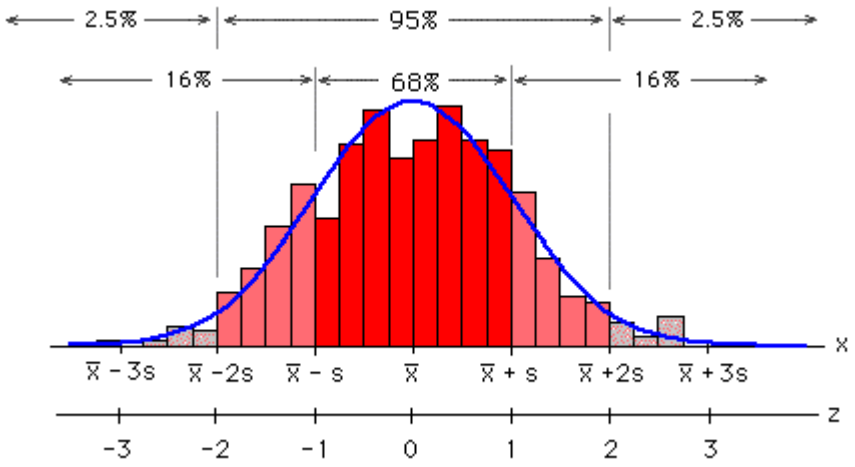
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

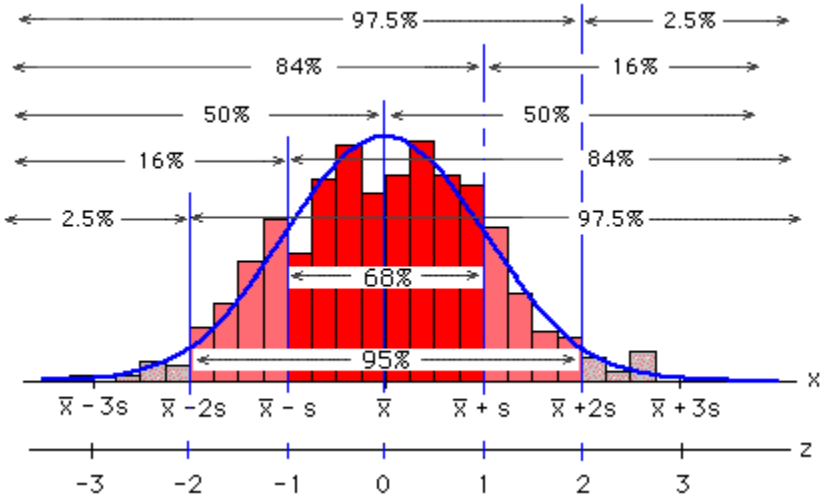
Inference:

The probability that this data came from a distribution of normal coffee intake is XX%

The Empirical Rule



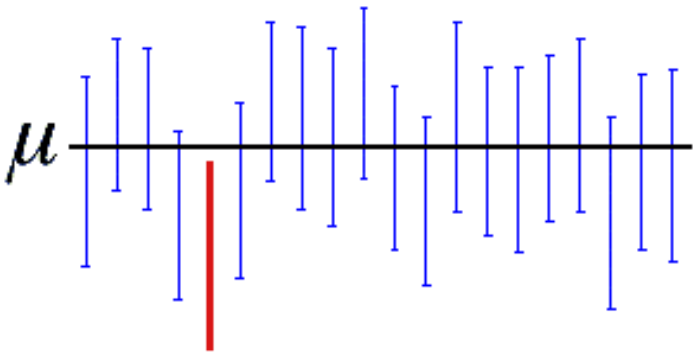




http://www.math.armstrong.edu/statsonline/3/empirical_rule.html

Confidence Intervals

- 95% CI = $\bar{X} \pm 1.96\,\sigma/\sqrt{n}$
- This means:
  If I take 100 samples, the confidence intervals of 95 of them will contain the real $\mu_{\bar{X}}$
- It does NOT mean:
  There is a 95% chance that the real $\mu_{\bar{X}}$ is within this CI.



A 95% confidence interval indicates that 19 out of 20 samples (95%) from the same population will produce confidence intervals that contain the population parameter.

Null-Hypothesis Testing

Your new iPhone game “Sultry Pigeons” has a mean score of 4.3 stars with a standard deviation of 1.5 stars from 100 users. You know that the average game receives 4.0 stars.

Are you the next Zuckerberg?

Null-Hypothesis Testing

Your new iPhone game “Sultry Pigeons” has a mean score of 4.3 stars with a standard deviation of 1.5 stars from 100 users. You know that the average game receives 4.0 stars.
Are you the next Zuckerberg?

H0: My game is no different from most mobile games

H1: My game is different from most mobile games

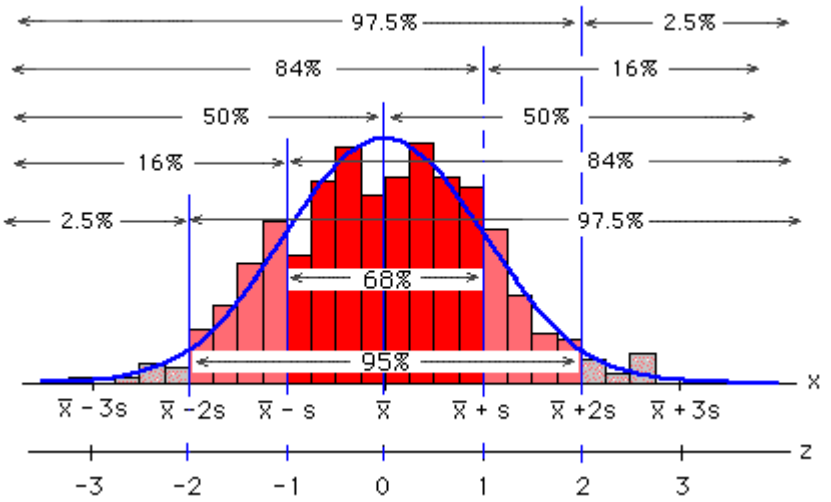
Null-Hypothesis Testing

Your new iPhone game “Sultry Pigeons” has a mean score of 4.3 stars with a standard deviation of 1.5 stars from 100 users. You know that the average game receives 4.0 stars.
Are you the next Zuckerberg?

~~H0: My game is no different from most mobile games~~

H1: My game is different from n

z=2.0; P(D|H0)=0.025



Null-Hypothesis Testing

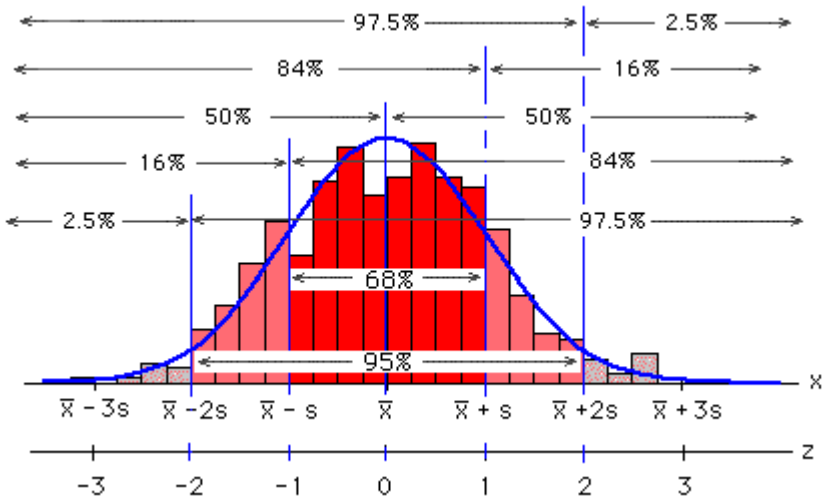
Your new iPhone game “Sultry Pigeons” has a mean score of 4.25 stars with a standard deviation of 1.5 stars from 100 users. You know that the average game receives 4.0 stars. **Are you the next Zuckerberg?**

H0: My game is no better from most mobile games

H1: My game is better from most mobile games

z=1.67; P(D|H0)=0.05

Can we still reject the null-hypothesis now?



