Statistics for Health Care

Unit 5:

Overview/Teasers

Overview

- Statistical inference: <mark>confidence intervals,</mark> hypothesis tests, and p-values

p-
(   )
http://blog.naver.com/chcher/70183449033

Teaser 1, Unit 5, polls

- New York Times/CBS News Poll (Feb. 2012) found: "59 percent said the health insurance plans of religiously affiliated employers should cover the cost of birth control."

- "The nationwide telephone poll included 1,064 registered U.S. voters. The margin of sampling error is plus or minus three percentage points."

Where does the margin of error come from?

Teaser 2, Unit 5

- Thai HIV vaccine trial (2009)
    - 8197 randomized to vaccine (RV144)
    - 8198 randomized to placebo

- 51 infections in the vaccine group vs. 74 in the placebo group ("statistically significant difference")
- Generated a lot of public discussion about p-values!
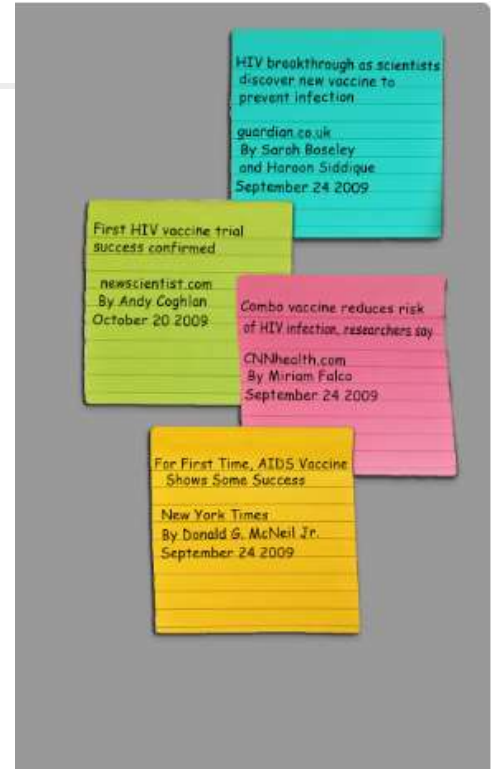
Statistics in Medicine

Module 1:

Review of Z-distribution/Introduction to The T-distribution

Review: Z distribution

Question:
Bone density of women:
mean = 1.0, Std dev = 0.1,
Prob. < 0.9?

z	0.00	0.01	0.02	0.03	0.04
0.0	0.5000	0.5040	0.5080	0.5120	0.5159
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0.6554	0.6591	0.6628	0.6664	0.6700
0.5	0.6915	0.6950	0.6985	0.7019	0.7054
0.6	0.7257	0.7291	0.7324	0.7357	0.7389
0.7	0.7580	0.7611	0.7642	0.7673	0.7704
0.8	0.7881	0.7910	0.7939	0.7967	0.7995
0.9	0.8159	0.8186	0.8212	0.8238	0.8264
1.0	0.8413	0.8438	0.8461	0.8485	0.8508
1.1	0.8643	0.8665	0.8686	0.8708	0.8729
1.2	0.8849	0.8869	0.8888	0.8907	0.8925
1.3	0.9032	0.9049	0.9066	0.9082	0.9099

ue Calculation

Z score (z) = -1.0

Calculate   Reset

Left-tailed P-value ($Z < 1$)   = 0.1587

Right-tailed P-value ($Z > 1$)   = 0.8413

http://easycalculation.com/statistics/p-value-for-z-score.php

What is a T-distribution?

- A t-distribution is like a standard normal (Z) distribution, except has slightly fatter tails for small samples (n<100).

- The bigger the sample size, then the closer t becomes to Z.

- If n≥100, t approaches Z.

The T probability density function

What does t look like mathematically?

$$f(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \frac{1}{\sqrt{v\pi}} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}}$$
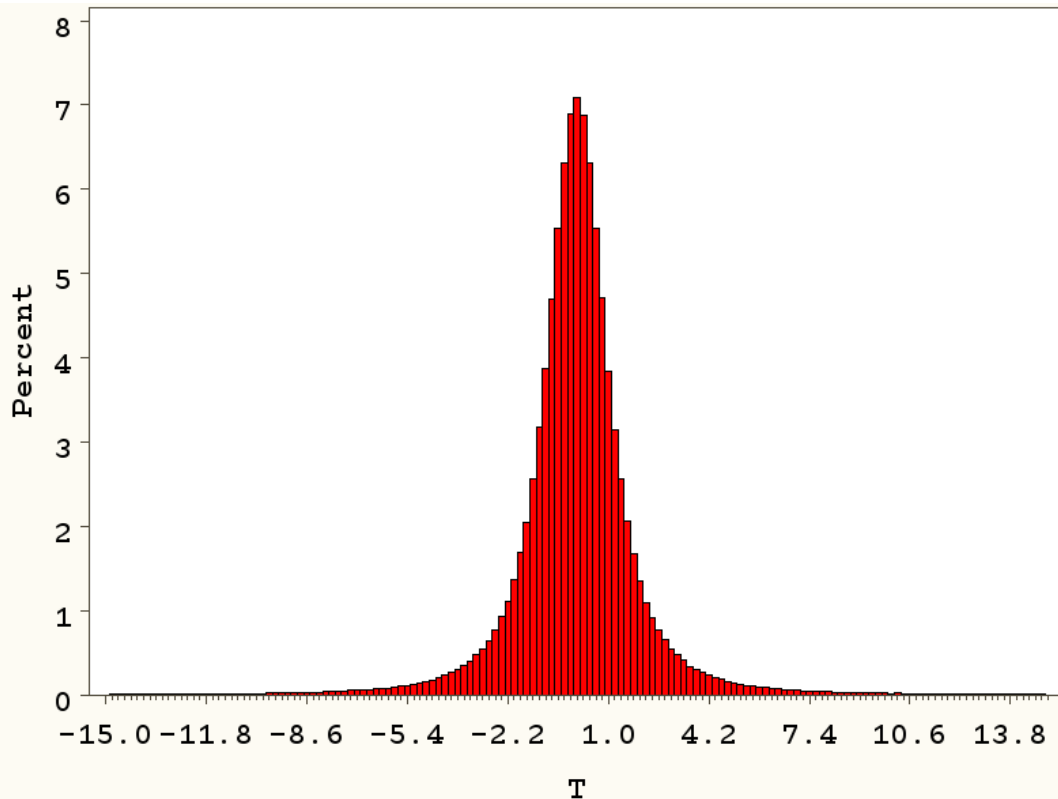
Where:

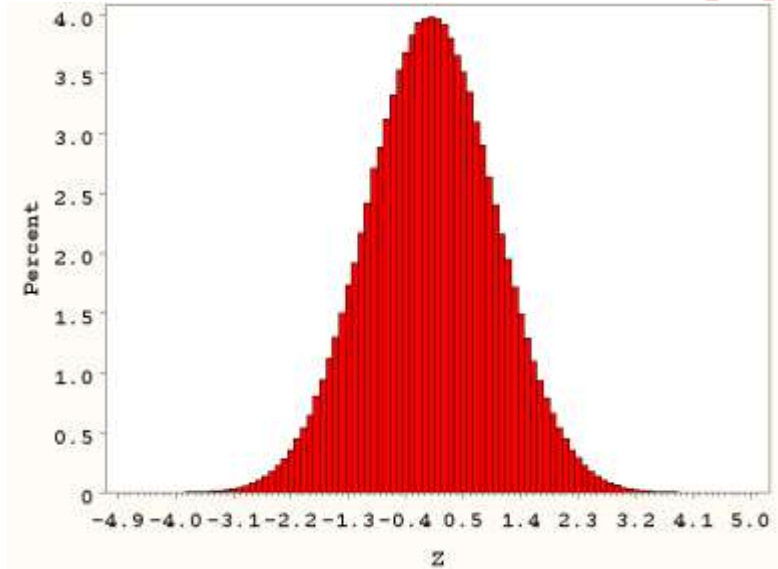v is the degrees of freedom (which is directly related to sample size)

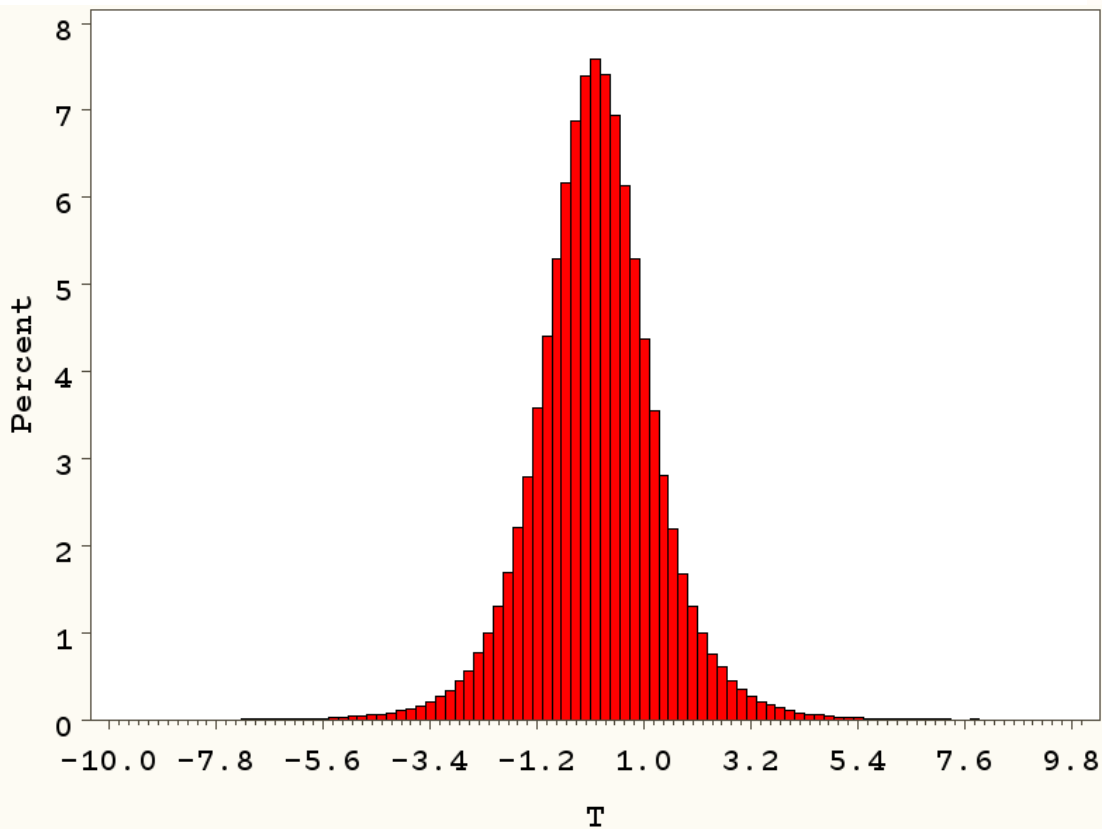Γ (gamma) is the Gamma function $\Gamma(n) = (n-1)!$.
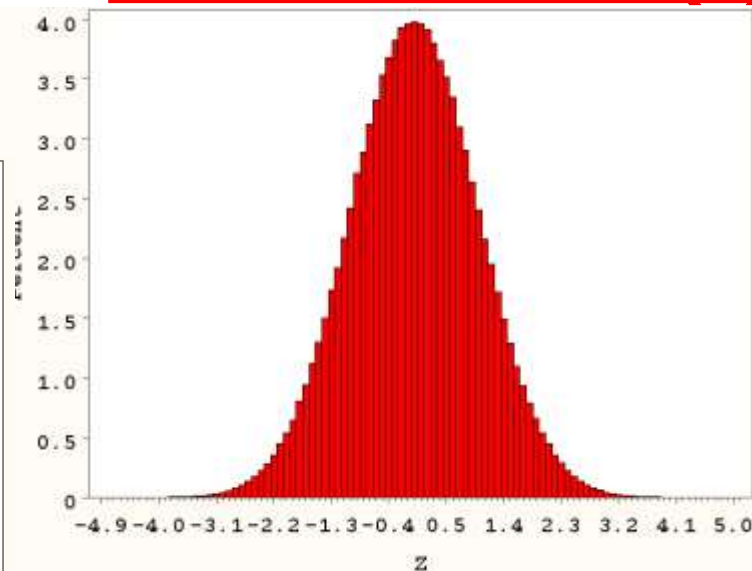
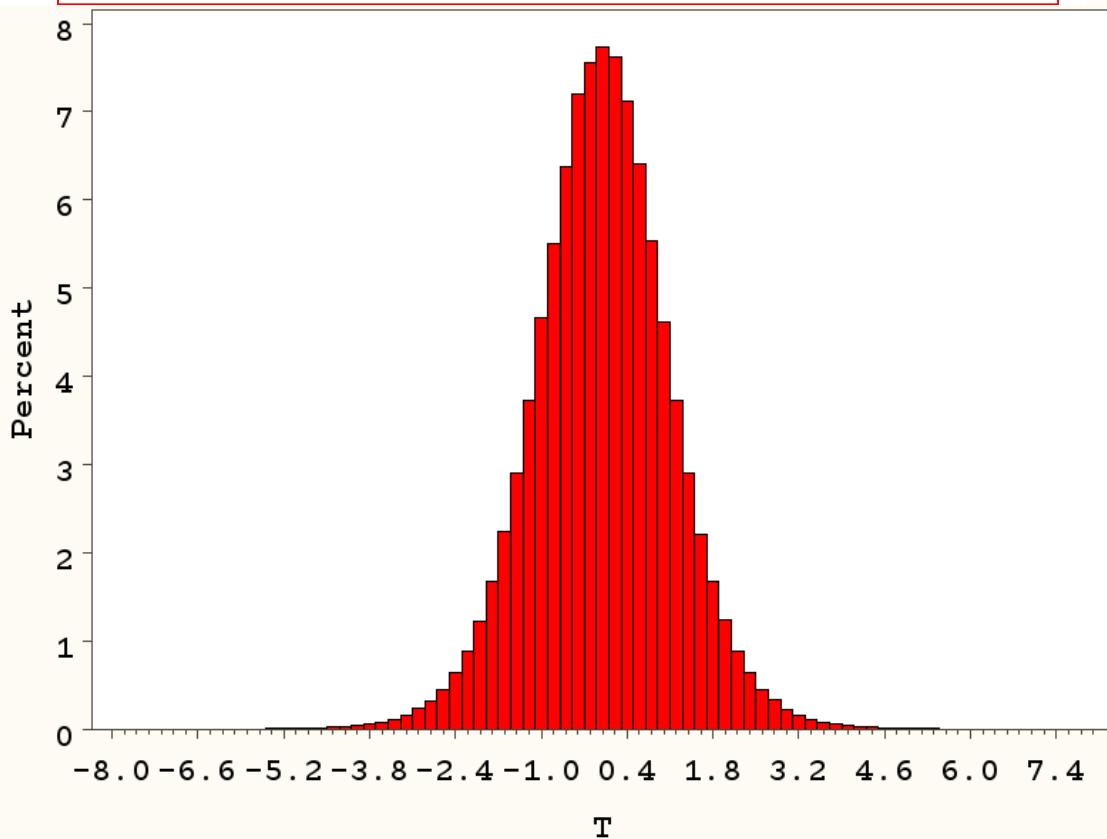T-distribution with 2 degrees of freedom (n=3).

Standard Normal (Z)
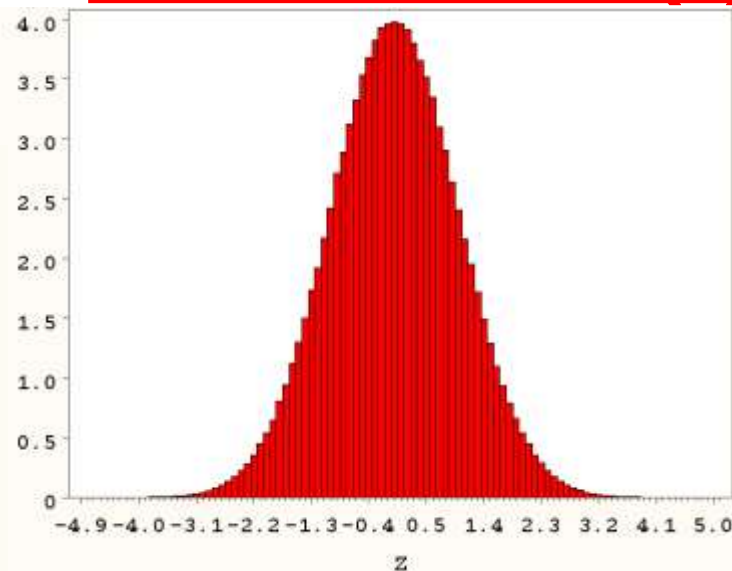
T-distribution with 5 degrees of freedom (n=6).

Standard Normal (Z)

T-distribution with 10 degrees of freedom (n=11).

Standard Normal (Z)

T-distribution with 20 degrees of freedom (n=21).

Standard Normal (Z)

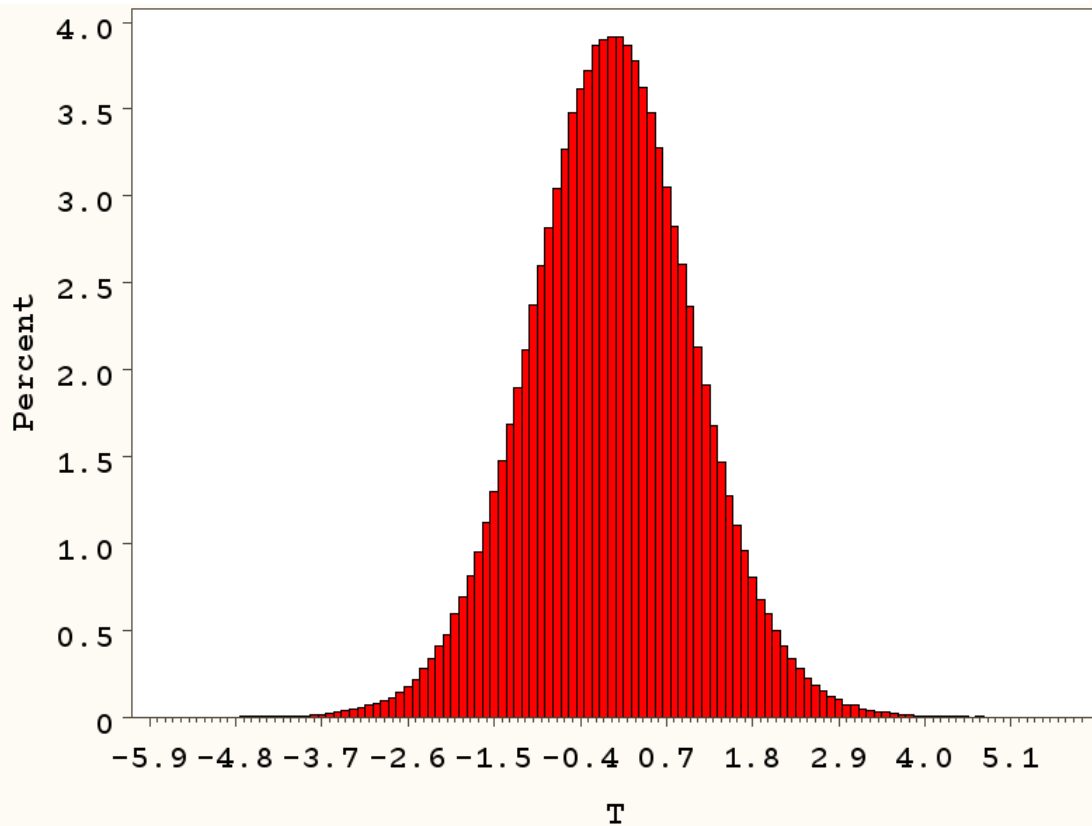**T-distribution with 30 degrees of freedom (n=31).**

Standard Normal (Z)
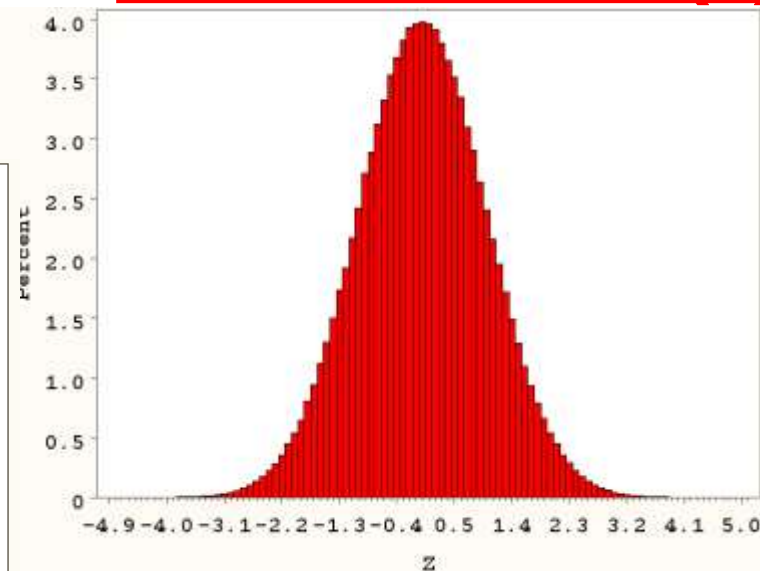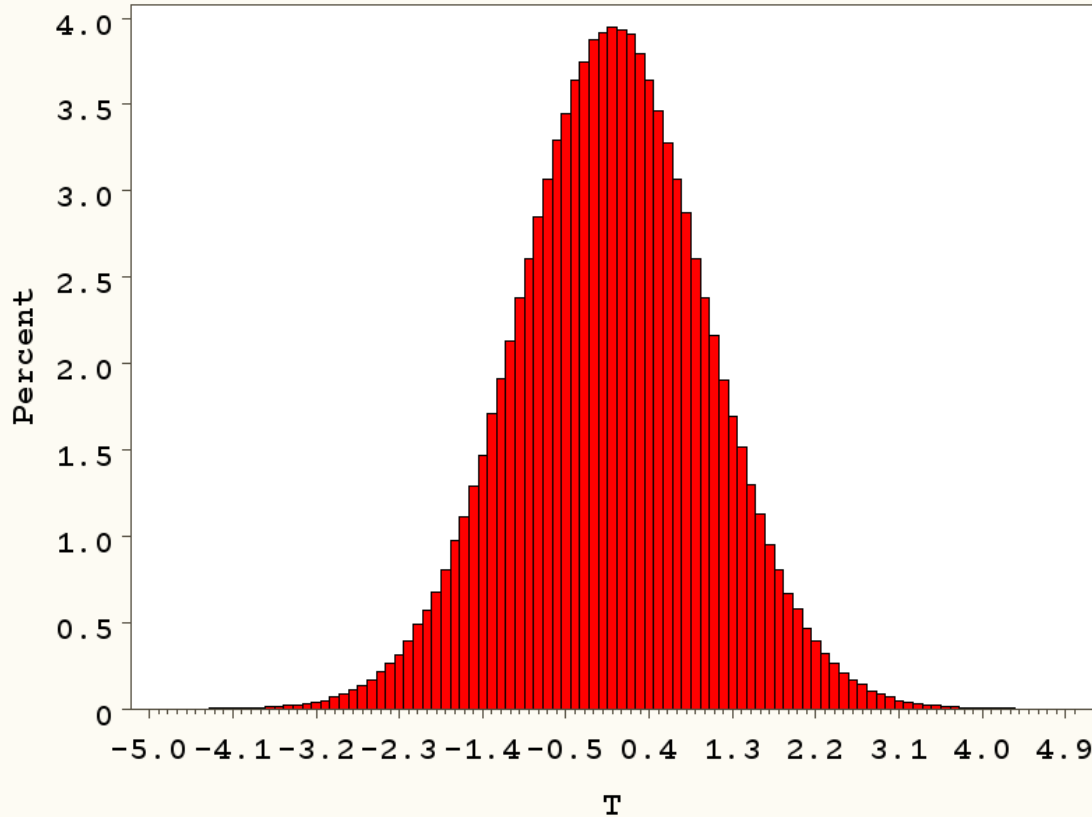
T-distribution with 100 degrees of freedom (n=101).
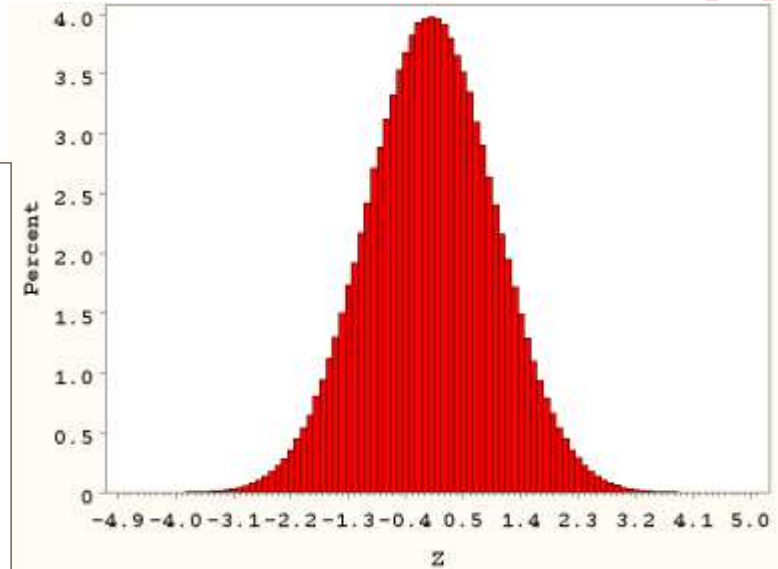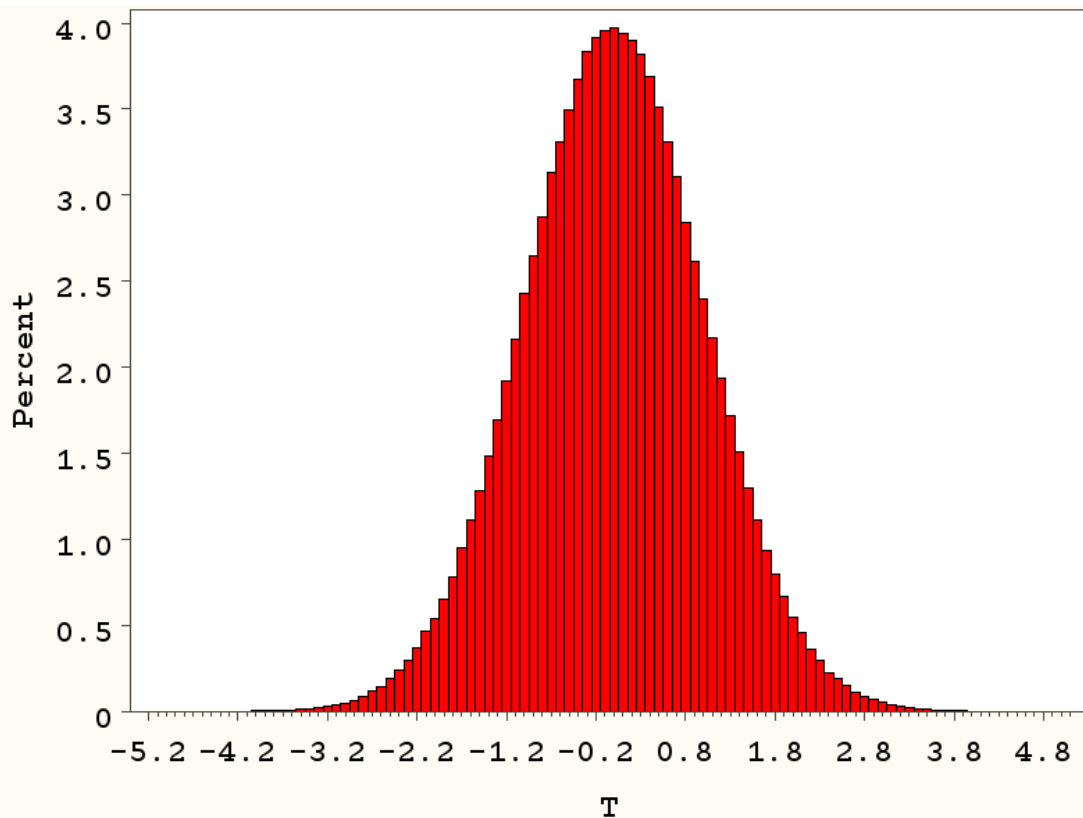
Standard Normal (Z)
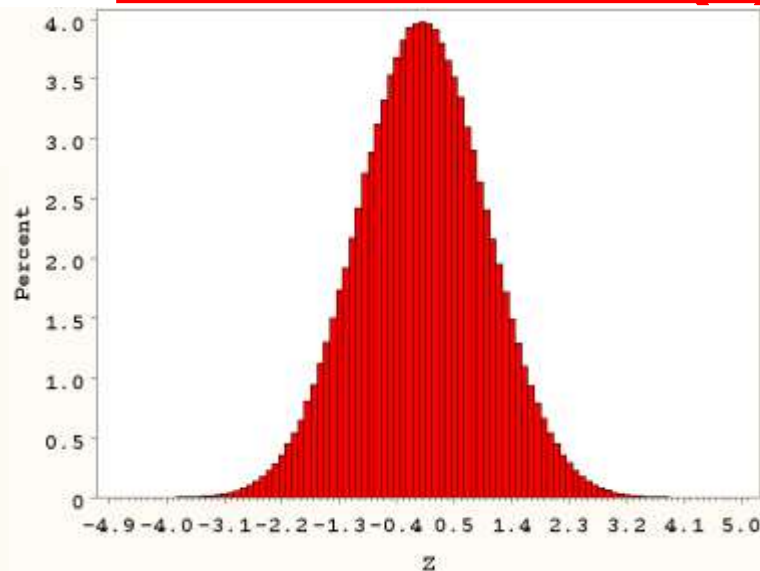
T-distribution with 500 degrees of freedom (n=501).
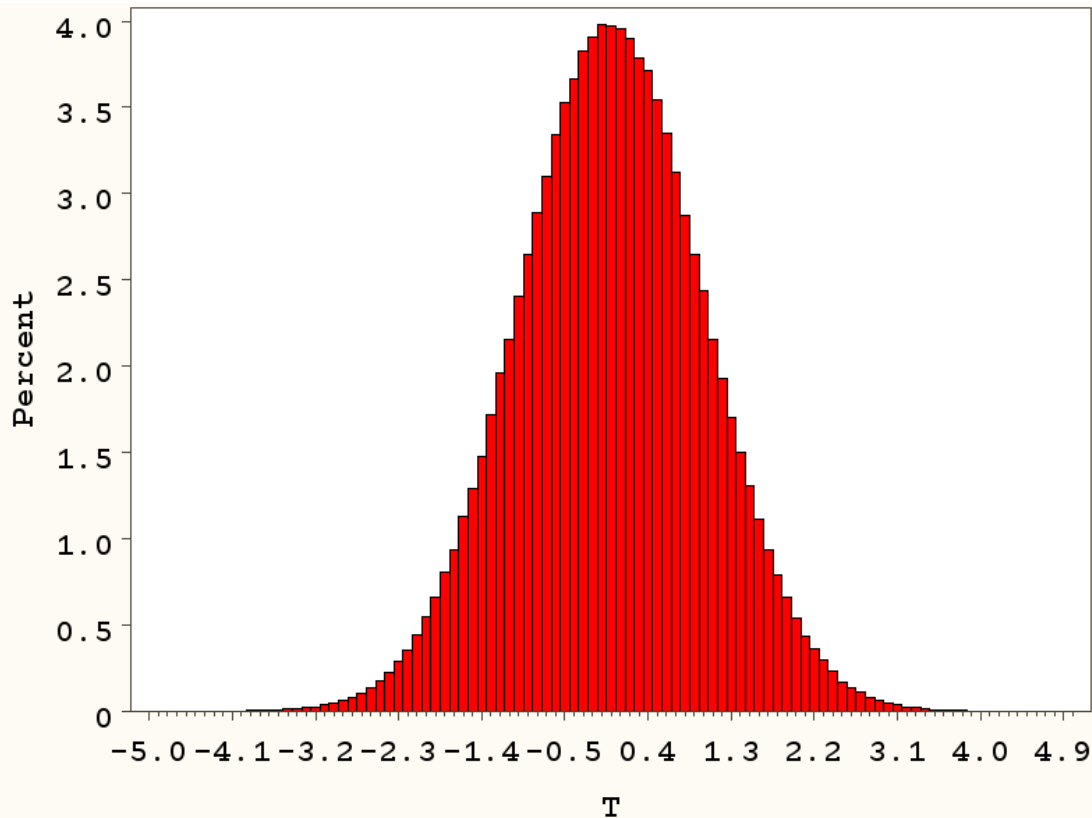
Standard Normal (Z)
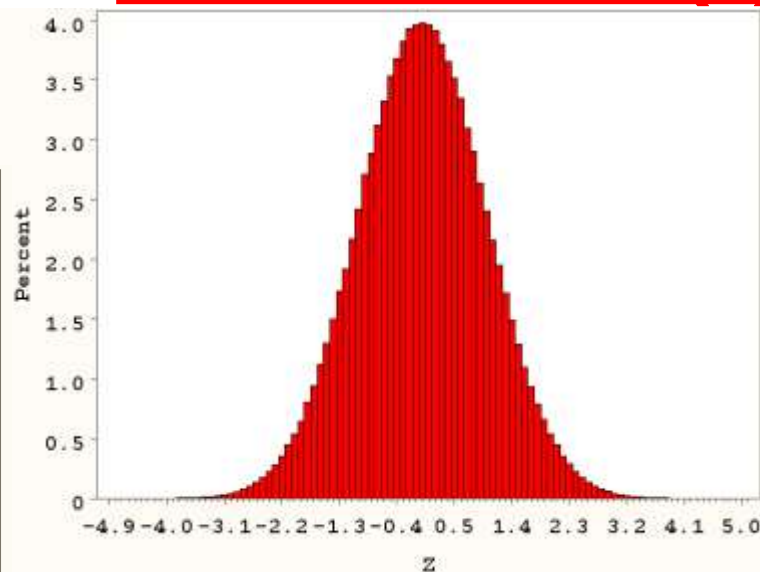
Statistics in Medicine

Module 2:
Introduction to statistical inference

Statistical inference: making about the population from a sample.

$$\overline{X} = \frac{\sum\limits_{i=1}^{n} x}{n}$$

$$s^2 = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{X}_n)^2}{n-1}$$

Truth (not observable)

Sample (observation)

Population parameters

$$\mu = \frac{\sum\limits_{i=1}^{N} x}{N} \qquad \sigma^2 = \frac{\sum\limits_{i=1}^{N}(x_i - \mu)^2}{N}$$

Inference: Make guesses about the whole population
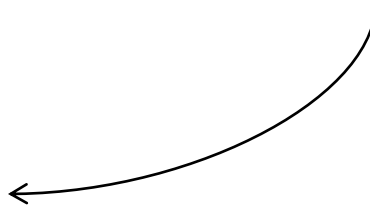
Statistics vs. Parameters

- **<u>Sample Statistic</u>** – any summary measure calculated from data; e.g., could be a mean, a difference in means or proportions, an odds ratio, or a correlation coefficient
  - E.g., the mean vitamin D level in a sample of 100 men is 63 nmol/L

- **<u>Population parameter</u>** – the true value/true effect in the entire population of interest
  - E.g., the true mean vitamin D in all middle-aged and older European men is 62 nmol/L

Examples of Sample Statistics:

Mean

Rate

Risk (proportion)

Difference in means (ttest)

Difference in risks (Z-test)

Relative risk (odds ratio/risk ratio/rate ratio)

Correlation coefficient

Regression coefficient

…

Statistical inference

- Two possible goals:
  - 1. Estimation (confidence intervals)
    - Give a plausible range of values for the true effect size.
  - 2. Hypothesis testing (p-values)
    - Is there an effect at all?

Statistics in Medicine

Module 3:

Introduction to the distribution of a statistic

Statistics follow distributions!

- These distributions are defined by:
  - 1. Shape (e.g., normal distribution, T-distribution)
  - 2. Mean
  - 3. Standard deviation (called a standard error!)

How can a statistic have a distribution??

- A statistic is a single number calculated from our sample data (e.g., a mean or risk ratio).

- How can a single number have a distribution?
  - Answer: It's a theoretical concept!

Distribution of a statistic…

- *The distribution of a statistic is a theoretical construct.*

- Statisticians ask a thought experiment: how much would the value of the statistic fluctuate if one could repeat a particular study over and over again with different samples of the same size?

- By answering this question, statisticians are able to pinpoint exactly how much uncertainty is associated with a given statistic.

Distribution of a statistic

- Two approaches to determine the distribution of a statistic:
    - 1. Computer simulation
        - Repeat the experiment over and over again virtually!
        - More intuitive; can directly observe the behavior of statistics.
    - 2. Mathematical theory
        - Proofs and formulas!
        - More practical; use general formulas to solve problems.

Example of computer simulation: virtual coin flipping

- Computer simulation:
  - Flip a virtual coin 100 times; count the number of heads.
    - Statistic = number of heads
  - Repeat this over and over again a large number of times (we'll try 30,000 repeats!)
  - Plot the 30,000 results.