Hypothesis testing is educated guessing

H0: My game is no different from most mobile games

H1: My game is different from most mobile games

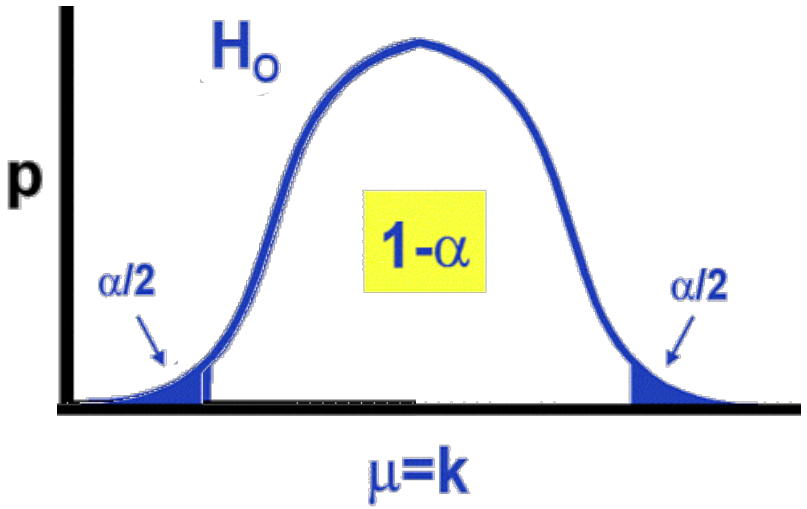
Possibilities		**Actual Situation**	
		H0 is True	H0 is False
Educated Guess	Reject H0	*Type 1 error*	*Correct*
	Do not reject H0	*Correct but so what?*	*Type 2 error*

Hypothesis testing is educated guessing

H0: My game is no different from most mobile games

H1: My game is different from most mobile games

- α-level determines the risk of rejecting the null-hypothesis when H0 is true. (Type 1 error)
- α-level is typically 0.05.
- We decide to reject the null-hypothesis for extreme values larger than this.



http://www4.uwsp.edu/psych/stat/10/hyptestc.htm#cd1

Hypothesis testing is educated guessing

H0: My game is no different from most mobile games

H1: My game is different from most mobile games

- α-level determines the risk of rejecting the null-hypothesis when H0 is true. (Type 1 error)
- β-level determines the risk of accepting the null-hypothesis when H1 is true. (Type 2 error)

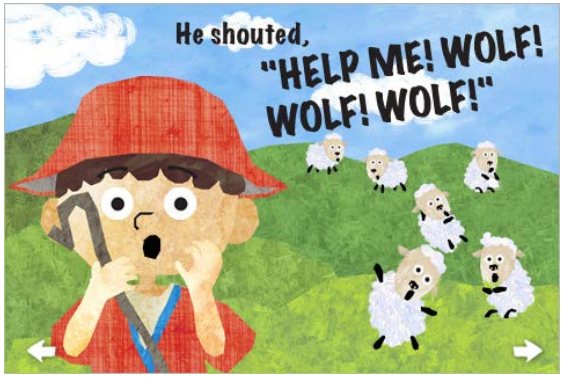
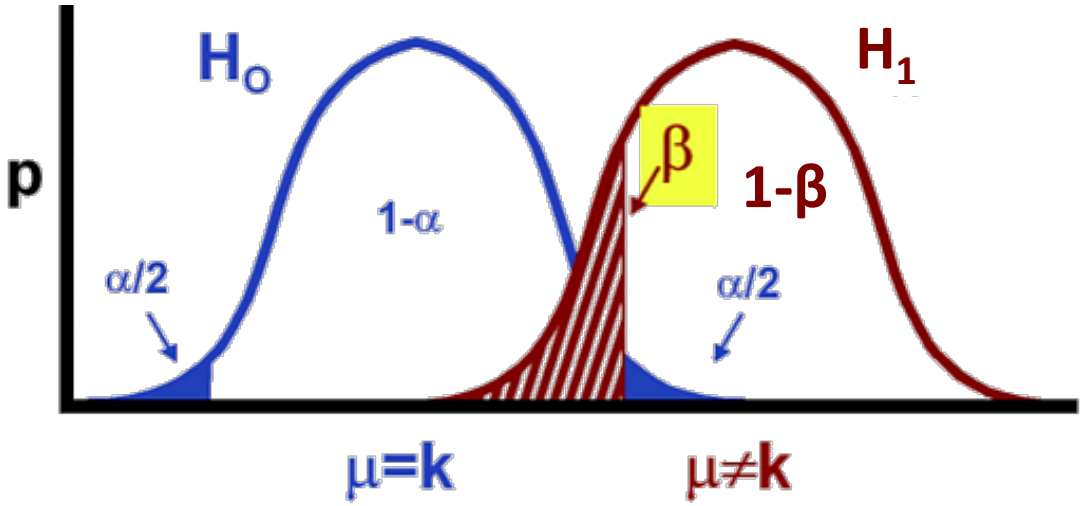


Image: www.kidzstory.com



http://www4.uwsp.edu/psych/stat/10/hyptestc.htm#cd1

Hypothesis testing is educated guessing

H0: My game is no different from most mobile games

H1: My game is different from most mobile games

- α-level determines the risk of rejecting the null-hypothesis when H0 is true. (Type 1 error)
- β-level determines the risk of accepting the null-hypothesis when H1 is true. (Type 2 error)

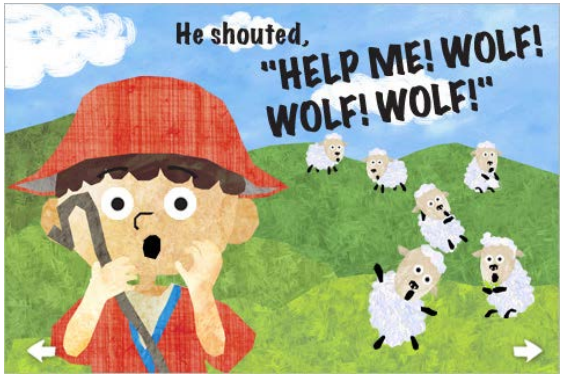
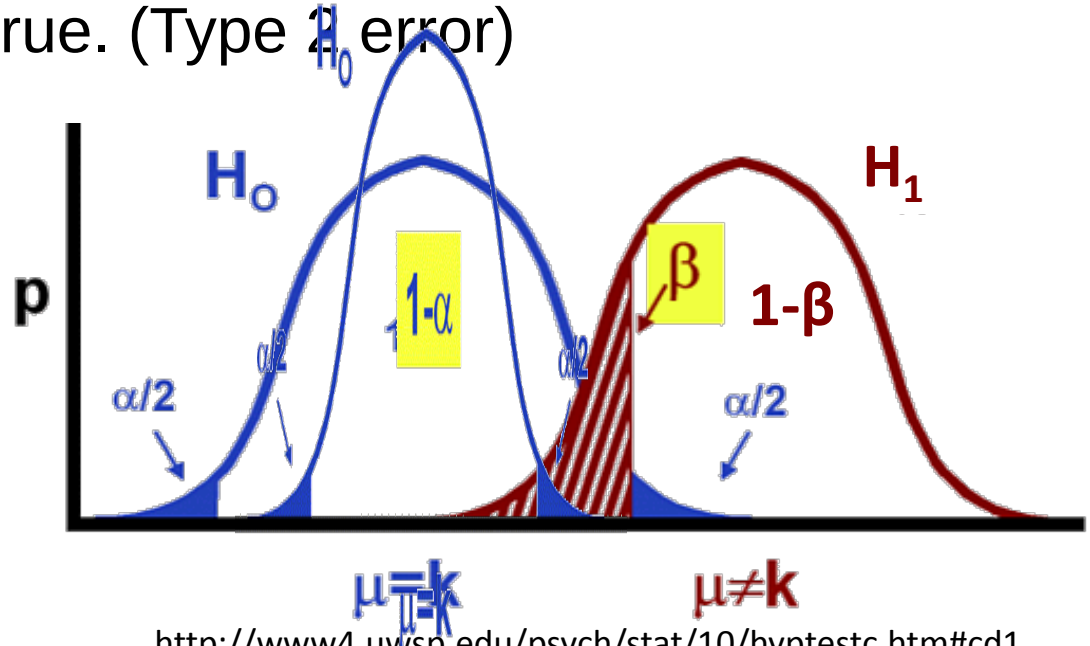


Image: www.kidzstory.com

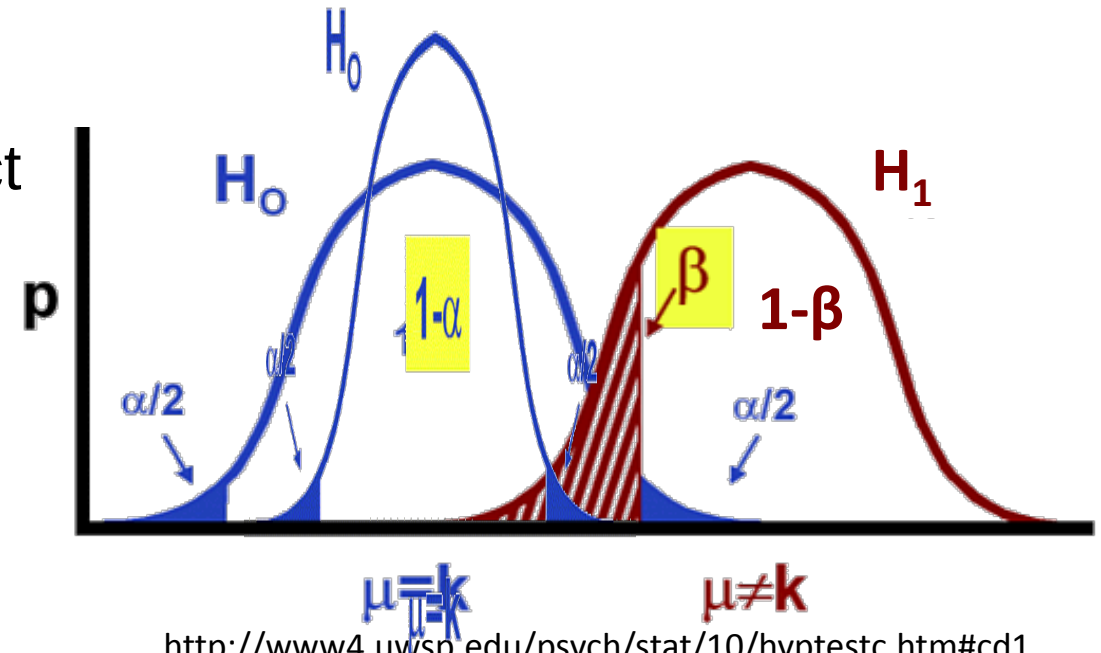


http://www4.uwsp.edu/psych/stat/10/hyptestc.htm#cd1

Power (1-β)