30,000 repeated experiments!
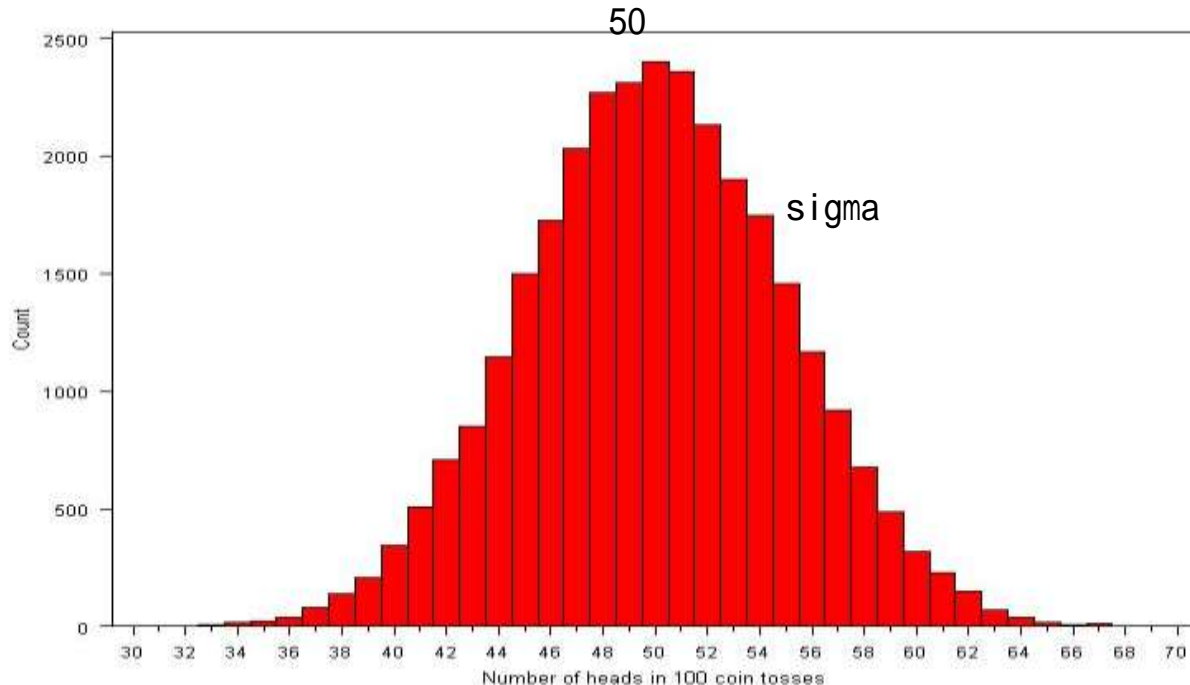
Shape: normally distributed

Mean = 50

Std. dev = 5

∴ 95% of the time, we get between 40 and 60 heads…

Same example, using mathematical theory:

X~Bin(n=100,p= .5)

- 1. Shape: binomial can be approximated by a normal for np>5
- 2. Mean = np = 100*.5 = 50
- 3. Variance = np(1-p) = 25
  - Therefore, SD = 5

Statistics in Medicine

Module 4:

Distributions of some common statistics

Example dataset: cognitive function and vitamin D

- Hypothetical data loosely based on [1]; cross-sectional study of 100 middle-aged and older European men.
  - Sample statistic: mean vitamin D levels (true mean=62 nmol/L)
  - Sample statistic: correlation coefficient between vitamin D and cognitive function, measured by the Digit Symbol Substitution Test (DSST). (true correlation coefficient=0.15)

1. Lee DM, Tajar A, Ulubaev A, et al. Association between 25-hydroxyvitamin D levels and cognitive performance in middle-aged and older European men. J Neurol Neurosurg Psychiatry. 2009 Jul;80(7):722-9.

Sample data: vitamin D (n=100)

Right-skewed!

Mean= 63 nmol/L

Standard deviation = 33 nmol/L
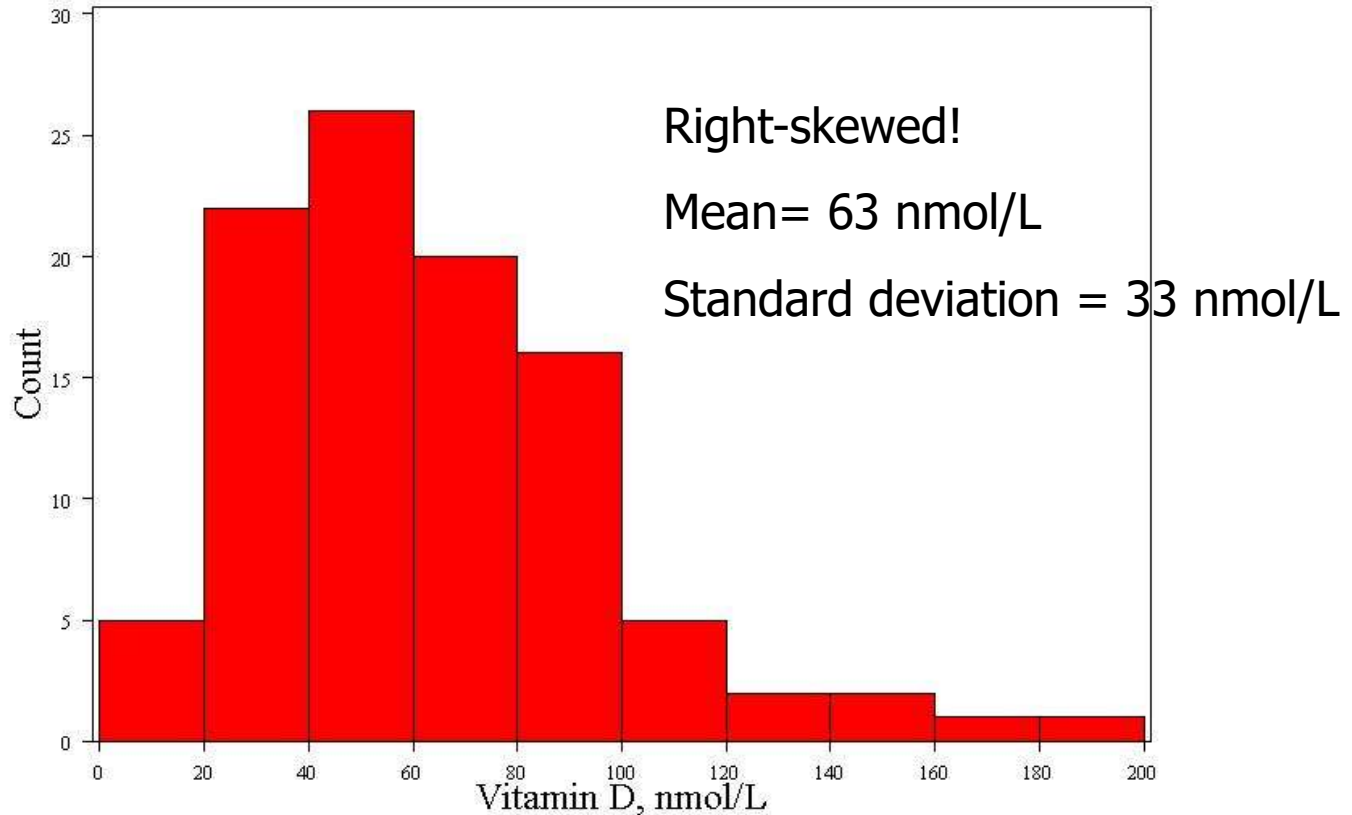
Correlation between vitamin D and cognitive function (DSST score), n=100
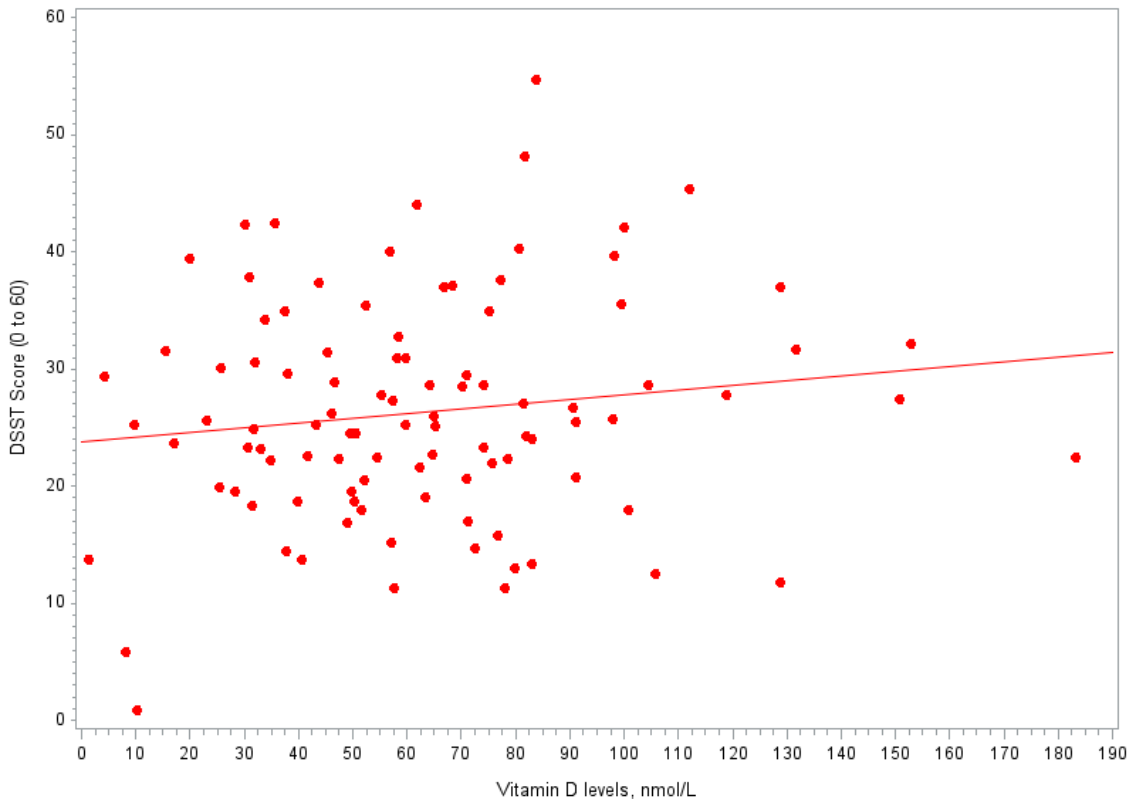
Correlation coefficient r =0.15

Correlation coefficient (r) ranges from -1 to +1.

0=uncorrelated

+1 = perfect positive correlation

-1 = perfect inverse correlation

Correlation coefficient e.g.
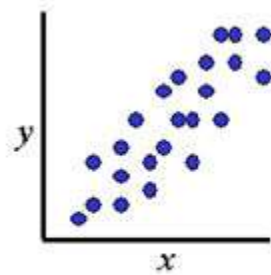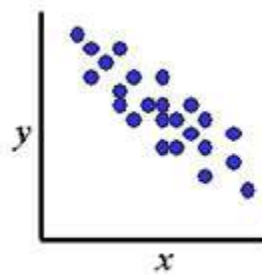
Positive correlation   Negative correlation   No correlation

$\rho = -1$ $-1 < \rho < 0$

$0 < \rho < +1$ $\rho = +1$ $\rho = 0$
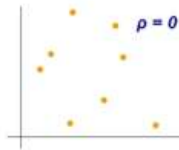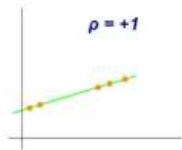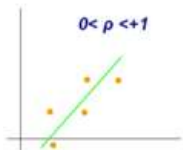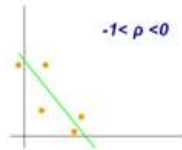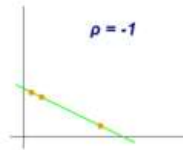
Correlation Coefficient
Shows Strength & Direction of Correlation

Strong ← Weak | Weak → Strong

-1.0   -0.5   0.0   +0.5   +1.0

Negative Correlation   Zero   Positive Correlation

What's the distribution of the **mean of** vitamin D?

Computer simulation of the distribution of a statistic (sample mean)

- 1. Specify the underlying distribution of vitamin D in all European men aged 40 to 79.
  - Right-skewed
  - Standard deviation = 33 nmol/L
  - True mean = 62 nmol/L (this is arbitrary; does not affect the distribution)
- 2. Select a random sample of 100 virtual men from the population.
- 3. Calculate the mean vitamin D for the sample.
- 4. Repeat steps (2) and (3) a large number of times (say 10,000 times).
- 5. Explore the distribution of the 10,000 means.

Distribution of mean vitamin D (a sample statistic)
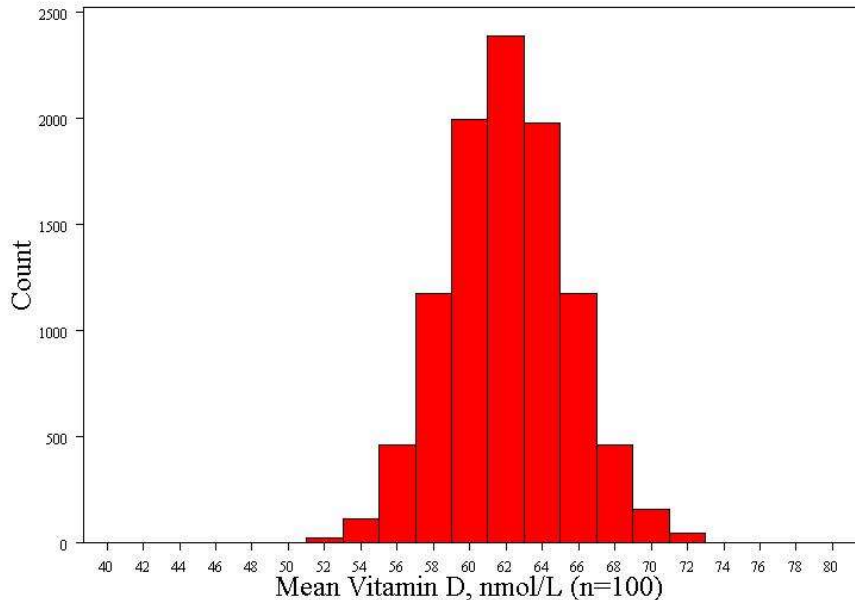
Normally distributed! Surprise!

Mean (of means)= 62 nmol/L (the true mean)

Standard deviation (error) = 3.3 nmol/L

(Note: Standard deviation = 33 nmol/L)

Distribution of mean vitamin D (a sample statistic)

- Normally distributed (even though the trait is right-skewed!)

- Mean = true (population) mean

- Standard deviation = 3.3 nmol/L

  - The standard deviation of a statistic is called a standard error

Distribution of a sample mean in general (from math theory):

- T-distribution
  - Same as a normal distribution for larger n (≥100)
  - Fatter tails for small n (n<100)
- Mean = true mean
- The standard error = $\dfrac{s}{\sqrt{n}}$

Standard error of the mean:

- Standard error of the mean = $\dfrac{s}{\sqrt{n}}$

  - Standard error decreases with bigger sample size (n).
  - Standard error increases with greater trait variability (s).

If I increase the sample size to n=400...

Standard error = 1.7 nmol/L

$$\frac{s}{\sqrt{n}} = \frac{33}{\sqrt{400}} = 1.7$$

Mean Vitamin D, nmol/L (n=400)
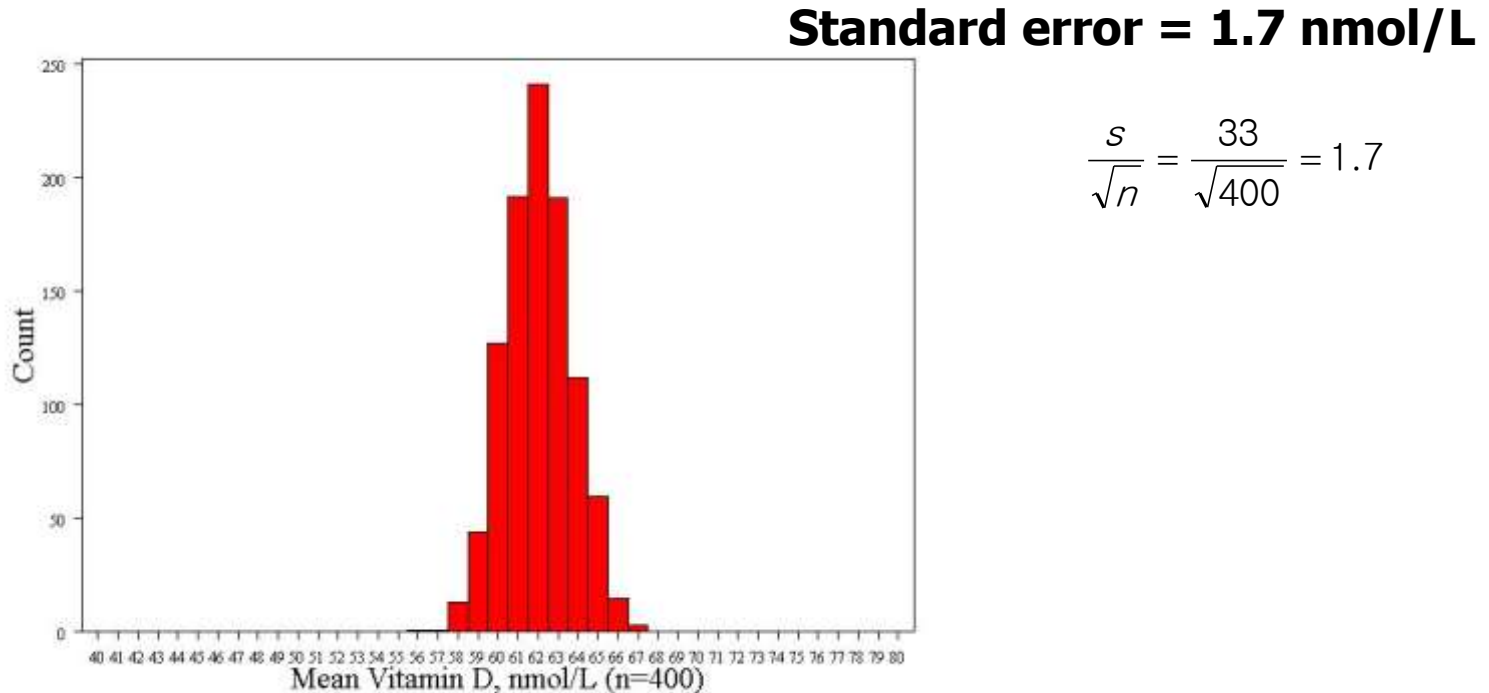
If I increase the variability of vitamin D (the trait) to SD=40...

Standard error = 4.0 nmol/L
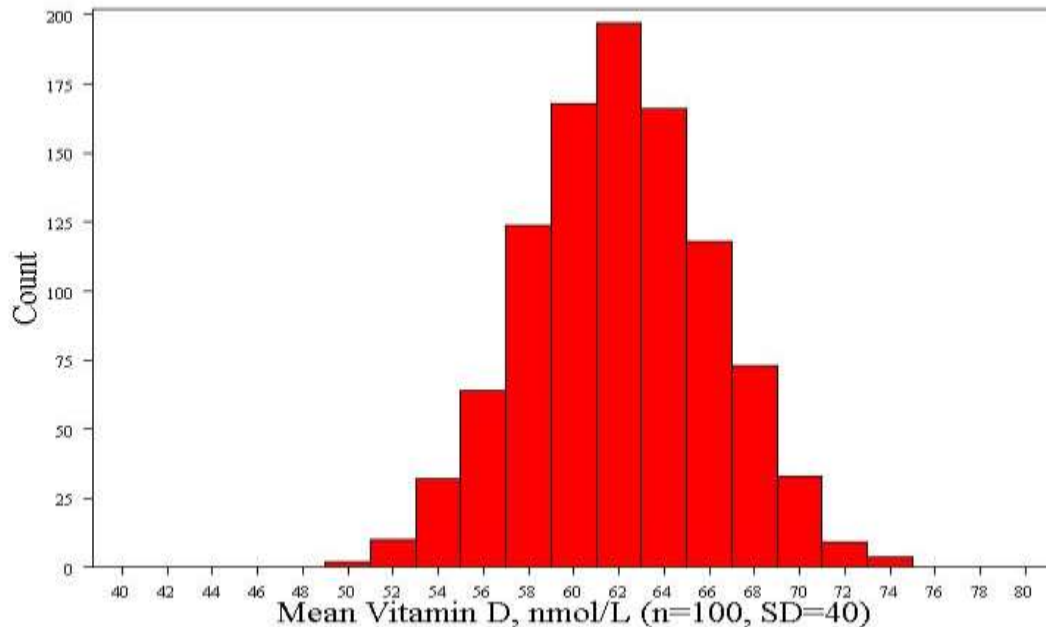
$$\frac{s}{\sqrt{n}} = \frac{40}{\sqrt{100}} = 4.0$$

Mathematical Theory… The Central Limit Theorem!

If all possible random samples, each of size n, are taken from any population with a mean μ and a standard deviation σ, the sampling distribution of the sample means (averages) will:

1. have mean: $$\mu_{\bar{x}} = \mu$$

2. have standard deviation (="standard error"): $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

3. be approximately normally distributed regardless of the shape of the parent population (normality improves with larger n). It all comes back to Z!

Symbol Check

$\mu_{\overline{x}}$ **The mean of the sample means.**

$\sigma_{\overline{x}}$ **The standard deviation of the sample means. *Also called "the standard error of the mean."***

Mathematical Proof

If X is a random variable from any distribution with known mean, $E(x)$, and variance, $Var(x)$, then the expected value and variance of the average of n observations of X is:

$$E(\overline{X}_n) = E\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{\sum_{i=1}^{n} E(x)}{n} = \frac{nE(x)}{n} = E(x)$$

$$Var(\overline{X}_n) = Var\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{\sum_{i=1}^{n} Var(x)}{n^2} = \frac{nVar(x)}{n^2} = \frac{Var(x)}{n}$$

Computer simulation of the CLT:

1. Pick any probability distribution and specify a mean and standard deviation.
2. Tell the computer to randomly generate 1000 observations from that probability distributions

   E.g., the computer is more likely to spit out values with high probabilities
3. Plot the "observed" values in a histogram.
4. Next, tell the computer to randomly generate 1000 averages-of-2 (randomly pick 2 and take their average) from that probability distribution. Plot "observed" averages in histograms.
5. Repeat for averages-of-10, and averages-of-100.

Uniform on [0,1]: 1000 observations
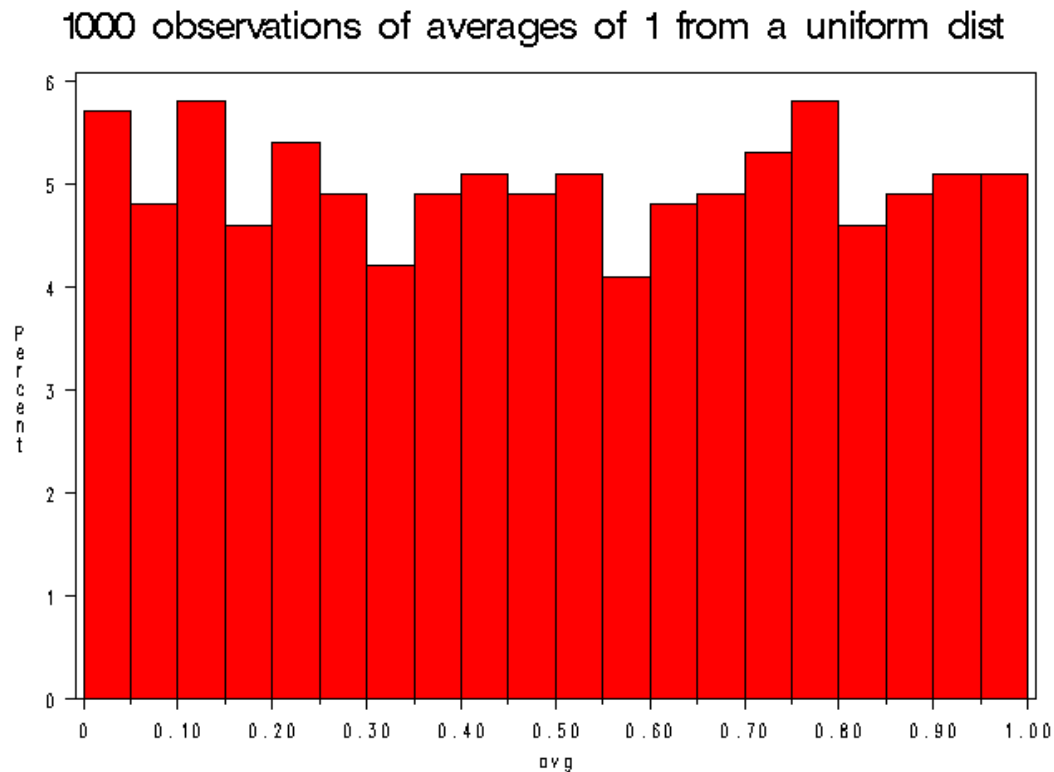
Uniform: 1000 averages of 2 values

1000 observations of averages of 2 from a uniform distribution
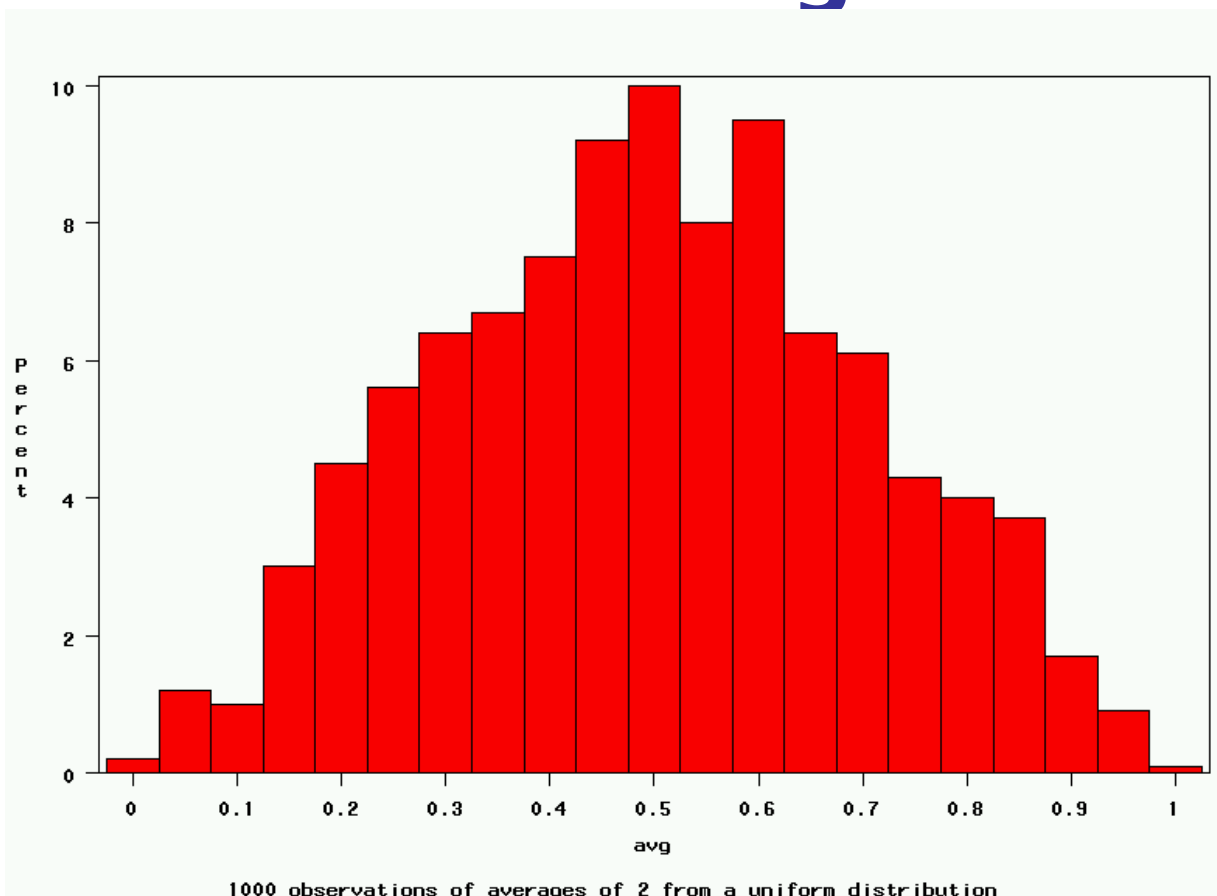
Uniform: 1000 averages of 5 values

1000 observations of averages of 5 from a uniform distribution

Uniform: 1000 averages of 100

1000 observations of averages of 100 from a uniform distribution
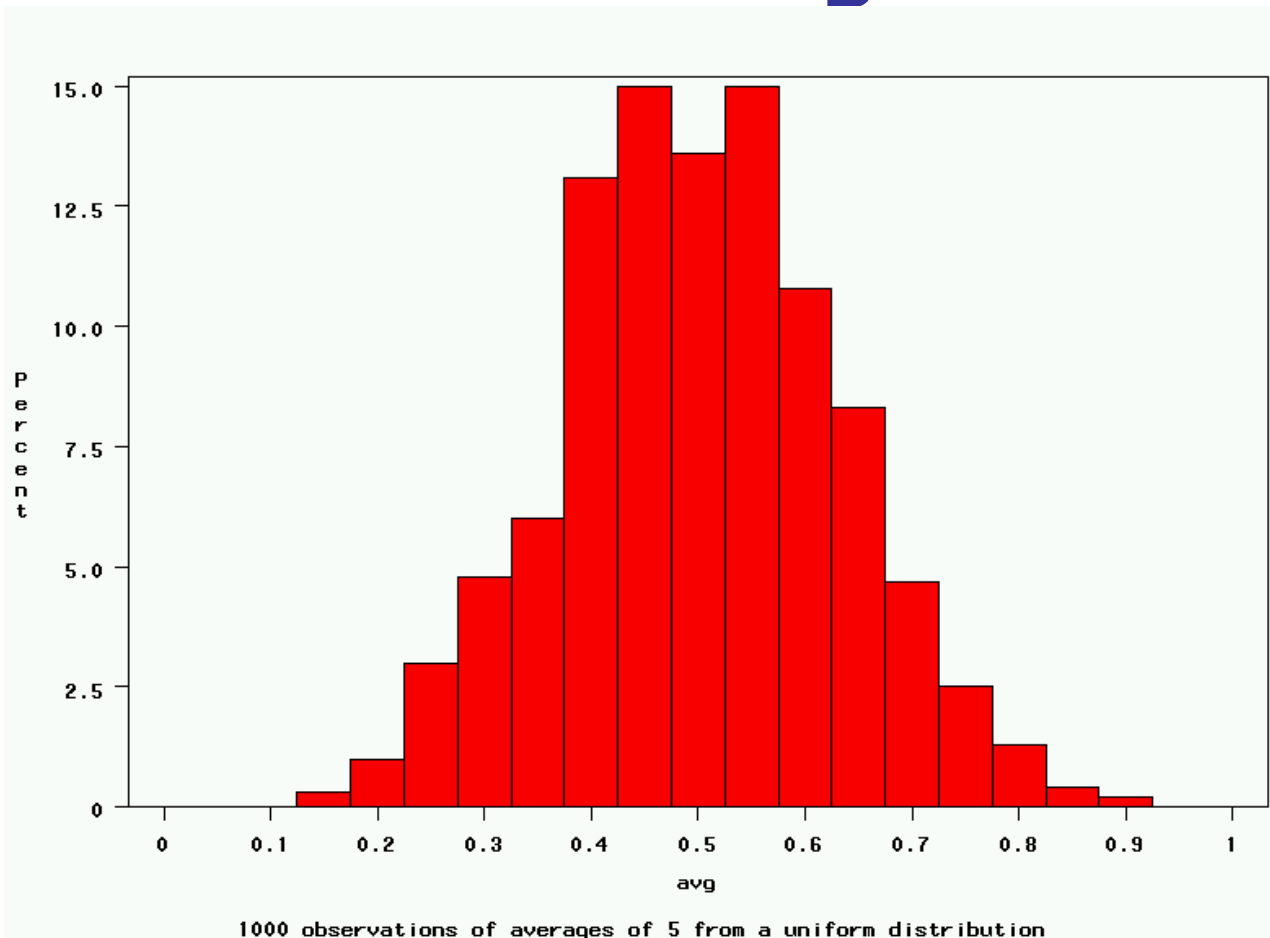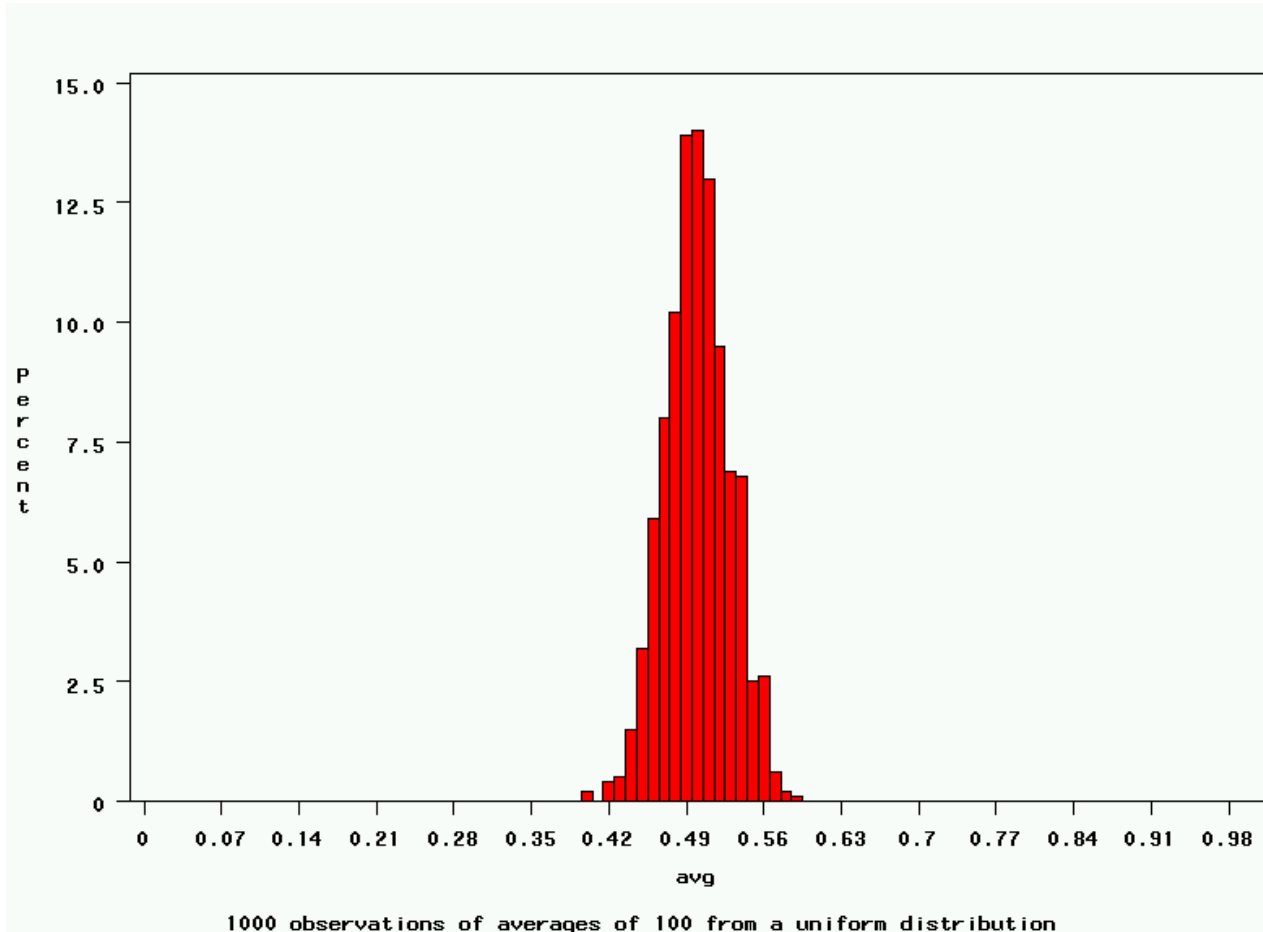
Exponential distribution: 1000 observations