Test sensitivity (1-β) is influenced by:

- alpha (α)
- sample size (N)
- Type of test (parametric or non-parametric)
- Variability
- Test-Directionality
- Robustness of the effect



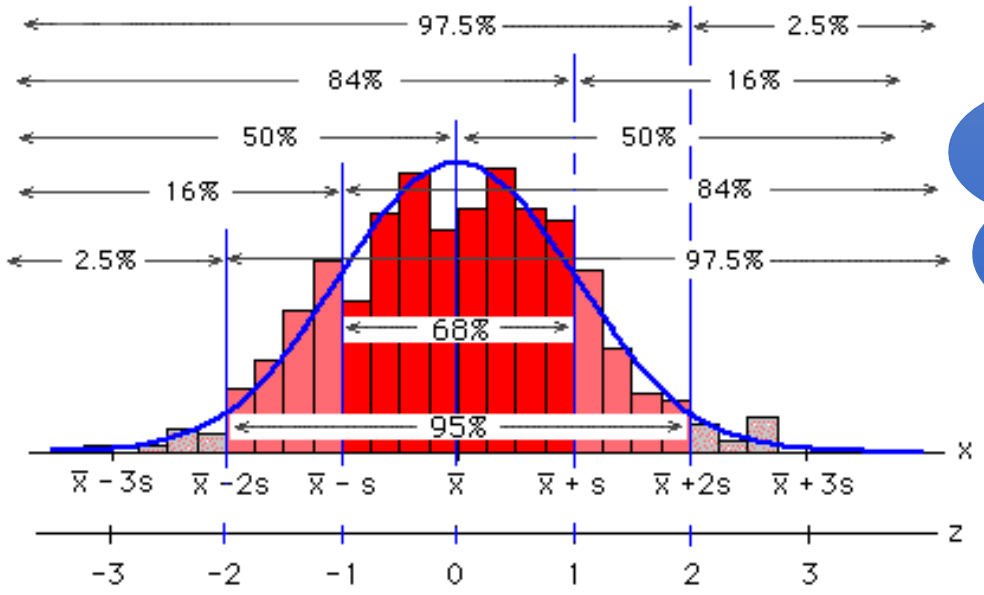
http://www4.uwsp.edu/psych/stat/10/hyptestc.htm#cd1

One-tailed vs Two-tailed testing

H0: My game is no different from most mobile games

H1: My game is different from most mobile games

critical z-score for one-tailed testing: 1.95



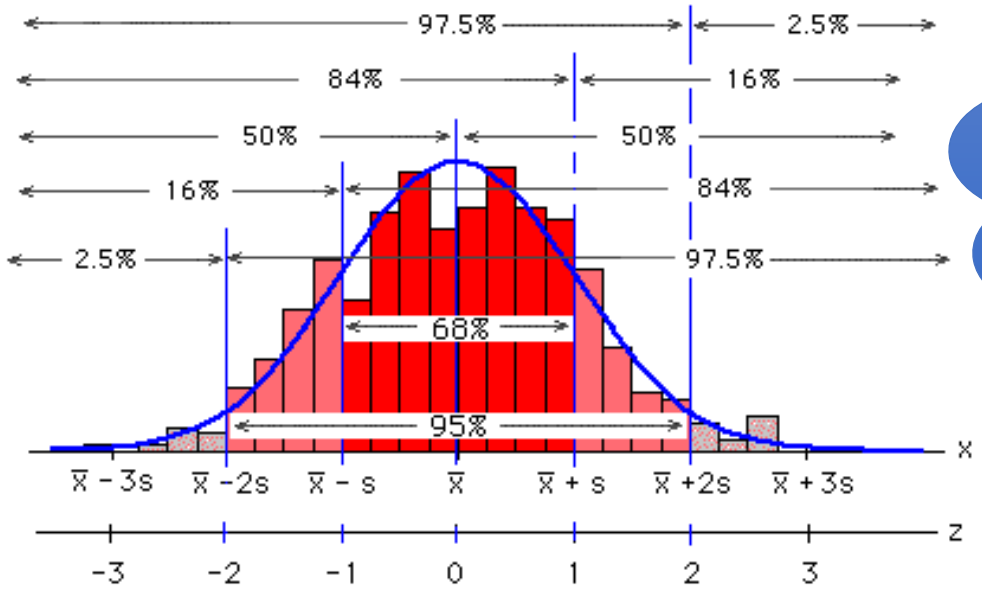
BEFORE I run my test, I have decided that I will reject the null hypothesis so long as the data average is different from the population

One-tailed vs Two-tailed testing

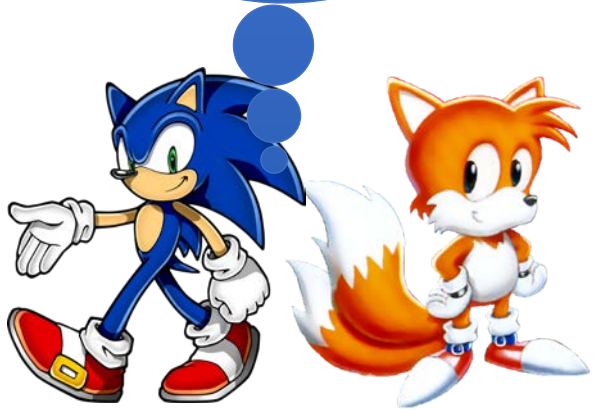
H0: My game is no different from most mobile games

H1: My game is less popular than most mobile games

critical z-score for one-tailed testing, 1.65



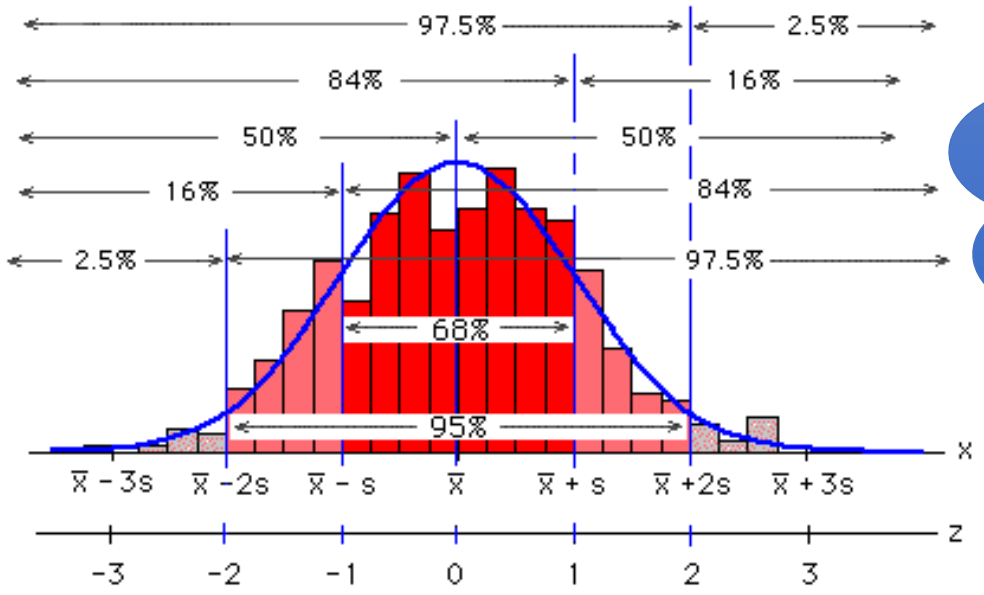
BEFORE I run my test, I have decided that I will reject the null hypothesis only if my data average is *smaller* than the population



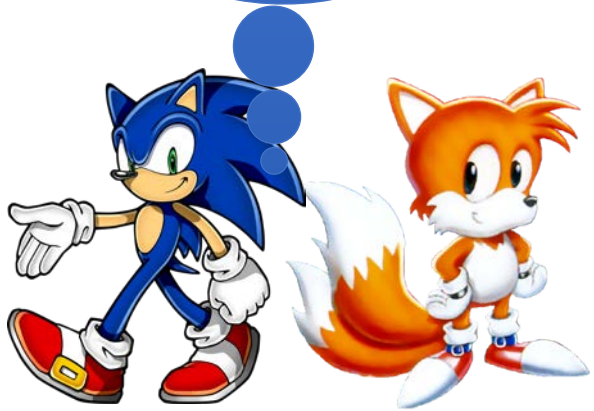
One-tailed vs Two-tailed testing

H0: My game is no different from most mobile games

H1: My game is more popular than most mobile games



BEFORE I run my test, I have decided that I will reject the null hypothesis only if my data average is *larger* than the population

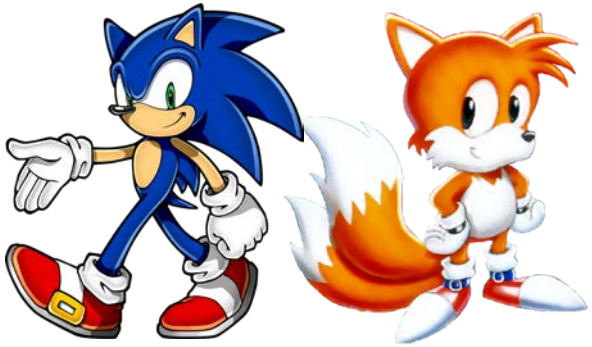


One-tailed vs Two-tailed testing

H0: My game is no different from most mobile games

H1: My game is more popular than most mobile games

If the purpose of the statistical test is one-tailed testing, then it only makes sense to perform inference if the sign of the z-score agrees with the test hypothesis (H1).