Average of 1 from an EXPONENTIAL distribution
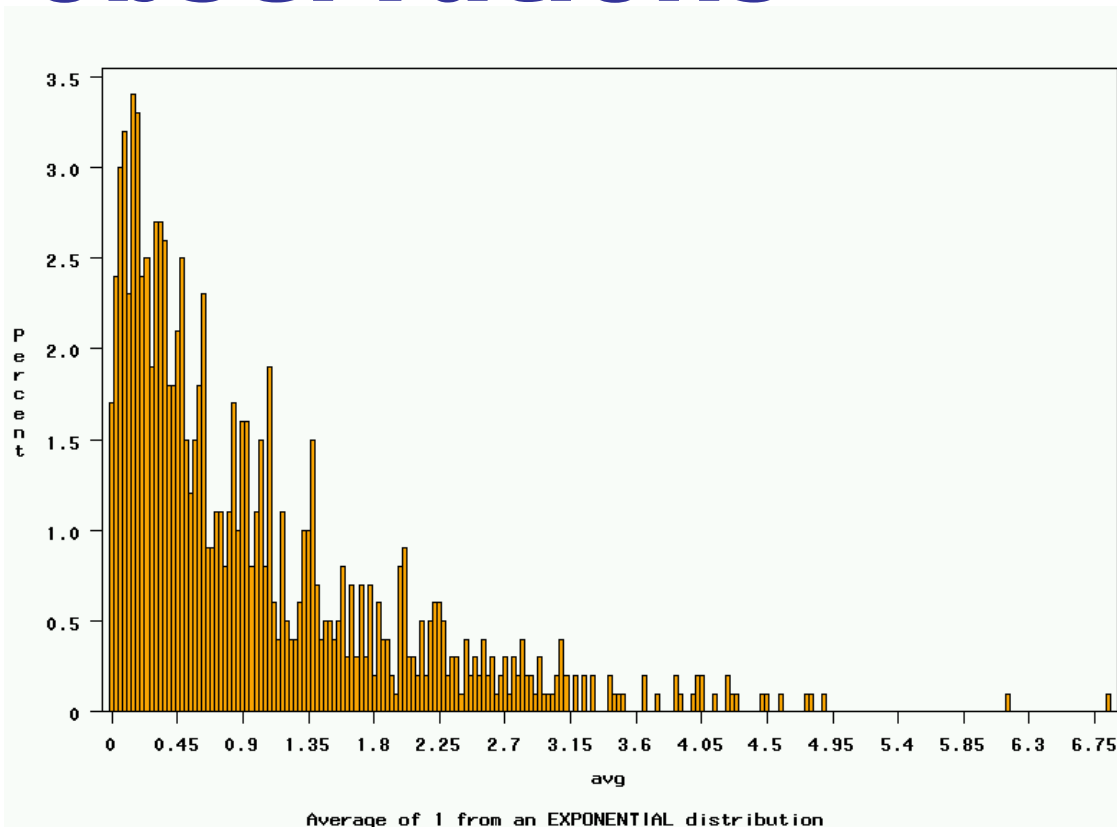
Exponential: 1000 averages of 2
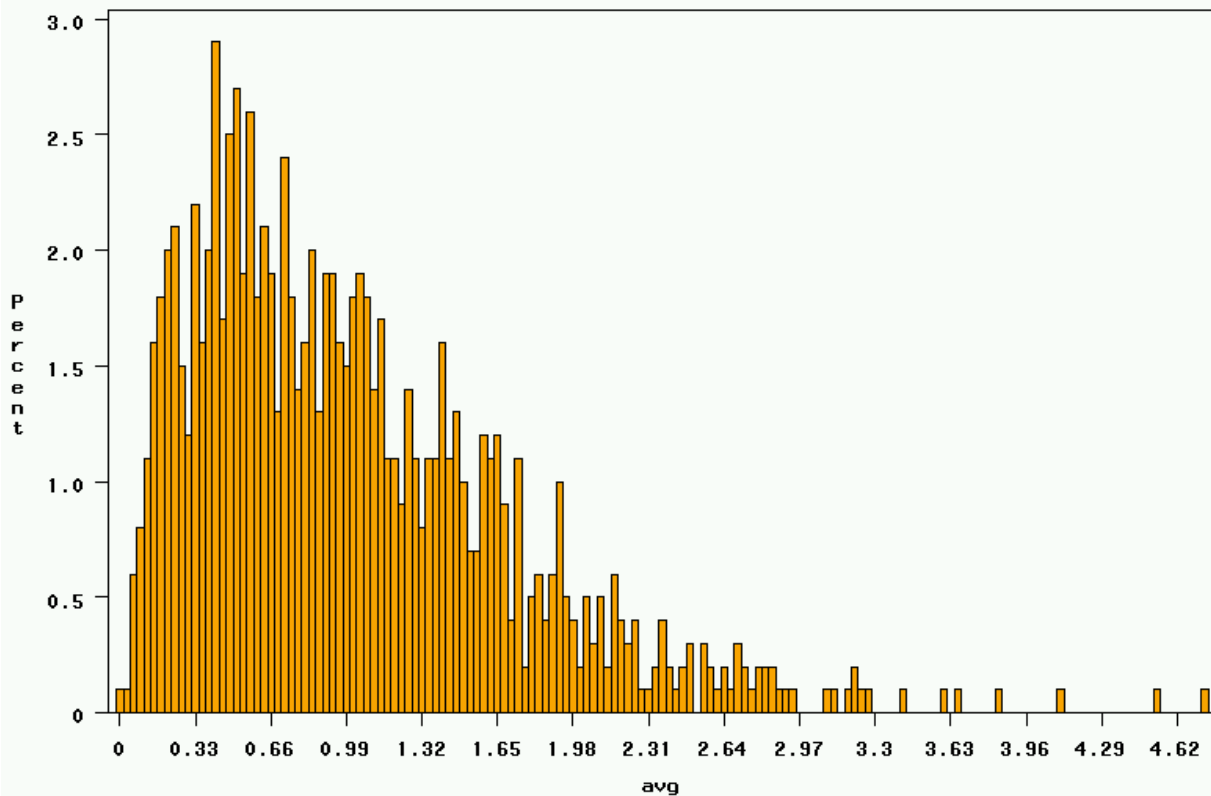
Average of 2 from an EXPONENTIAL distribution

Exponential: 1000 averages of 5
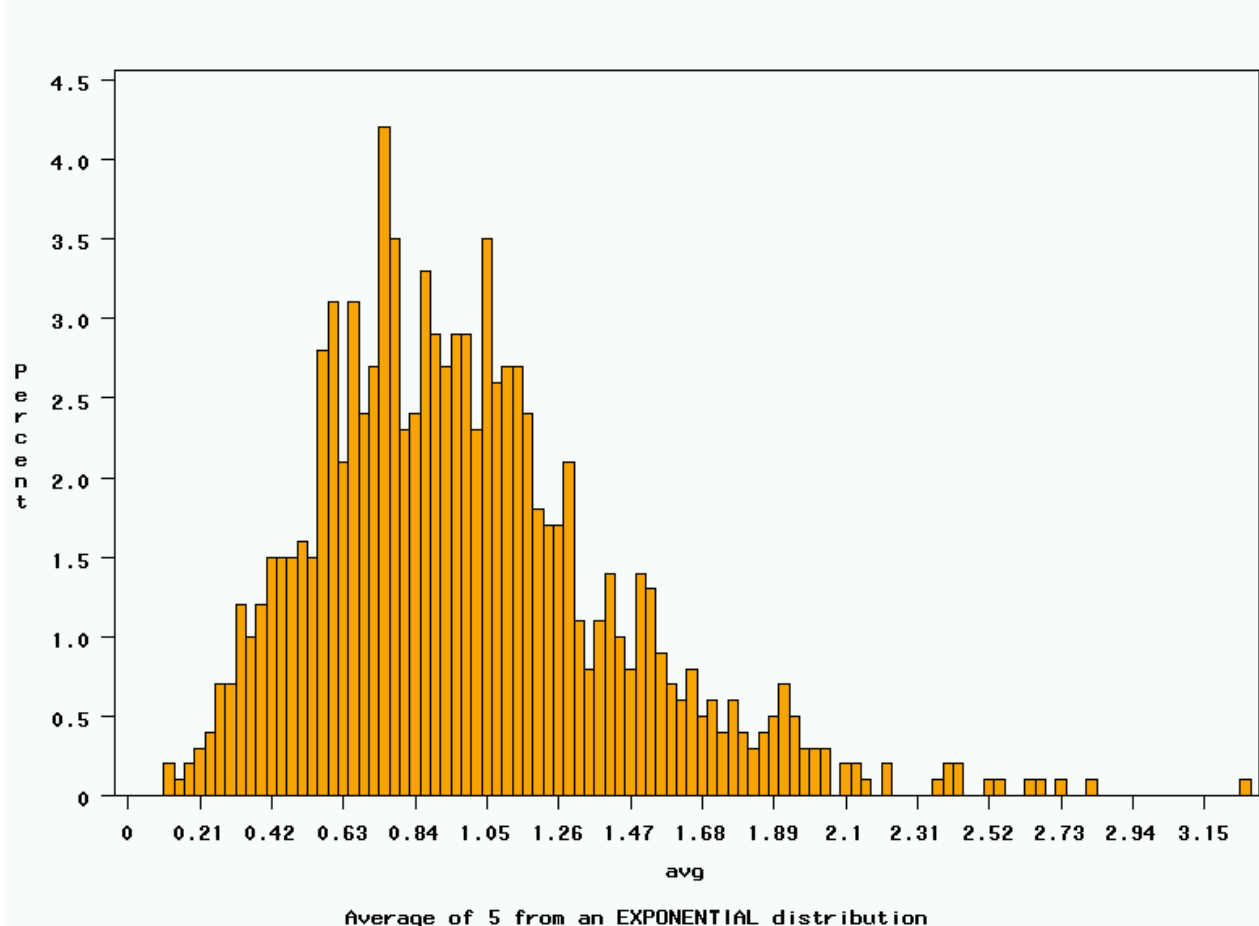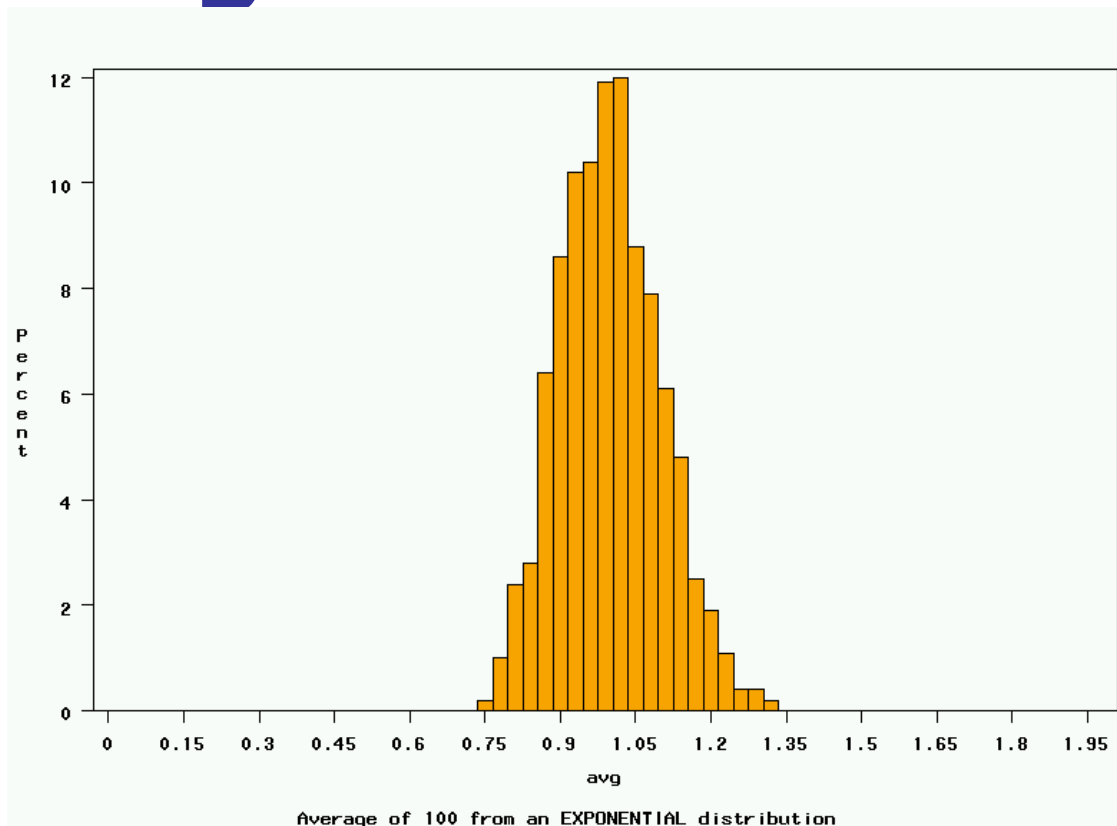
Average of 5 from an EXPONENTIAL distribution

Exponential: 1000 averages of 100

Average of 100 from an EXPONENTIAL distribution

Mathematical Theory… The Central Limit Theorem!

If all possible random samples, each of size n, are taken from any population with a mean μ and a standard deviation σ, the sampling distribution of the sample means (averages) will:

1. have mean: $$\mu_{\bar{x}} = \mu$$

2. have standard deviation (="standard error"): $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

3. be approximately normally distributed regardless of the shape of the parent population (normality improves with larger n). It all comes back to Z!

Central Limit Theorem caveats:

- For real data, sample means follow a T distribution rather than a normal distribution.
  - Since we usually don't know the true standard deviation (σ), we have to estimate it using the sample standard deviation (s); this imprecision changes the distribution to a T-distribution.
- Central Limit Theorem is a "limit theorem." For small samples (<100), if distribution of the trait is non-normal, the distribution of the means may also be non-normal (more on this in Unit 7).

Distribution of a correlation coefficient?
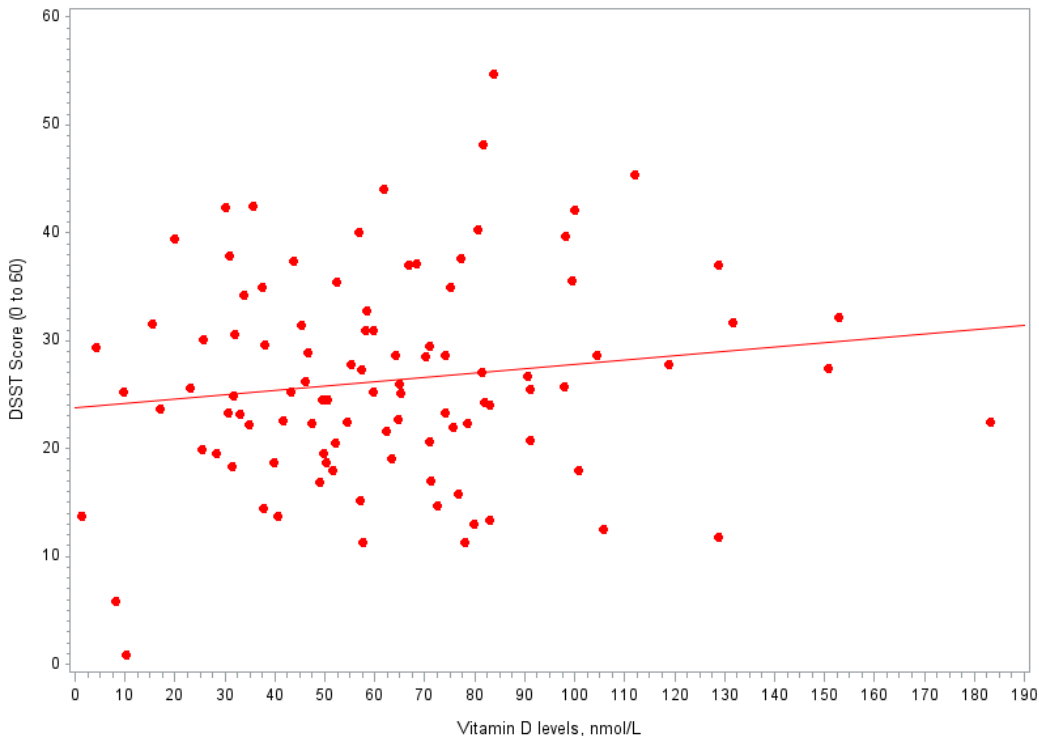
Correlation coefficient (r) ranges from -1 to +1.

0=uncorrelated

+1 = perfect positive correlation

-1 = perfect inverse correlation

Distribution of a correlation coefficient?? Computer simulation...

- 1. Specify the true correlation coefficient
  - Correlation coefficient = 0.15
- 2. Select a random sample of 100 virtual men from the population.
- 3. Calculate the correlation coefficient for the sample.
- 4. Repeat steps (2) and (3) 15,000 times
- 5. Explore the distribution of the 15,000 correlation coefficients.

Distribution of a correlation coefficient…
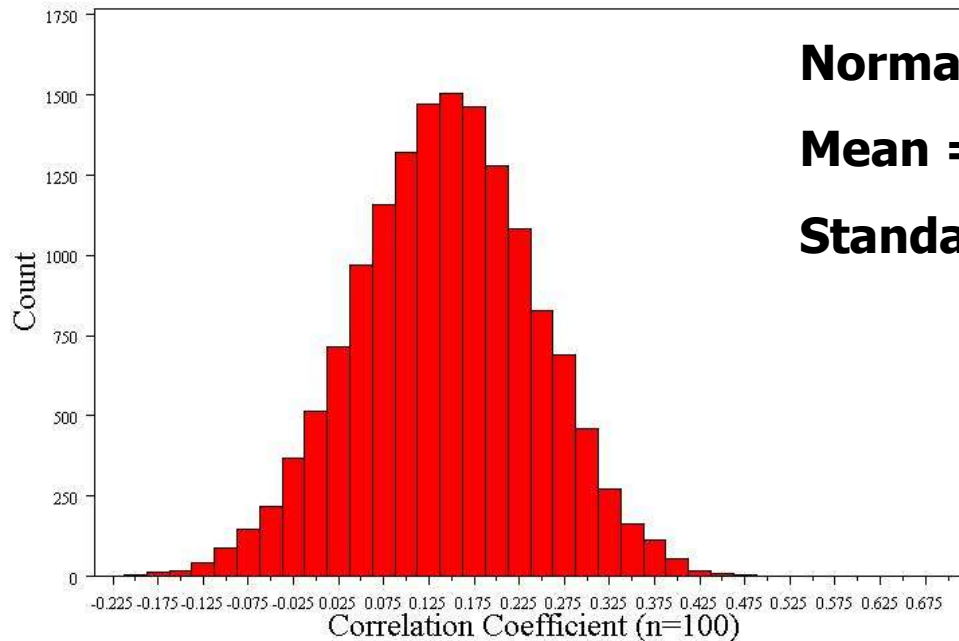
Normally distributed!

Mean = 0.15 (true correlation)

Standard error = 0.10

Distribution of a correlation coefficient in general...

- **1. Shape of the distribution**
    - Normal distribution for larger n!
    - T-distribution for smaller n (<100).
- **2. Mean = true correlation coefficient (r)**
- **3. Standard error** $\approx \sqrt{\dfrac{1-r^2}{n}}$

Many statistics follow normal or T distributions

- T-distribution (normal for large samples)
  - Means/difference in means
  - Correlation coefficients
- Normal distribution (for small or large n)
  - Proportions/difference in proportions
  - Natural log of the odds ratio

Simulation for an odds ratio…

- 1. Assume an infinite population of cases and controls with equal proportion of smokers (exposure), p=.20

- 2. Randomly select n=50 cases and n=50 controls each with p=.20 chance of being a smoker. (binomial distribution!)

- 3. Calculate the observed odds ratio.

- 4. Repeat this 1000 times (or some large number of times).

- 5. Observe the distribution of odds ratios.

Distribution of the odds ratio

Simulated results from 1000 case−control studies
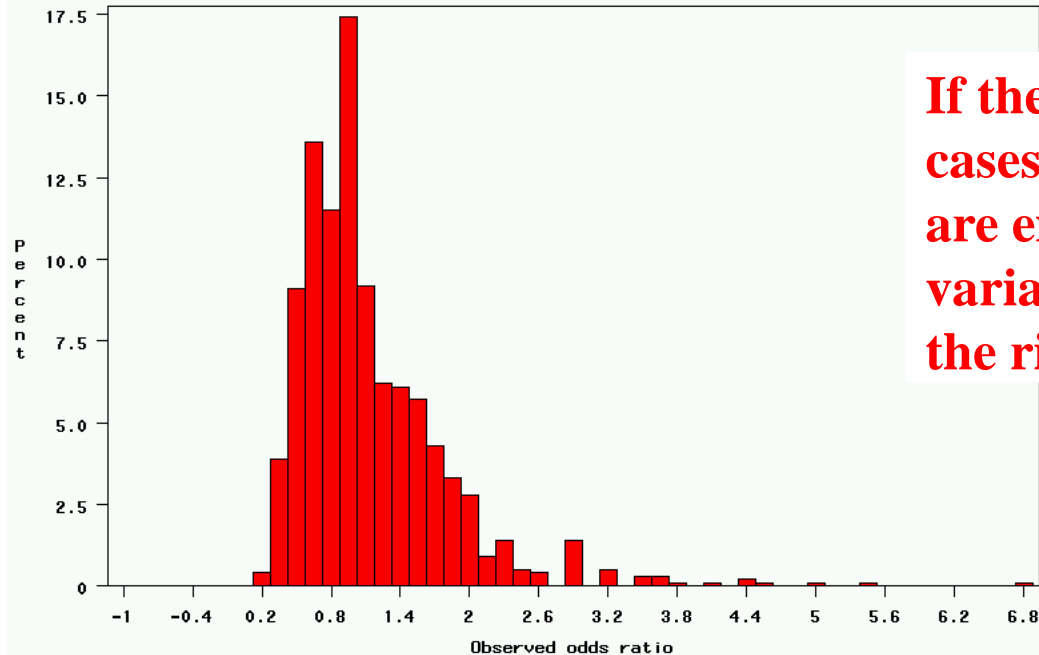
If the Odds Ratio=1.0 then with 50 cases and 50 controls, of whom 20% are exposed, this is the expected variability of the sample OR→note the right skew

Distribution of the natural log of the odds ratio...

Simlulated results from 1000 case–control studies
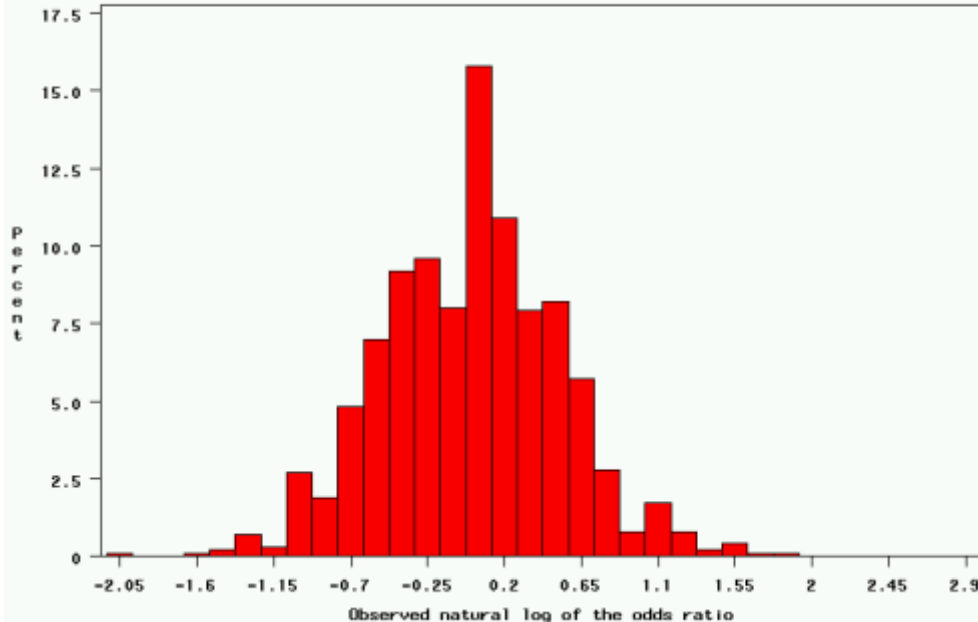
Standard error of the ln(OR)=

$$\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

a=exposed cases

b=exposed controls

c=unexposed cases

d=unexposed controls

Summary: the distribution of a statistic

- 1. Shape
    - Many statistics follow a normal distribution or a T-distribution.
- 2. Mean = true value of the statistic
- 3. The standard error varies by statistic, but is always inversely related to sample size. Three examples:
    - Standard error of the mean: $\frac{s}{\sqrt{n}}$
    - Standard error of a correlation coefficient: $\sqrt{\frac{1-r^2}{n}}$
    - Standard error of an odds ratio: $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$

Statistics in Medicine

Module 5:

Confidence Intervals (estimation)

Sample data on vitamin D: 100 men

Right-skewed!

Mean= 63 nmol/L

Standard deviation = 33 nmol/L
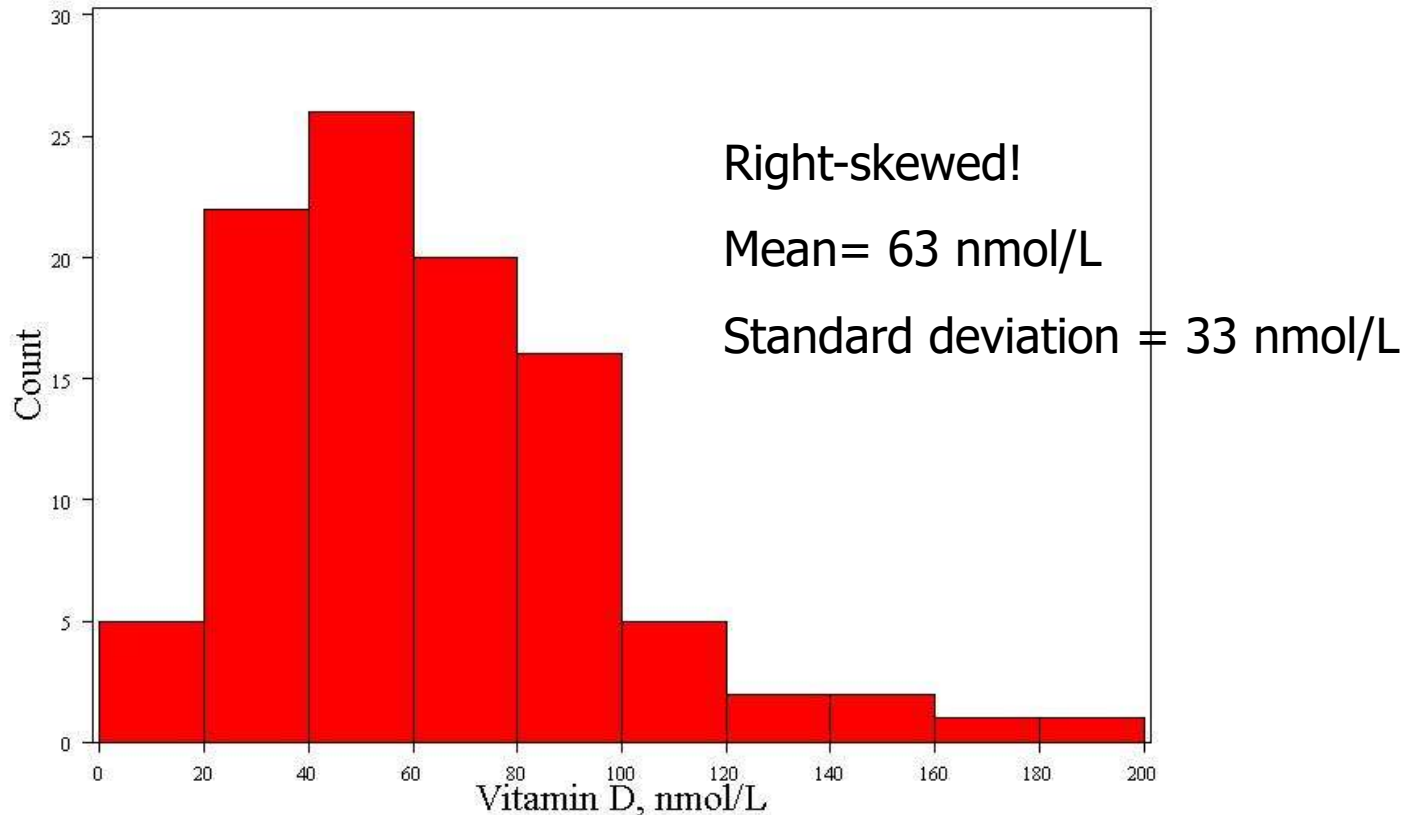
Estimation (confidence intervals)…

- What is a good estimate for the true mean vitamin D in the population (the population parameter)?
  - 63 nmol/L +/- margin of error

The confidence interval

- Goal: capture the true effect (e.g., the true mean) most of the time.
- A 95% confidence interval should include the true effect about 95% of the time.
- A 99% confidence interval should include the true effect about 99% of the time.

Recall: 68-95-99.7 rule for normal distributions! These is a 95% chance that the sample mean will fall within two standard errors of the true mean.

For example, if the true mean is 62, then 95% chance that the sample mean will be between 62 +/- 2*3.3 = 55.4 nmol/L to 68.6 nmol/L
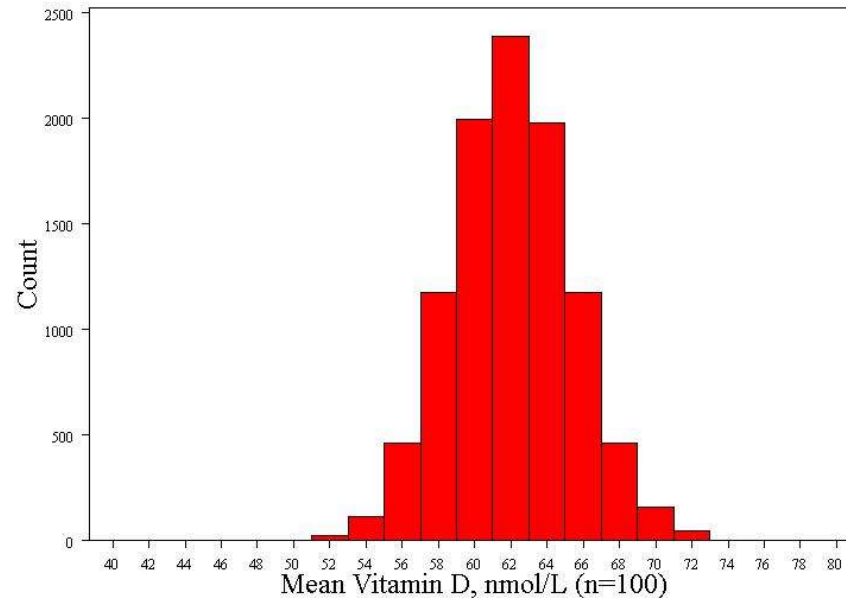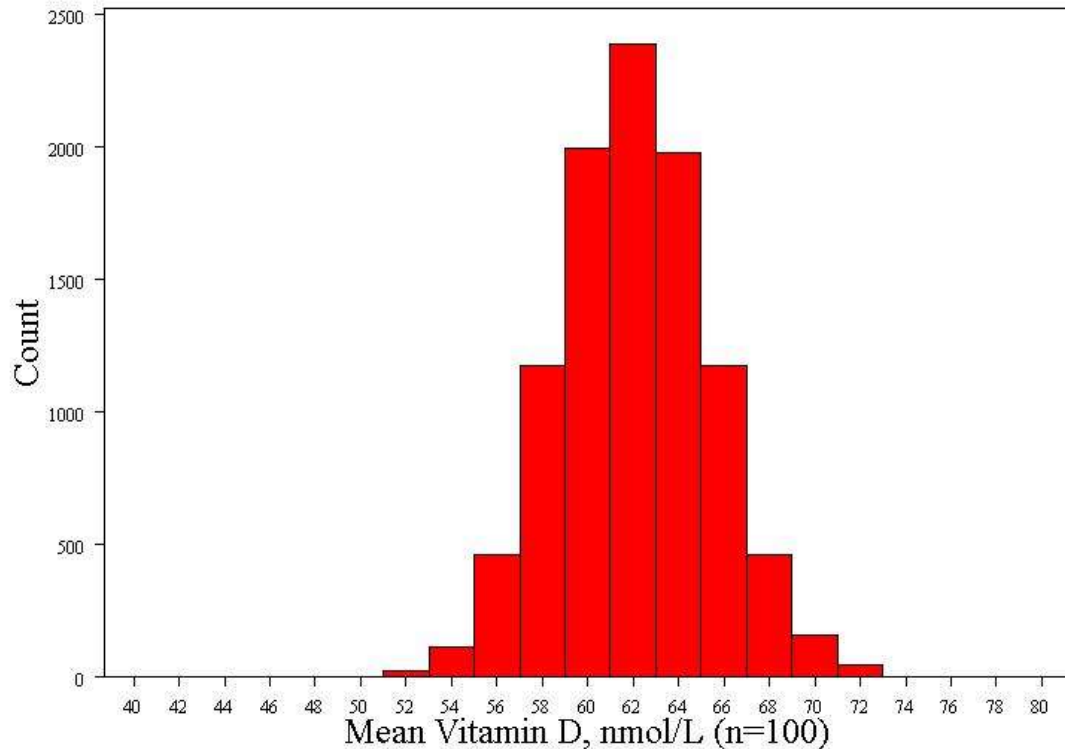
There is a 95% chance that the sample mean will fall between 55.4 nmol/L and 68.6 nmol/L.

Thus, for every sample mean in this range, sample mean +/- 2 standard errors will include the true mean.

95% confidence interval

- Thus, for normally distributed statistics, the formula for the 95% confidence interval is:

- sample statistic $\pm$ 2 x (standard error)

- Example:

  - 95% CI for mean vitamin D:

    - 63 nmol/L $\pm$ 2 x (3.3) = 56.4 – 69.6 nmol/L

    Does it contain the true mean? YES!

More precisely, use 1.96:

- For normally distributed statistics, the formula for the 95% confidence interval is, more precisely:
- sample statistic $\pm$ 1.96 x (standard error)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750

To be precise, if we're on a normal curve, 95% of observations fall between Z=-1.96 and Z= +1.96 (so the "2" is a rounded number)...

CI for correlation coefficient

- 95% CI for the correlation coefficient:
  - $0.15 \pm 2\ x\ (0.1) = -.05 - .35$
  - Does it contain the true correlation coefficient? YES!

Simulation of 20 studies of 100 men, mean vitamin D...

Vertical line indicates the true mean (62)

95% confidence intervals for the mean vitamin D for each of the 20 simulated studies.

Only 1 confidence interval missed the true mean.
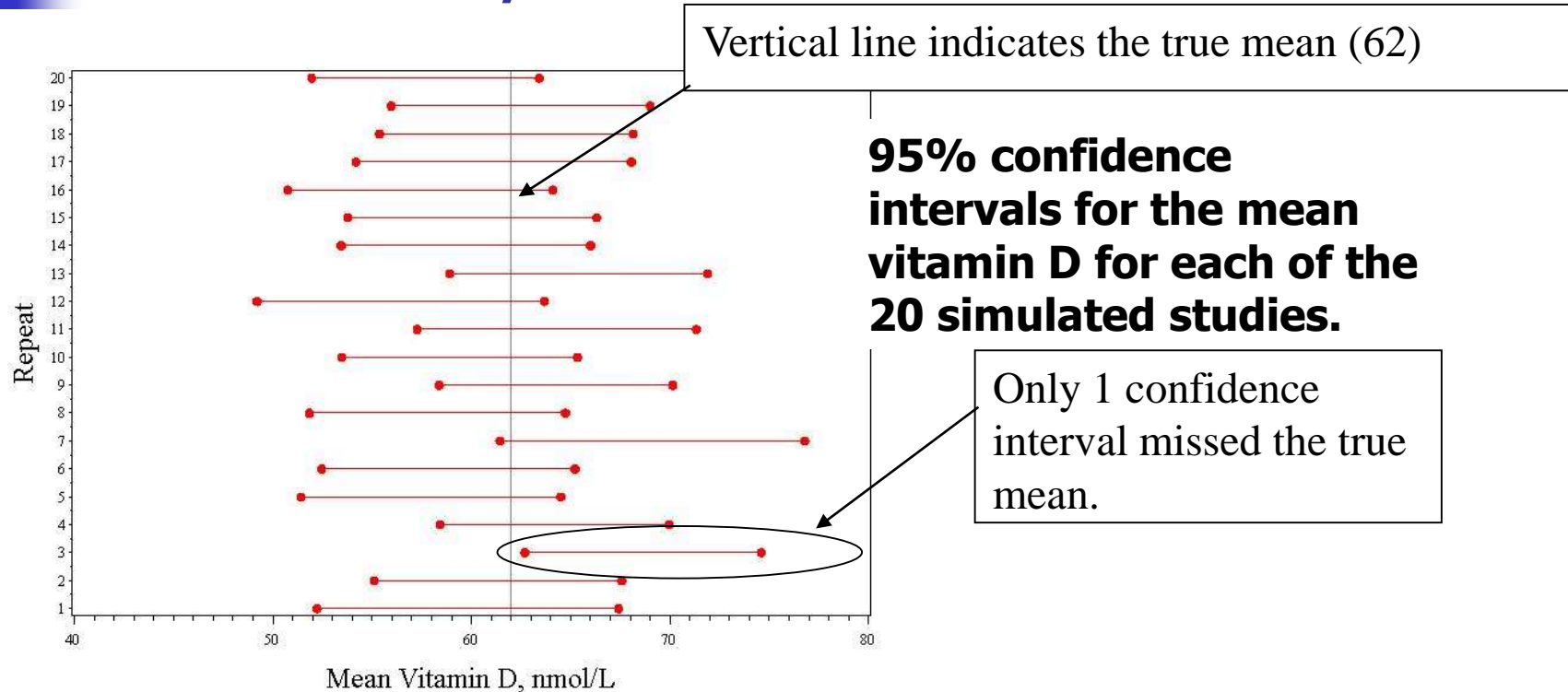
Confidence Intervals, general formula:

The value of the statistic in my sample (eg., mean, odds ratio, etc.)

$$\textit{point estimate} \pm \textit{(measure of how confident we want to be)} \times \textit{(standard error)}$$

From a Z table or a T table, depending on the sampling distribution of the statistic (Z-value).

Standard error of the statistic.

Precise Z values for various confidence levels

Confidence level	Z-value
90%	1.64
95%	1.96
99%	2.58

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z i.e.

$$P[\ Z < z\ ] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\ dZ$$

P[Z < z]

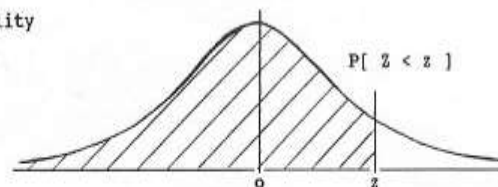| z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

99% confidence intervals…

- 99% CI for mean vitamin D:
  - 63 nmol/L $\pm$ 2.6 x (3.3) = 54.4 − 71.6 nmol/L
- 99% CI for the correlation coefficient:
  - 0.15 $\pm$ 2.6 x (0.1) = -.11 − .41

Confidence Intervals give:

*A plausible range of values for a population parameter.

*The precision of an estimate. (When sampling variability is high, the confidence interval will be wide to reflect the uncertainty of the observation.)

*Statistical significance (if the 95% CI does not cross the null value, it is significant at .05)

Statistics in Medicine

Module 6:

Where does the margin of error come from in polls?

Where does the "margin of error" come from in polls?

- New York Times/CBS News Poll (Feb. 2012) found: "59 percent said the health insurance plans of religiously affiliated employers should cover the cost of birth control."

- "The nationwide telephone poll included 1,064 registered voters. The margin of sampling error is plus or minus three percentage points."