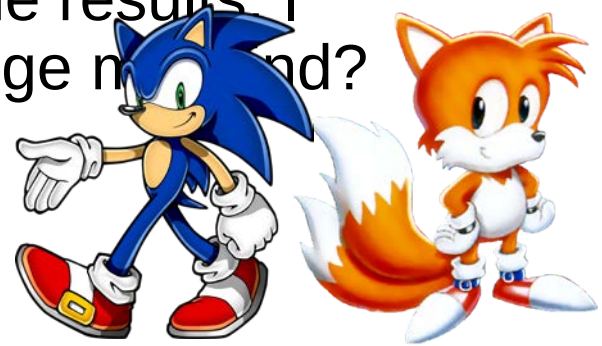
One-tailed vs Two-tailed testing

H0: My game is no different from most mobile games

H1: My game is more popular than most mobile games

If the purpose of the statistical test is one-tailed testing, then it only makes sense to perform inference if the sign of the z-score agrees with the test hypothesis (H1).

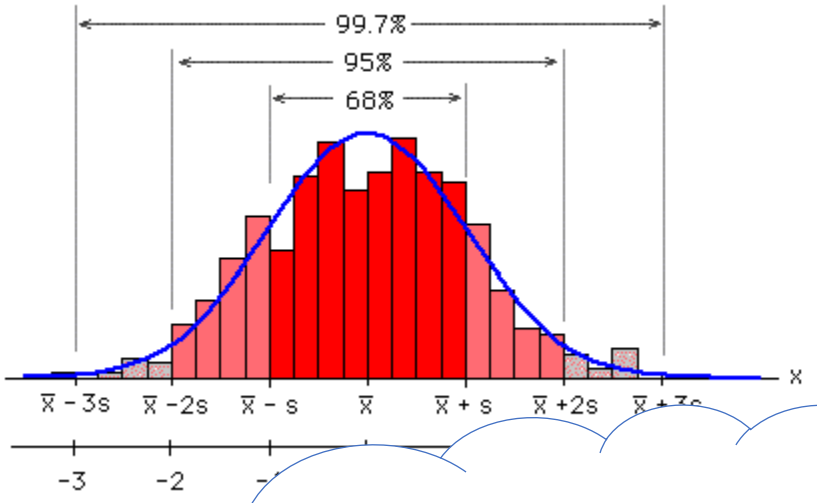
What if I wanted to test it for whether or not it is worse than most games but when I looked at the results, I actually had a positive score. Can I change my mind?




NO

Confirmatory testing: α-level

- Before confirmatory testing, we need to set a criterion for accepting or rejecting H0.
- This critierion is referred to as the α-level (e.g., 0.05*)

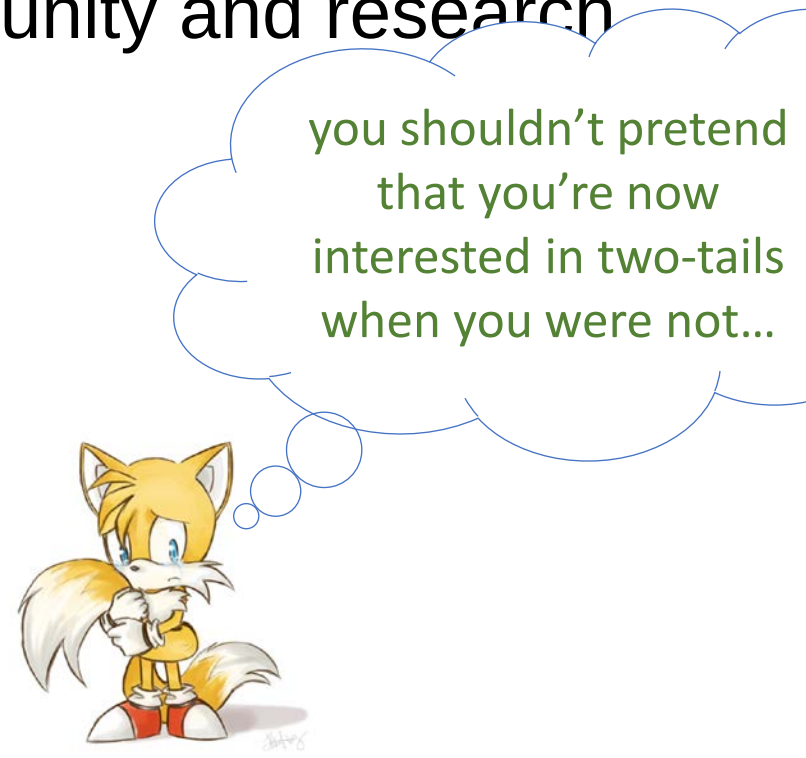


* 0.05 means that 1 out of 20 test samples will results in a false H0 rejection.
The α-level can be anything you want. So long as you can convince your peers and your conscience. If you don't decide the α-level before you decide, then you are not really deciding.

you shouldn't pretend that you're now interested in two-tails when you were not…

Confirmatory testing: α-level

- The z-score is just a statistic. It is not a criterion.
- The critierion is subjective. It is determined by the standards of your community and research field.
- This criterion is the α-level.



Exercise: Are UUstudents better?

The newspaper reports that the mean exam score of German undergraduates is 72. You suspect that UU undergraduates score better than other undergraduates. You sample 100 UU undergraduates (mean=76; s^2=484). What is your new opinion? *(critical z: $z_{0.05}$=1.645; $z_{0.025}$=1.96)*

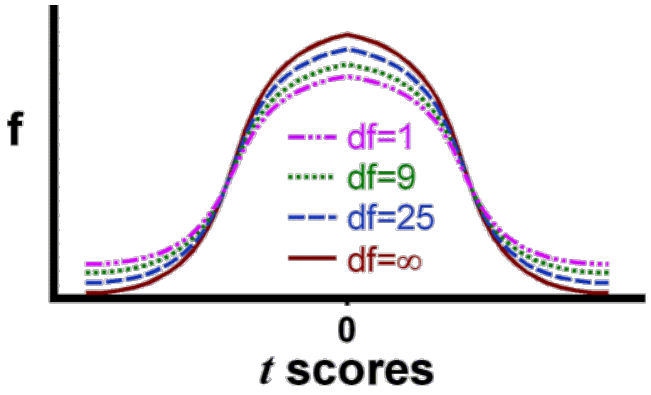
Hints:

1. What is the standard deviation of your test sample?
2. (22) What is the standard error of the mean (aka the standard deviation of the sampling distribution)?
3. (2.2) What is the difference between your sample and the population mean?
4. (4) What is the z-score of your test sample within the null sampling distribution?
5. (1.82) Was this a one-tailed or two-tailed z-test?

One-sample t-test

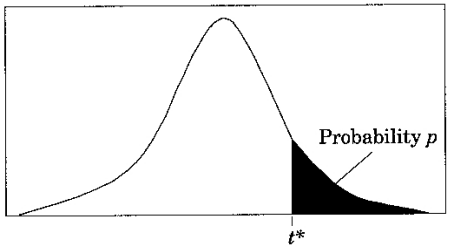
- What if you only tested 20 students?
- The central limit theorem is no longer valid
- Instead of z-distribution, use a t-distribution
- Which t-distribution?
- It depends on the *degrees of freedom* (n-1)

See http://rpsychologist.com/d3/tdist/



Exercise: One-sample t-test

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.



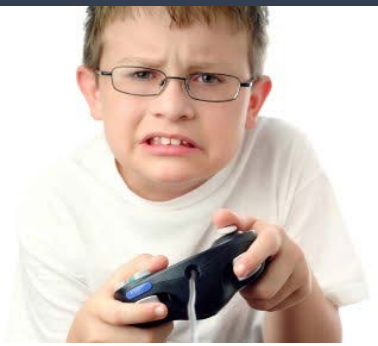
The newspaper reports that the mean exam score of undergraduates is 72. You suspect that UU undergraduates score better than other undergraduates. You sample 20 UU undergraduates (mean=76; s^2=484). What is your new opinion?

Table B t distribution critical values

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

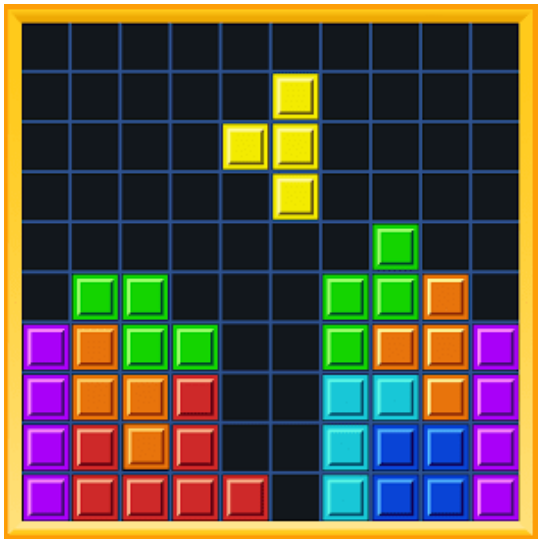
Example Study

Research question: Do different types of computer games result in different levels of stress ?



http://news.cheatcc.com/394703

Doom (c)

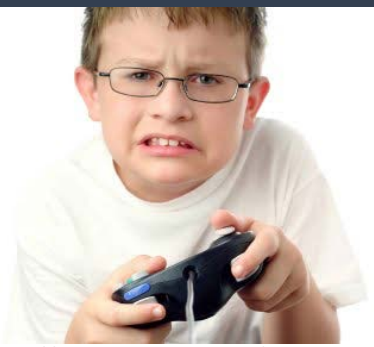


Tetris (c)

Discussion: Design a study for statistical analysis

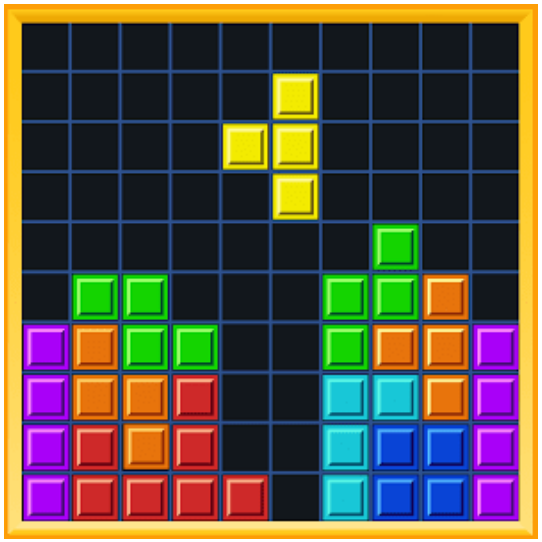
Example Study

Research question: Do different types of computer games result in different levels of stress ?



http://news.cheatcc.com/394703

Doom (c)



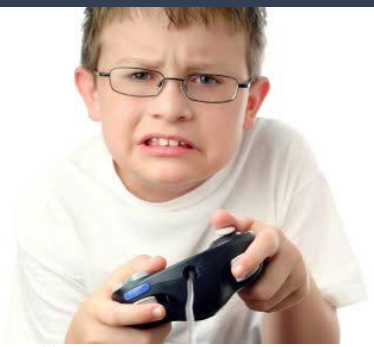
Tetris (c)

IV = ??

DV = ??

Example Study

Research question: Do different types of computer games result in different levels of stress ?



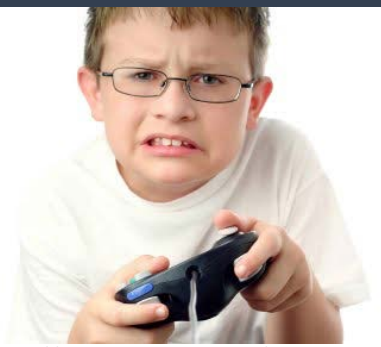
http://news.cheatcc.com/394703

- Recruit 24 students
- Procedure:
  - They will play a shooter (or puzzle) game
  - They will give a stress score (1-10)
  - They will play a puzzle (or shooter) game
  - They will give a stress score (1-10)

- Results:
  - 24 x $Stress_{shooter}$
  - 24 x $Stress_{puzzle}$

Example Study

Research question: Do different types of computer games result in different levels of stress ?



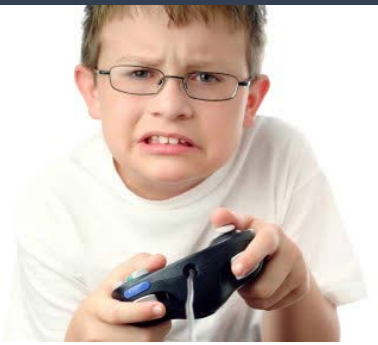
http://news.cheatcc.com/394703

- IV = Game_Type
- DV = Stress_Score

- Results:
  - 24 x $Stress_{shooter}$
  - 24 x $Stress_{puzzle}$

Example Study

Research question: Do different types of computer games result in different levels of stress ?

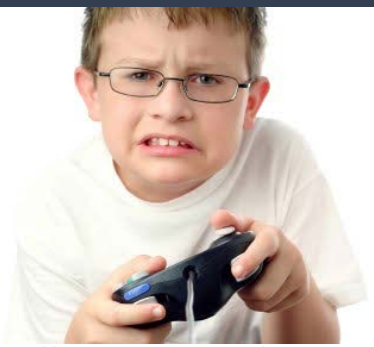


http://news.cheatcc.com/394703

IV = Game_Type
DV = Stress_Score

H0 = ???
H1 = ???

Example Study



http://news.cheatcc.com/394703

Research question: Do different types of computer games result in different levels of stress ?

IV = Game_Type
DV = Stress_Score

$H0 = Stress_{shooter} = Stress_{puzzle}$
$H1 = Stress_{shooter} \neq Stress_{puzzle}$

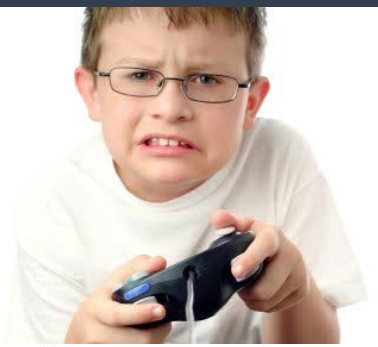
Problem?

- Until now, we compared one sample to the sampling distribution
- Here, we have two samples of stress scores.

Hint:

- We are not really interested in the sampling distribution of stress scores

Example Study



http://news.cheatcc.com/394703

Research question: Do different types of computer games result in different levels of stress ?

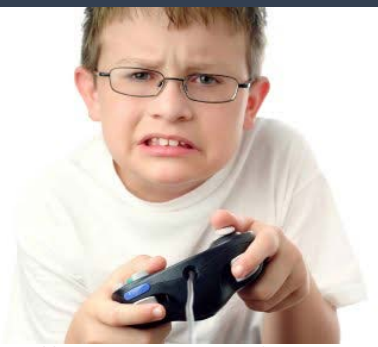
IV = Game_Type
DV = Stress_Score

H0 = ($Stress_{shooter}$ – $Stress_{puzzle}$) is from a sampling distribution with μ=0
H1 = ($Stress_{shooter}$ – $Stress_{puzzle}$) is **NOT** from a sampling distribution with μ=0

[See example, Excel Sheet]

Example Study

Research question: Do different types of computer games result in different levels of stress ?



http://news.cheatcc.com/394703

- Recruit **48** students
- Procedure:
  - One group will play a shooter (or puzzle) game
  - They will give a stress score (1-10)
  - Another group will play a puzzle (or shooter) game
  - They will give a stress score (1-10)

- Results (two independent samples):
  - 24 x $Stress_{shooter}$
  - 24 x $Stress_{puzzle}$

Independent samples t-test

H0 = ($Stress_{shooter}$ – $Stress_{puzzle}$) is from a sampling distribution with μ=0

H1 = ($Stress_{shooter}$ – $Stress_{puzzle}$) is **NOT** from a sampling distribution with μ=0

Problems:

- Which $Stress_{shooter}$ should be deleted from which $Stress_{puzzle}$
- What is the degree of freedom?
  - (N1-1) + (N2-1)
- What is the standard error of the mean (of $Stress_{shooter}$ – $Stress_{puzzle}$)
  - $$\sqrt{(\text{SE for the first quantity})^2 + (\text{SE for the second quantity})^2}$$
- What is the t-score?
  - $$\frac{\text{observed difference - expected difference}}{\text{SE for difference}}$$

Evidence?

Do you have enough evidence to support your hypothesis?!?!

Effect Sizes (Cohen's d)

- d = (diff. of means)/(pooled s.d.)

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}.$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- 0.2 (small), 0.5 (medium), 0.8 (large)

Correct procedure

- Come up with a research question
- Determine the IV and DV
- Determine the number of participants (<30?)
- Determine if it is paired or independent samples test
- Phrase your hypothesis as:
  - null-hypothesis
  - alternative hypothesis
- Decide on α-level
- Decide if it is a one-tail or two-tails test
  - determine the critical t-score

OKAY

Time for our little experiment

YAY!

One-way Analysis of Variance

1. One-way ANOVA
   - determines differences between means of two or more levels of an independent variable

You investigate if game genres have an influence on subjective anxiety. Your independent variable is _____, with 3 _____. Your _____ variable is a self-reported score on a questionnaire for anxiety. You have ____ experimental conditions and run 10 participants per condition. This is a ____ group experiment design. Your null hypothesis (H0) is that _____.