Where does the margin of error come from?

MARJORIE CONNELLY. Support Is Found for Birth Control Coverage and Gay Unions. *New York Times*, February 14, 2012

What is the statistic here?

- Statistic = proportion (out of n=1064)
  - Proportion who believe that religiously affiliated employers should cover the cost of birth control.

What's the distribution of a proportion?

- ## 1. Shape: Normal distribution!
  - A proportion is just a binomial divided by N (the total count); and a binomial can be approximated by a normal if np>5

- ## 2. Mean: true proportion
  - True proportion (p) of all registered voters who believe that religiously affiliated employers should cover the cost of birth control

- ## 3. Standard Error?
  - Related to the variance of a binomial…

Standard error of a proportion:

For binomial:
$$\sigma_x = \sqrt{np(1-p)}$$

Differs by a factor of n.

For proportion:
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

What's the distribution of a proportion?

- ## 1. Shape: Normal distribution!
  - A proportion is just a binomial divided by N (the total count); and a binomial can be approximated by a normal if np>5

- ## 2. Mean: true proportion
  - True proportion (p) of all registered voters who believe that religiously affiliated employers should cover the cost of birth control

- ## 3. Standard Error
  - $= \sqrt{\dfrac{p(1-p)}{n}}$

95% confidence interval for a proportion

point estimate ± 1.96 × (standard error)

95% confidence interval for this survey:
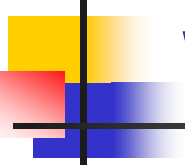
point estimate ± 1.96 × (standard error)

$$\text{standard error} = \sqrt{\frac{.59 * .41}{1064}} = .015$$

$$0.59 \pm 1.96 * .015 = 0.59 \pm .03$$

Margin of error for any survey with sample size N…

- p(1-p) reaches a maximum when p=.5
- Thus, the maximum possible standard error is:

$$\sqrt{\frac{.5 * .5}{N}} = \sqrt{\frac{.25}{N}} = \frac{.5}{\sqrt{N}}$$

- The width of the 95% confidence interval $= \pm \; 2 \times$ standard error
- Maximum width of the 95% confidence interval is thus:

$$2 * \frac{.5}{\sqrt{N}} = \frac{1}{\sqrt{N}}$$

Thus a general margin of error is:

$$\frac{1}{\sqrt{N}}$$

$$\frac{1}{\sqrt{1064}} = .03 = 3\%$$

Statistics in Medicine

Module 7:

Hypothesis testing (p-values)

Testing Hypotheses

- 1. Is the mean vitamin D in middle-aged and older European men lower than 100 nmol/L (the "desirable" level)?

- 2. Is cognitive function correlated with vitamin D?

Is the mean vitamin D lower than 100 nmol/L

- Start by assuming that the mean = 100
- This is the "null hypothesis"
- This is usually the "t arget" that we are interested in
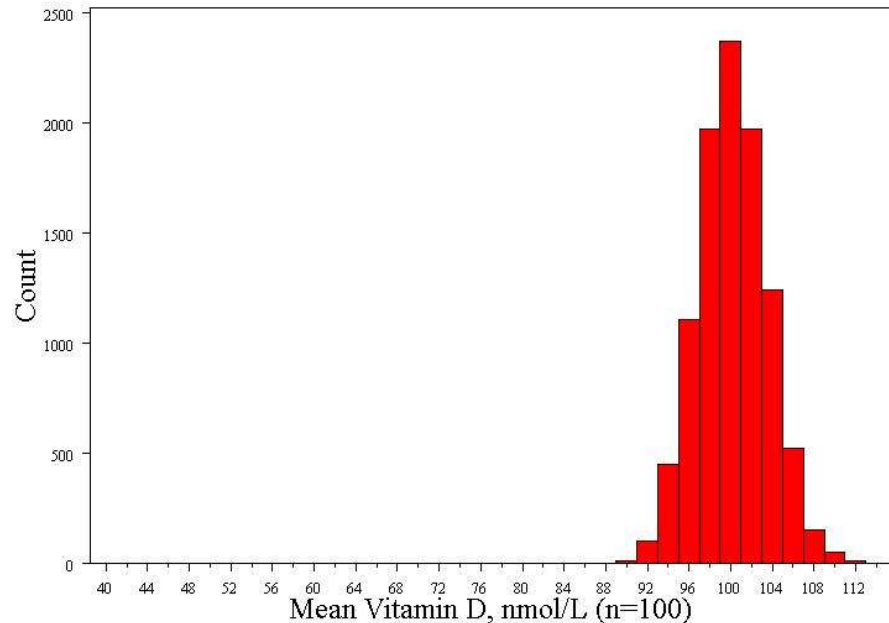- Determine the distribution of statistics assuming that the null is true…

Computer simulation (10,000 repeats)…

This is called the null distribution!
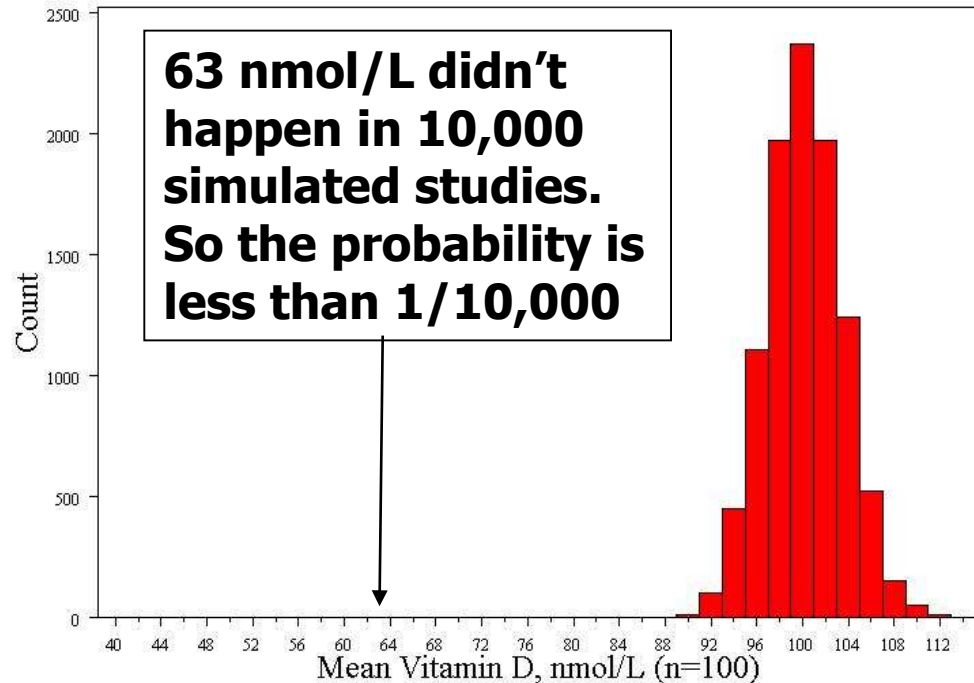
Normally distributed
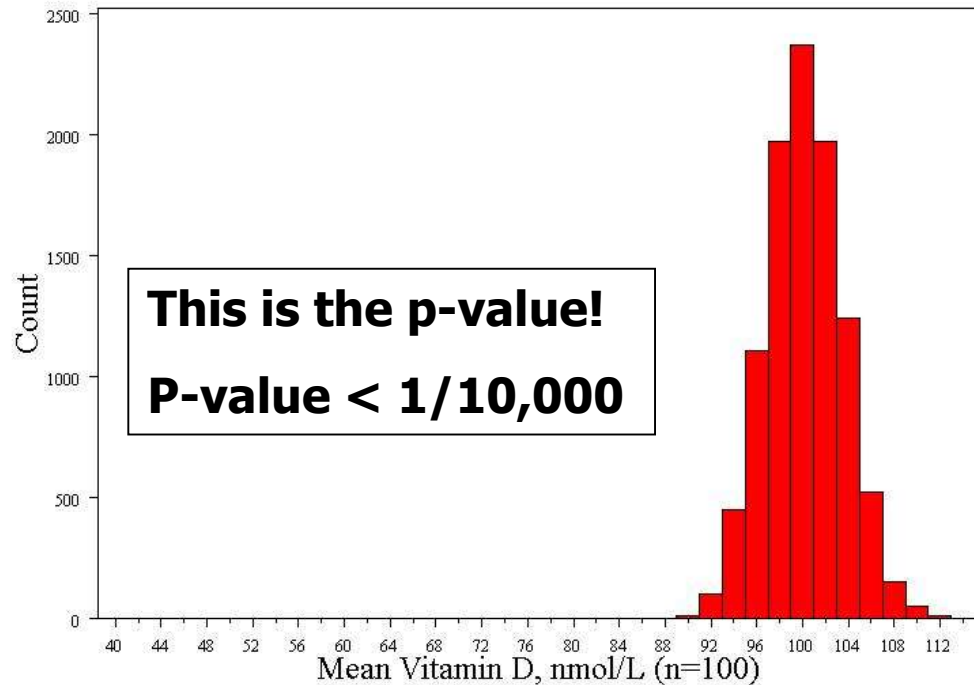
Std error = 3.3

Mean = 100

Compare the null distribution to the observed value...

What's the probability of seeing a sample mean of 63 nmol/L if the true mean is 100 nmol/L?

63 nmol/L didn't happen in 10,000 simulated studies. So the probability is less than 1/10,000

Compare the null distribution to the observed value...

This is the p-value!

P-value < 1/10,000

P-value from a math formula:

Because we know the general distribution of a sample mean, we know:

1. We are on a nearly normal curve.
2. The mean of this normal curve = 100 (the null mean)
3. The standard error is: $\dfrac{s}{\sqrt{n}}$

$$Z = \frac{63 - 100}{\frac{33}{\sqrt{100}}} = \frac{63 - 100}{3.3} = -11.2; p < .00001$$

P-values

The p-value = P(data/null hypothesis)

P-values give the probability that the observed effect could have arisen by chance (when the null hypothesis is true).

The P-value

In this case, our data are so unlikely given the null hypothesis (p<.0001) that I'm going to reject the null hypothesis!

Hypothesis Testing

The Steps:

1. **Define your hypotheses (null, alternative)**

    ➢ The null hypothesis is the "target" that we are trying to shoot down.

    ➢ Null here: "mean vitamin D level = 100 nmol/L"

    ➢ Alternative here: "mean vit D < 100 nmol/L" (<u>one-sided</u>)

2. **Specify your null distribution**

    ➢ If we repeated this experiment many, many times, the mean vitamin D would be normally distributed around 100 nmol/L with a standard error of 3.3.

$$\frac{33}{\sqrt{100}} = 3.3$$

Hypothesis Testing

3.   Do an experiment

➢   observed sample mean = 63 nmol/L

4.   Calculate the p-value of what you observed

➢   p<.0001

5.   Reject or fail to reject the null hypothesis

➢   Reject; conclude that average vitamin D is "significantly" lower than 100 nmol/L

Summary: Hypothesis Testing

The Steps:
1. Define your hypotheses (null, alternative)
2. Specify your null distribution
3. Do an experiment
4. Calculate the p-value of what you observed
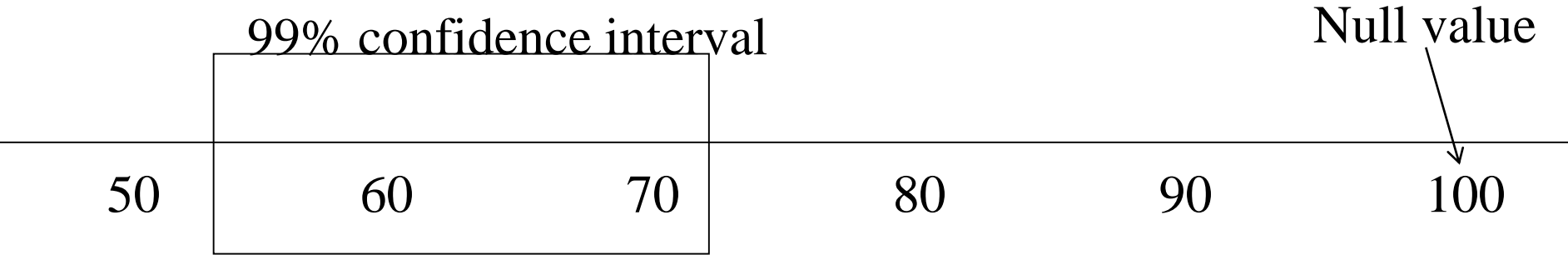5. Reject or fail to reject the null hypothesis

- Confidence intervals give the same information (and more) than hypothesis tests…

Duality between confidence intervals and p-values.

95% confidence interval

Null value

50   60   70   80   90   100
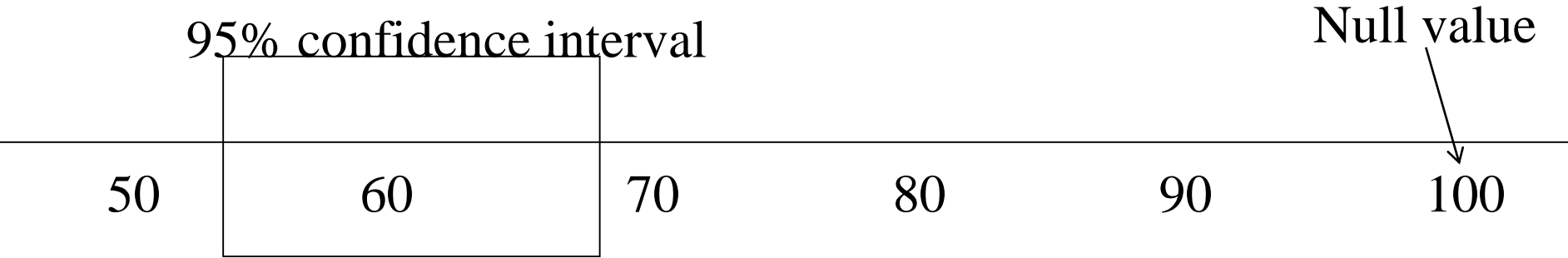
P-value < .05

Duality between confidence intervals and p-values.

99% confidence interval

Null value

50   60   70   80   90   100

P-value < .01

Example 2: Is cognitive function correlated with vitamin D?

- Null hypothesis: $r = 0$ (no correlation)
- Alternative hypothesis: $r \neq 0$
  - Two-sided hypothesis
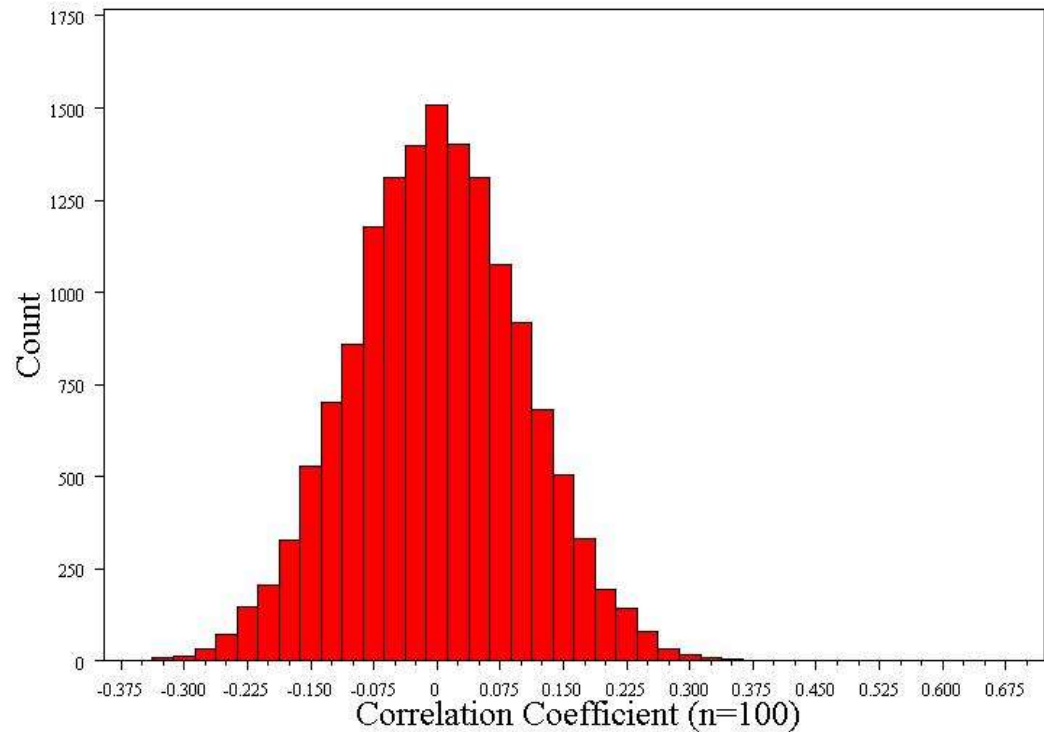 - Doesn't assume that the correlation will be positive or negative.

Computer simulation (15,000 repeats)
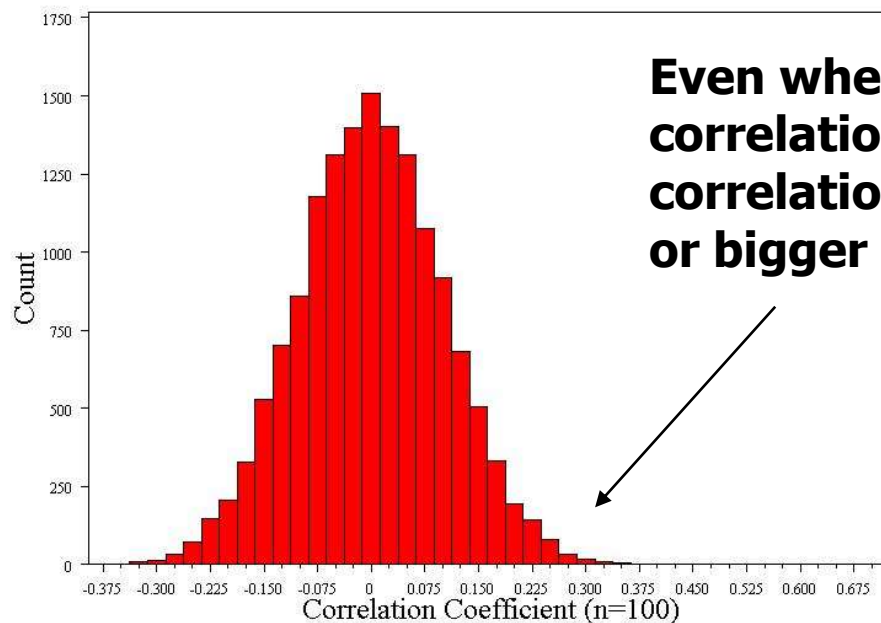
Null distribution:

Normally distributed
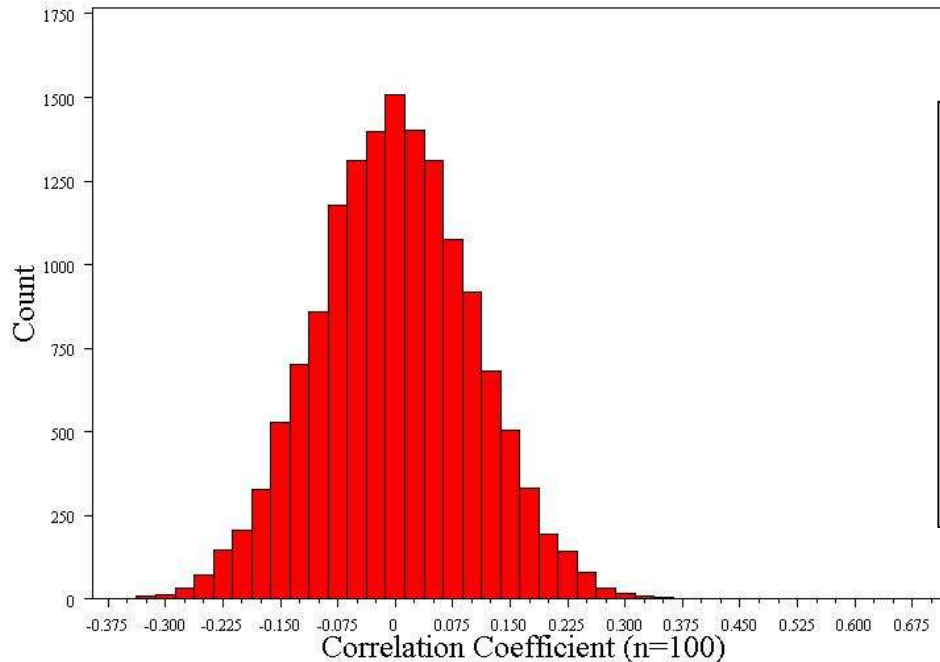
Std error = 0.10

Mean = 0

What's the probability of our data?