One-way Analysis of Variance

1. One-way ANOVA
   - determines differences between means of two or more levels of an independent variable

You investigate if game genres have an influence on subjective anxiety. Your independent variable is "GameType", with 3 levels. Your dependent variable is a self-reported score on a questionnaire for anxiety. You have three experimental conditions and run 10 participants per condition. This is a between group experiment design. Your null hypothesis (H0) is that the data of the three conditions come from the SAME sampling distribution.

One-way Analysis of Variance

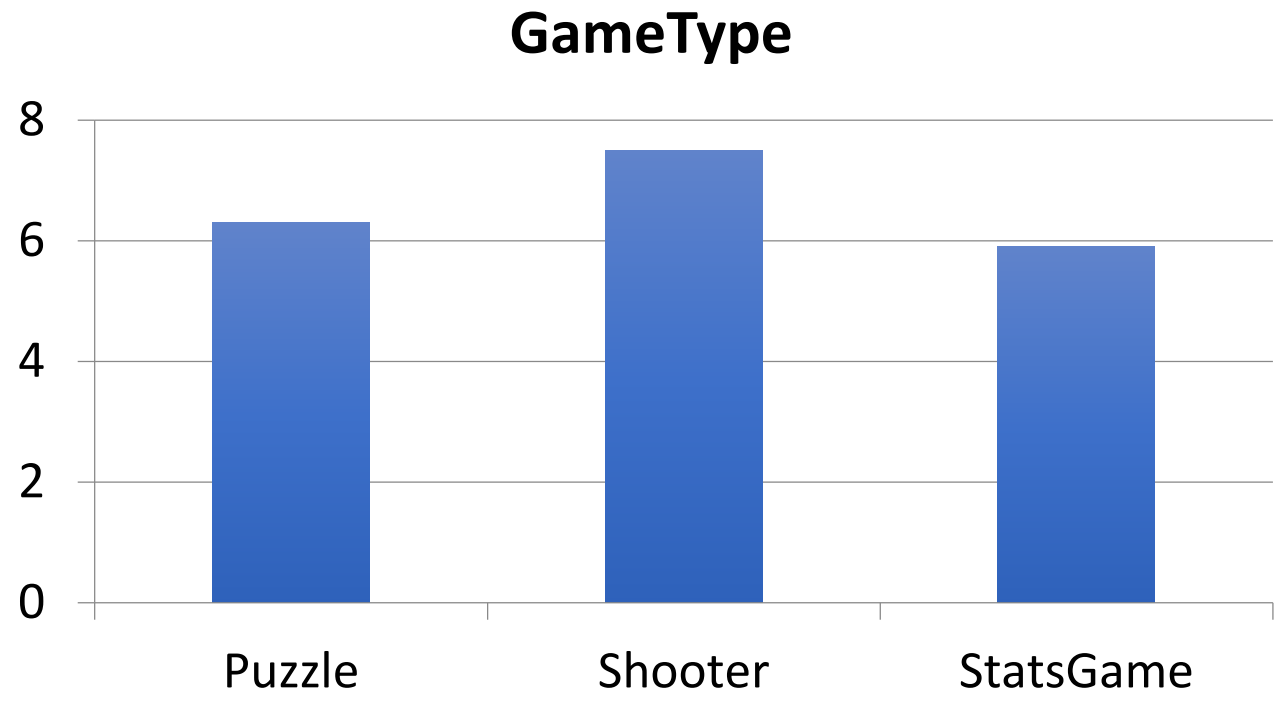
- determines differences between means of two or more levels of an independent variable

N=30	Puzzle	Shooter	StatsGame
Mean Anxiety Score	6.3	7.5	5.9

H0: All populations that were sampled have the same μ

H1: At least one population is different from the others.

One-way ANOVA (aka F-test)



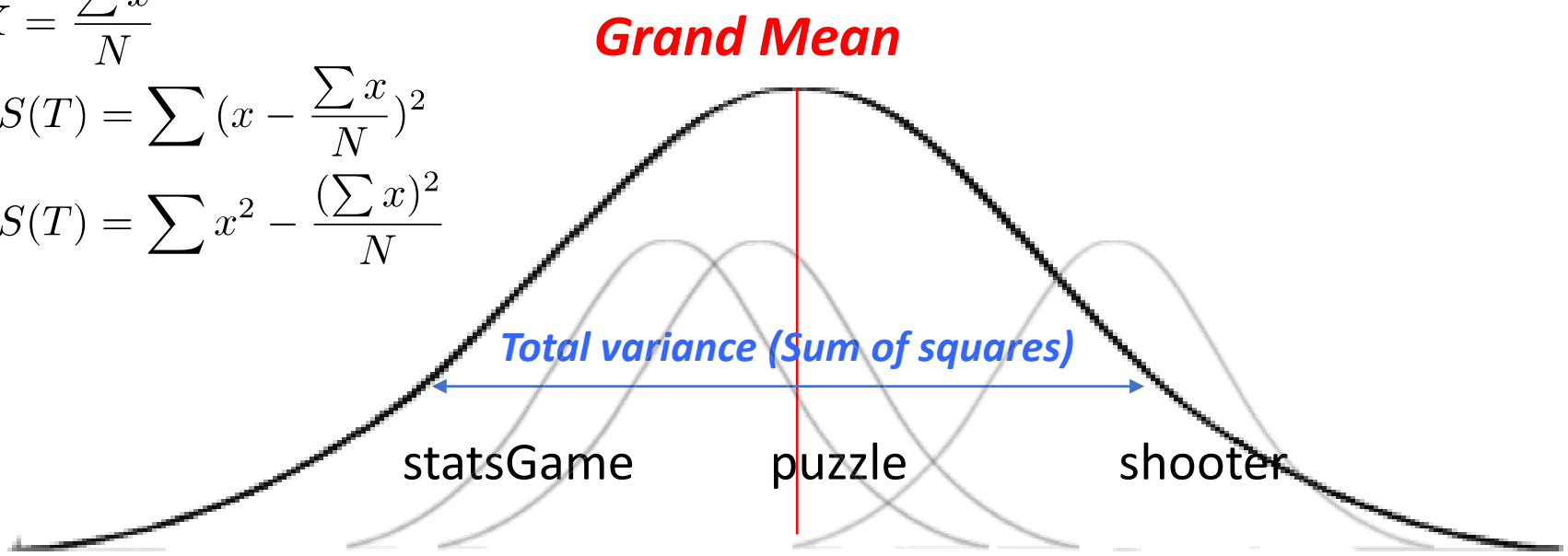
N=30	Puzzle	Shooter	StatsGame
Mean Anxiety Score	6.3	7.5	5.9

One-way ANOVA (aka F-test)

$$\bar{X} = \frac{\sum x}{N}$$

$$SS(T) = \sum (x - \frac{\sum x}{N})^2$$

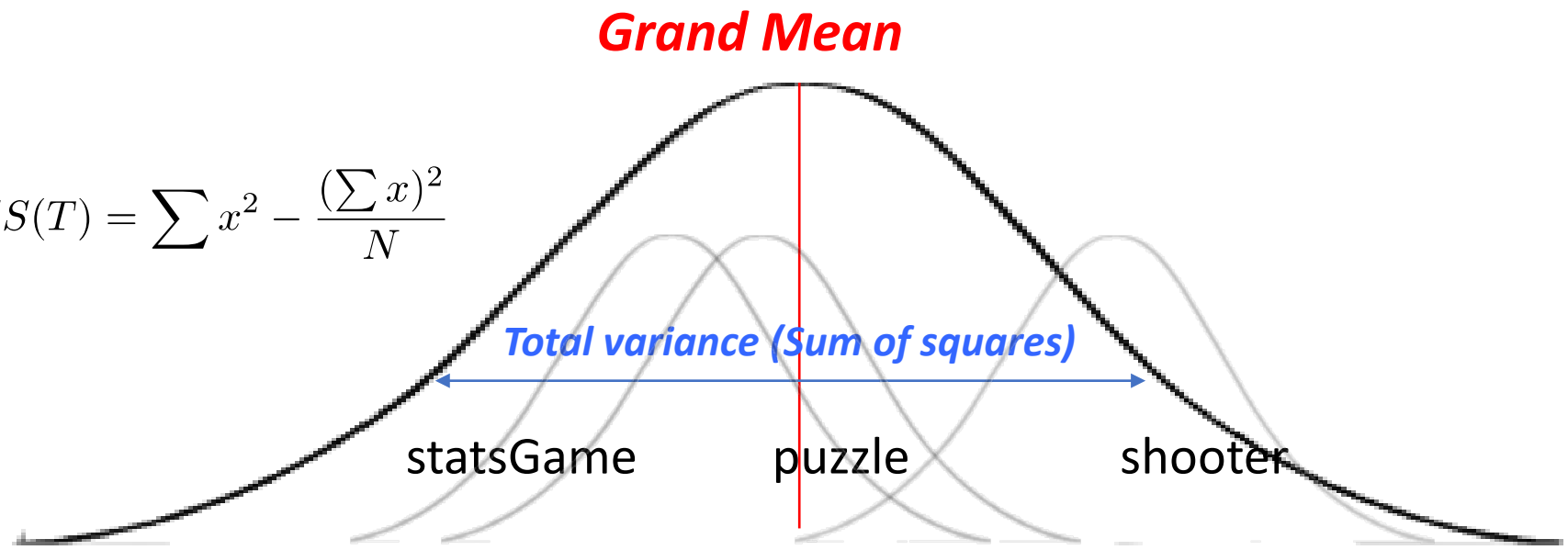
$$SS(T) = \sum x^2 - \frac{(\sum x)^2}{N}$$



Total Sum of Squares (or SS(T)) refers to the sum of squares of difference for all the data.

One-way ANOVA (aka F-test)

$$SS(T) = \sum x^2 - \frac{(\sum x)^2}{N}$$

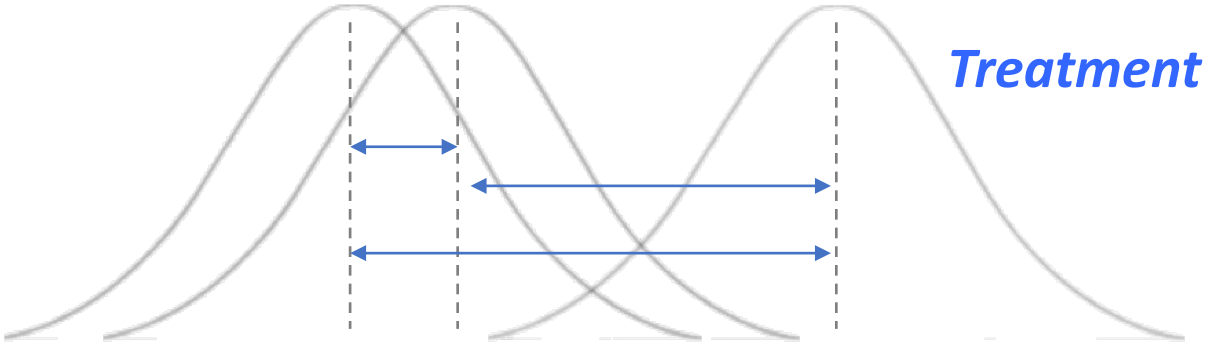


SS(T) (i.e., total variance) has two parts:

a. effect of the independent variable

b. error due to chance

One-way ANOVA
(between-group variance)



SS(T) or Total Variance has two parts:

a. effect of the independent variable

b. error due to chance

One-way ANOVA
(within-group variance)

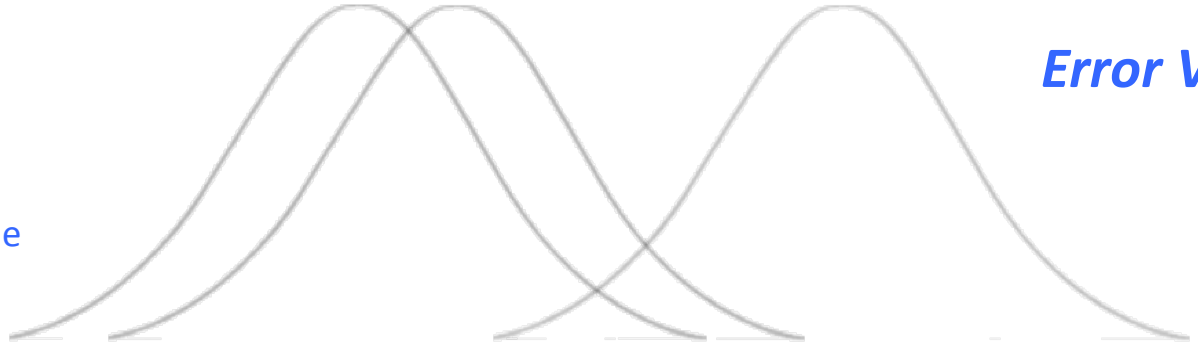
$$F = \frac{betweenGroup\ variance}{withinGroup\ variance}$$

$M_{shooter}$, $SS_{shooter}$

M_{puzzle}, SS_{puzzle}

Error Variance

$M_{statsgame}$, $SS_{statsgame}$



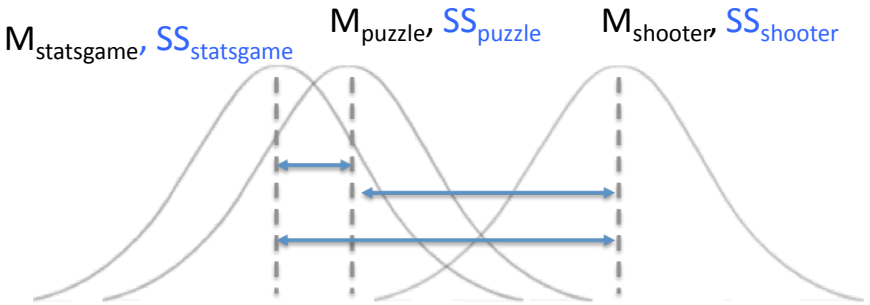
Total Variance has two parts:

a. effect of the independent variable

b. error due to chance

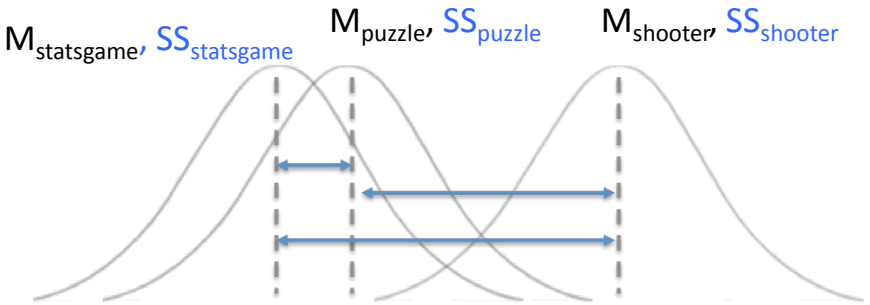
One-way ANOVA
Defining the F-ratio

$$F = \frac{betweenGroup\ variance}{withinGroup\ variance}$$
$$= \frac{treatment\ variance + error\ variance}{error\ variance}$$



One-way ANOVA
Defining the F-ratio

$$F = \frac{betweenGroup\ variance}{withinGroup\ variance}$$
$$= \frac{treatment\ variance + error\ variance}{error\ variance}$$
$$= \frac{\sigma^2_{between}}{\sigma^2_{within}}$$
$$= \frac{s^2_{between}}{s^2_{within}}$$



One-way ANOVA: Calculation of F

$$F = \frac{s^2_{between}}{s^2_{within}} \qquad F = \frac{MS_{between}}{MS_{within}}$$

- F ratio represents the relationship of between-sample and within-sample variance
- Each type of variance can be 'summarized' in terms of MS
- MS is "mean squares".
- This is the average of the sum of squared deviations.

One-way ANOVA: *Summary Table*

Source of Variation	Sum of Squares	df	Mean Square	F	*p*
Between groups	$SS_{between}$	$df_{between}$	$SS_{between}/df_{between}$	$MS_{between}/MS_{within}$	0.05
Within groups	SS_{within}	df_{within}	SS_{within}/df_{within}		
Total	SS_{total}	df_{total}			

- p<0.05 means that we can reject H0
  - *H0:* observed variance is entirely due to error variance
  - *H1:* not H0

One-way ANOVA: *Summary Table*

Source of Variation	Sum of Squares	df	Mean Square	F	*p*
Between groups	$SS_{between}$	$df_{between}$	$SS_{between}/df_{between}$	$MS_{between}/MS_{within}$	0.05
Within groups	SS_{within}	df_{within}	SS_{within}/df_{within}		
Total	SS_{total}	df_{total}			

- p<0.05 means that we can reject H0
  - *H1: at least one treatment mean is different from another treatment mean*
  - *H1: between-σ^2 ≠ within-σ^2*
- Total variance is the sum of Treatment and Error variance
- A large F suggests a significant Treatment variance

One-way ANOVA: *Summary Table*

Source of Variation	Sum of Squares	df	Mean Square	F	*p*
Between groups	$SS_{between}$	$df_{between}$	$SS_{between}/df_{between}$	$MS_{between}/MS_{within}$	0.05
Within groups	SS_{within}	df_{within}	SS_{within}/df_{within}		
Total	SS_{total}	df_{total}			

Structural model (aka General Linear Model)

Each observation is a sum of the mean, the treatment, and individual error

$X_{ij}=\mu + \tau_j + \varepsilon_{ij}$

ANOVAs assume that:

- the variance is the same for all treatment samples
- each sample is normally distributed
- the observations are independent from each other

Structural model
(aka General Linear Model)

Each observation is a sum of the mean, the treatment, and individual error

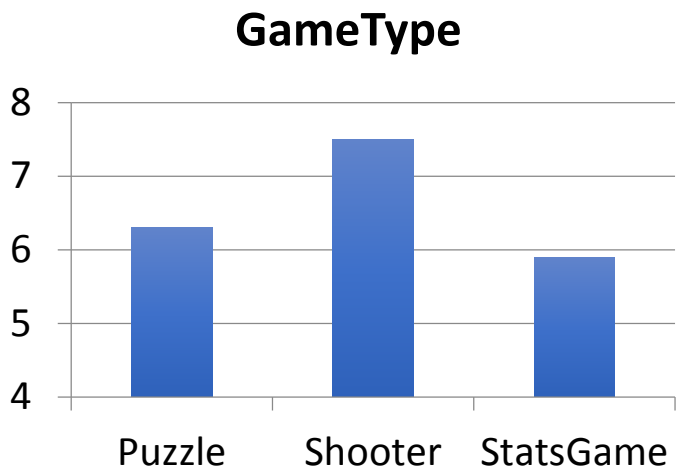
$X_{ij} = \mu + \tau_j + \varepsilon_{ij}$

ANOVAs assume that:

- the sampled populations are normally distributed
- the sampled populations have equal variance
- the observations are independent from each other

One-way ANOVA: *Summary Table*

Source of Variation	Sum of Squares	df	Mean Square	F	*p*
Between groups	13.87	2	6.93	8.36	0.00
Within groups	22.38	27	0.83		
Total	36.25	29			



There is a significant main effect of Game Type ($F(2,27)=8.36$, $p<0.05$).