Even when the true correlation is 0, we get correlations as big as 0.15 or bigger 7% of the time.
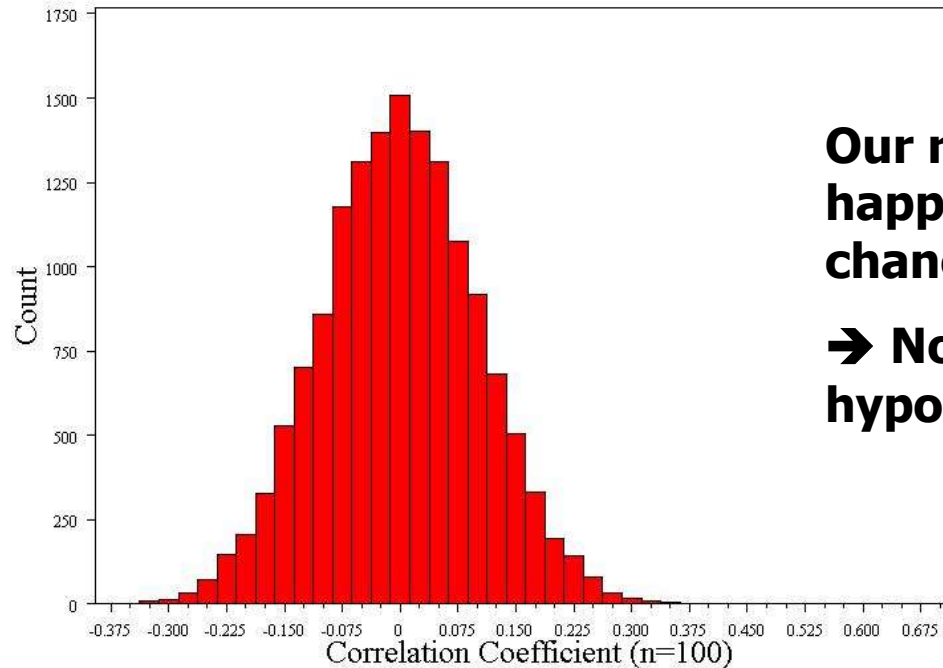
What's the probability of our data?

This is a two-sided hypothesis test, so "more extreme" includes as big or bigger negative correlations (<-0.15).

P-value = 7%+7%=14%

What's the probability of our data?

Our results could have happened purely due chance!

➔ **No evidence to reject null hypothesis**

Formal hypothesis test

- 1. Null hypothesis: r=0
  - Alternative: r $\neq$ 0 (two-sided)
- 2. Determine the null distribution
  - Normally distributed; r=0; standard error = 0.1
- 3. Collect Data, r=0.15
- 4. Calculate the p-value for the data:
  - Z = $\frac{0.15 - 0}{.1} = 1.5$ **p=.14**
- 5. Reject or fail to reject the null
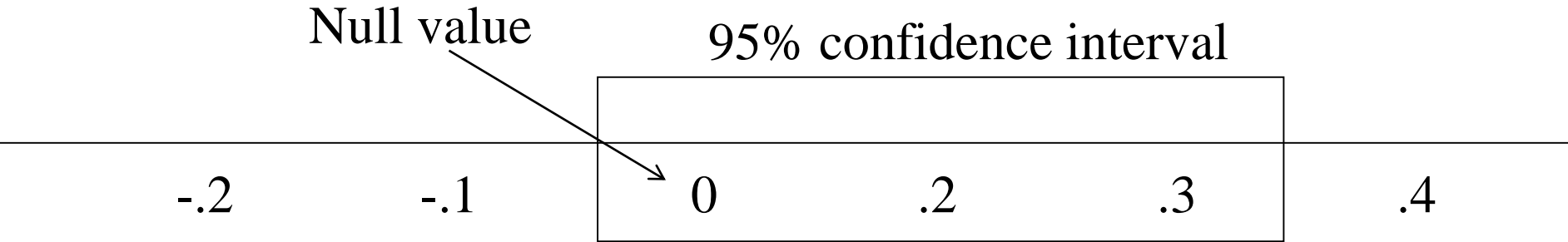 - Not enough evidence to reject the null hypothesis.

Or use confidence interval to gauge statistical significance...

- 95% CI = -0.05 to 0.35
- Thus, 0 (the null value) is a plausible value!
- P>.05

Duality between p-values and confidence intervals.

Null value

95% confidence interval

-.2   -.1   0   .2   .3   .4

P-value > .05

P-values mathematically

$$Z = \frac{\text{observed mean} - \text{null value}}{\text{standard error}}$$

P-value from a mathematical formula (mean)....

Effect size (difference of the observed mean from the null mean).

Larger **Z-values** correspond to smaller **p-values**. A Z-value of 11.2 corresponds to a p-value of <.0001.

$$Z = \frac{63 - 100}{\frac{33}{\sqrt{100}}} = 11.2$$

The **standard deviation** of vitamin D (of the trait).

Sample size

P-value from a mathematical formula (correlation coefficient)....

Effect size (weak correlation)

Larger **Z-values** correspond to smaller **p-values**. A Z-value of 1.5 corresponds to a p-value of 0.14.

$$Z = \frac{.15}{\sqrt{\dfrac{1-.15^2}{100}}} = 1.5$$

A measure of **variability** for a correlation coefficient.

Sample size

P-values

- P-values depend on: effect size, sample size, and variability.

The P-value

- By convention, p-values of <.05 are often accepted as "statistically significant" in the medical literature; but this is an arbitrary cut-off.

- A cut-off of p<.05 means that, when there are no effects, about 1 in 20 experiments will appear significant just by chance ("Type I error").
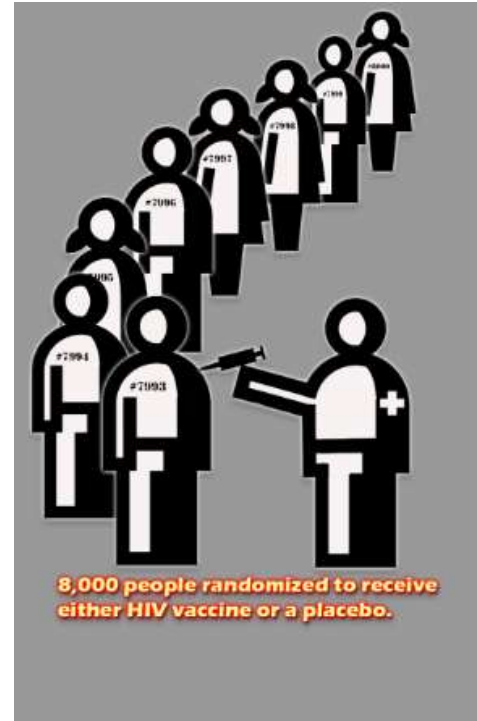
Statistics in Medicine

Module 8:

HIV vaccine trial/Bayesian inference

HIV vaccine trial

- Thai HIV vaccine trial (2009)
  - 8197 randomized to vaccine
    - RV144, a combination of two vaccines: ALVAC HIV (the prime) and AIDSVAX B/E (the boost)
  - 8198 randomized to placebo
- Generated a lot of public discussion about p-values!

8,000 people randomized to receive either HIV vaccine or a placebo.

Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, et al. Vaccination with ALVAC and AIDSVAX to Prevent HIV-1 Infection in Thailand. *N Engl J Med* 2009; 361:2209-2220.

Results:

51 infections in the vaccine group vs. 74 infections in the placebo group

Analogy: *If the vaccine doesn't work at all*, this outcome would be similar to flipping a coin 125 times and getting 51 heads and 74 tails.

The value of a p-value!

- Could this difference have arisen just due to chance, i.e. if the vaccine were completely ineffective?

- I don't have good intuition on this. Hence, the need for a p-value!
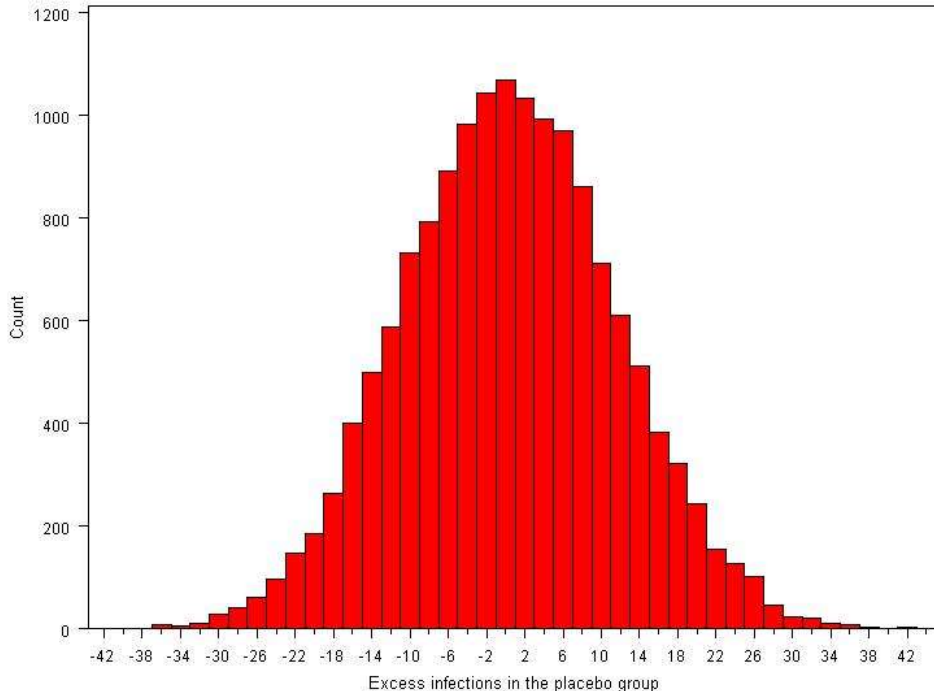
Start with the null hypothesis

- Null Hypothesis: The vaccine has no effect.
  - infection rate/risk in placebo group=infection rate/risk in vaccine group

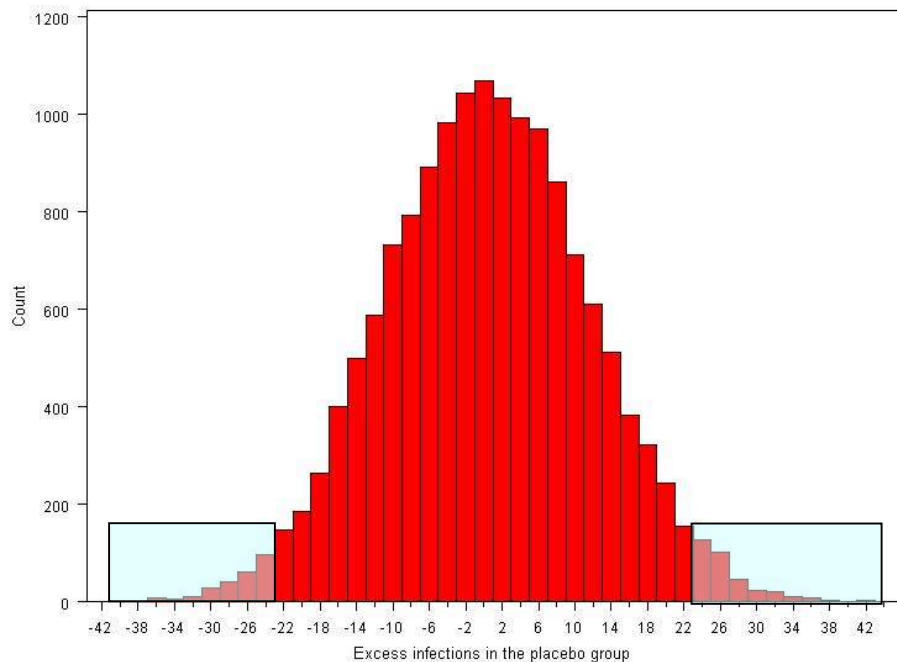# Computer simulation assuming the null (15,000 repeats)...

Distribution of the statistic (expressed as excess infections in the placebo group):

Shape: Normally distributed

Mean = 0 excess infections

Standard Error = 11.1 excess infections

Computer simulation assuming the null (15,000 repeats)...

If the vaccine has no effect, we could still get 23 or more excess infections just by chance.

Probability of 23 or more excess infections = 4%

That's the p-value! (two-sided)

Identical to the probability of getting 51 or fewer heads or 51 or fewer tails in 125 fair coin flips!

P-value: interpretation

- A p-value of 4% tells us that we could have observed an imbalance this big if the vaccine were ineffective, but the probability is low (4%). This gives us some reason to reject the null hypothesis.

- How low is low enough to be convinced that the vaccine has some real effect?

- By arbitrary convention, p-values < .05 are deemed "statistically significant."

- But nothing magical happens at .05!

Controversy over the trial!

- The authors used the "modified intention to treat population" (which excludes some randomized participants) for their primary analysis.

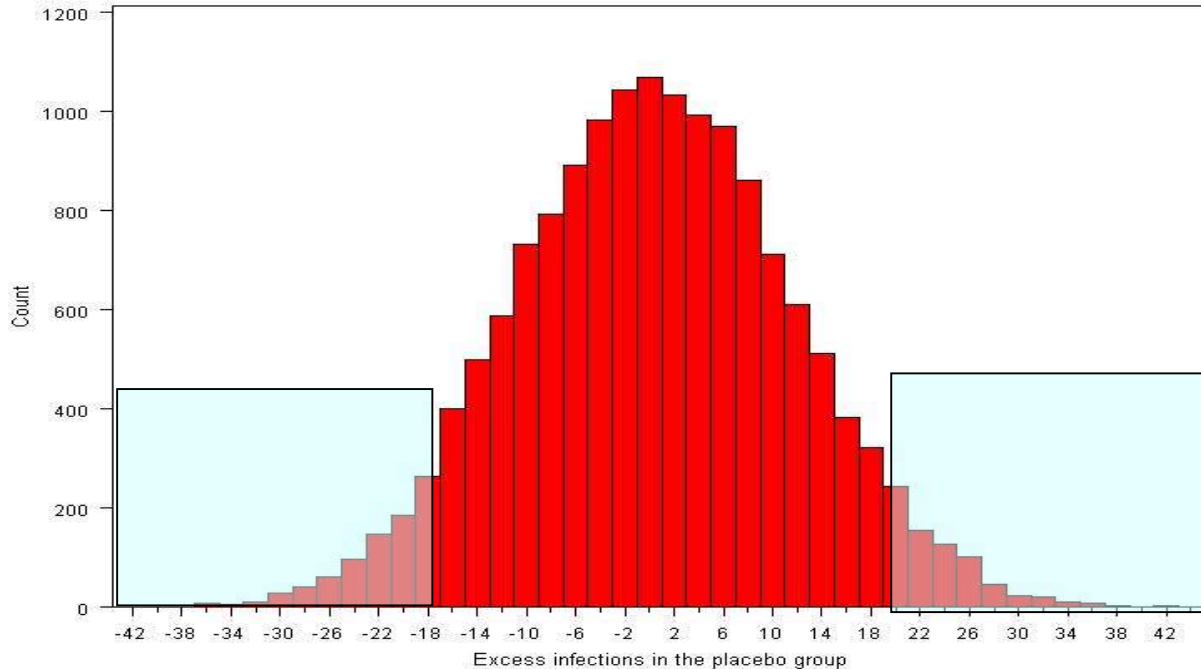- The "intention to treat" analysis (with no exclusions) had slightly less favorable results…

Alternative analysis of the data ("intention to treat")...

- 56/8202 infections in the vaccine group versus 76/8200

- P-value = 0.08

Computer simulation assuming the null (15,000 repeats)...

Probability of 20 or more excess infections = 0.08

P=.08 is only slightly different than p=.04!

Compare effect sizes, not p-values!!

Risk reduction expressed as:	Analysis 1 (modified ITT)	Analysis 2 (strict ITT)
Whole numbers	23 fewer infections (51 vs. 74)	20 fewer infections (56 vs. 76)
Absolute risk reduction	2.8 fewer infections per 1000 people vaccinated	2.4 fewer infections per 1000 people vaccinated*
Number needed to treat (NNT)**	357 people would have to be vaccinated to prevent 1 infection	410 people would have to be vaccinated to prevent 1 infection
Relative risk (RR) reduction	31% reduction in the risk of infection among those vaccinated	26% reduction in the risk of infection among those vaccinated

How to interpret p=.04…

- P(data/null) = .04
- P(null/data) ≠ .04

- "There is only a 4% chance that the vaccine doesn't work" is MISTAKEN!

- To find P(null/data), use Bayes' Rule (and prior data on the vaccine)

Applying Bayes' Rule

$$P(null\,/\,data) = \frac{P(data\,/\,null)P(null)}{P(data)} =$$

$$\frac{P(data\,/\,null)P(null)}{P(data\,/\,null)*P(null) + P(data\,/\sim null)*P(\sim null)} =$$

$$\frac{.04*P(null)}{.04*P(null) + P(data\,/\sim null)*P(\sim null)}$$

Results:

- P(null/data) $\geq$ 22%

  *estimated using Bayes' Rule (and prior data on the vaccine)...

*Gilbert PB, Berger JO, Stablein D, Becker S, Essex M, Hammer SM, Kim JH, DeGruttola VG. Statistical interpretation of the RV144 HIV vaccine efficacy trial in Thailand: a case study for statistical issues in efficacy trials. *J Infect Dis* 2011; 203: 969-975.