One-way ANOVA: *Effect size*

Source of Variation	Sum of Squares	df	Mean Square	F	*p*
Between groups	$SS_{between}$	$df_{between}$	$SS_{between}/df_{between}$	$MS_{between}/MS_{within}$	0.05
Within groups	SS_{within}	df_{within}	SS_{within}/df_{within}		
Total	SS_{total}	df_{total}			

Effect size

$$\eta^2 = \frac{SS_{between}}{SS_{total}}$$

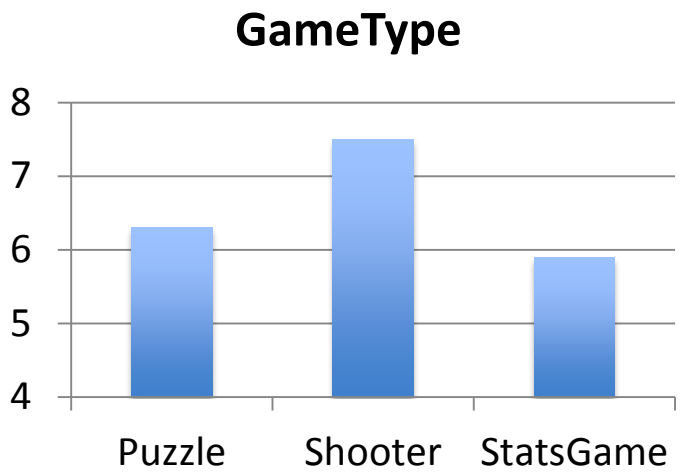
$$f = \sqrt{\frac{\eta^2}{1-\eta^2}}$$

the effect size (η^2) estimates how much of the observed variance is accounted for by the treatment.

the effect size of the F-ratio is estimated by **f**. small (f=0.10); medium (f=0.25); large (f>0.40)

One-way ANOVA: *Summary Table*

Source of Variation	Sum of Squares	df	Mean Square	F	*p*
Between groups	13.87	2	6.93	8.36	0.00
Within groups	22.38	27	0.83		
Total	36.25	29			



There is a significant main effect of Game Type ($F(2,27)=8.36$, $p<0.05$, $\eta^2=0.38$). The effect size of this result was large ($f=0.79$).

One-way ANOVA & t-tests

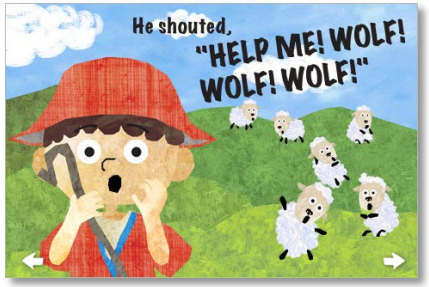
- F-test and t-test are related; $F=t^2$
  (see Excel example for 2-level tests)

- Like a t-value, an F ratio is likely to be significant if:
  - the sample size (i.e., n_j) is relatively large
  - the variability within each condition is relatively small
  - the difference(s) between conditions are relatively large

- Like a t-test, an F test assumes that:
  - the samples come from populations that have equal variances
  - the dependent measure is normally distributed
  - the data were randomly sampled
  - the dependent measure is an interval or ratio scale, which allows for mean calculation

Why not run multiple t-tests?

- α=0.05 means that we risk 1 false discovery per 20 tries
  (Type 1 error).
- Number of comparisons = k(k-1)/2,
  Prob(experiment-wise error) = $1-(1-\alpha)^{\text{comparisons}}$
  - 2 levels: 1 comparison, Prob(Type 1 error)=0.05
  - 3 levels: 3 comparisons, Prob(Type 1 error)=0.14
  - 5 levels: 10 comparisons Prob(Type 1 error)=0.40
  - 10 levels: 45 comparisons Prob(Type 1 error)=0.90



- Be cautious of any study that performs multiple t-tests on the same data-set.

Why ANOVA?

- An ANOVA is an omnibus statistical test.
  - This means it can deal with multiple variables.
  - It analyses patterns of variance, not sample differences.
- Experiment-wise error: It protects against Type 1 error
- It allows for comparisons across many different conditions and accelerate scientific discovery

Post-hoc Tests

Bonferroni correction

- $\alpha_{desired}$ = for wrongly rejecting H0 (Type 1 error)
- m = number of hypotheses
- $\alpha_{adjusted} = \frac{\alpha_{desired}}{m}$

Post-hoc Tests

- Tukey HSD test
- *honest significant difference*
- $q_{crit}(\alpha, k, df_{within})$

$$q = \frac{\bar{x}_{max} - \bar{x}_{min}}{\sqrt{\frac{MS_{within}}{n}}}$$

if $q > q_{crit}$, reject H0

if $\bar{x}_{max} - \bar{x}_{min} > q_{crit}\sqrt{\frac{MS_{within}}{n}}$, reject H0

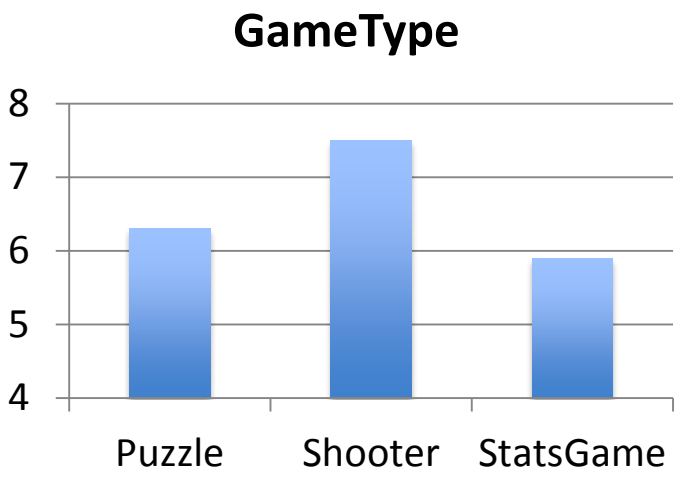
The studentized range statistic (q)*

*The critical values for q corresponding to alpha = .05 (top) and alpha = .01 (bottom)

df for Error Term	k= Number of Treatments								
	2	3	4	5	6	7	3	9	10
5	3.64 5.70	4.60 6.98	5.22 7.80	5.67 8.42	6.03 8.91	6.33 9.32	6.58 9.67	6.80 9.97	6.99 10.2 4
6	3.46 5.24	4.34 6.33	4.90 7.03	5.30 7.56	5.63 7.97	5.90 8.32	6.12 8.61	6.32 8.87	6.49 9.10
7	3.34 4.95	4.16 5.92	4.68 6.54	5.06 7.01	5.36 7.37	5.61 7.68	5.82 7.94	6.00 8.17	6.16 8.37
8	3.26 4.75	4.04 5.64	4.53 6.20	4.89 6.62	5.17 6.96	5.40 7.24	5.60 7.47	5.77 7.68	5.92 7.86
9	3.20 4.60	3.95 5.43	4.41 5.96	4.76 6.35	5.02 6.66	5.24 6.91	5.43 7.13	5.59 7.33	5.74 7.49
10	3.15 4.48	3.88 5.27	4.33 5.77	4.65 6.14	4.91 6.43	5.12 6.67	5.30 6.87	5.46 7.05	5.60 7.21

One-way ANOVA: *Summary Table*

Source of Variation	Sum of Squares	df	Mean Square	F	*p*
Between groups	13.87	2	6.93	8.36	0.00
Within groups	22.38	27	0.83		
Total	36.25	29			



There is a significant main effect of Game Type ($F(2,27)=8.36$, $p<0.05$, $\eta^2=0.38$). The effect size of this result was large ($f=0.79$). A post-hoc Tukey test showed that the anxiety levels induced by the Shooter game differed significantly from the other games at $p < 0.05$; the "Puzzle" and "Stats" game did not significantly differ from each other.

Summary

- NHST
  - research question, scientific hypothesis
  - null-hypothesis, alternative-hypothesis
- two sample test: between groups t-test
  - effect size (Cohen's d)
- One way analysis of variance for multiple levels
  - SS(T), SS(B), SS(W)
  - F-ratio
  - Summary table
  - effect size
  - post hoc tests for multiple comparisons