# Business Analytics

## Regression Analysis

Prof. Dr. Martin Bichler

Decision Sciences & Systems

Department of Informatics

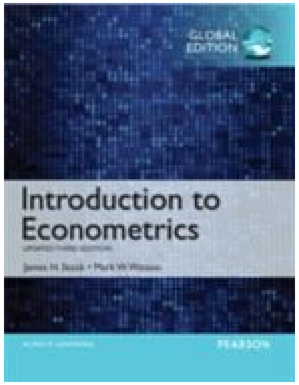
Technical University of Munich

# Course Content

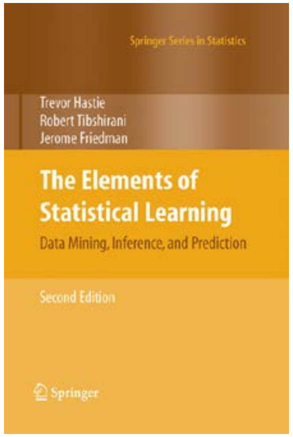
- Introduction
- **Regression Analysis**
- Regression Diagnostics
- Logistic and Poisson Regression
- Naive Bayes and Bayesian Networks
- Decision Tree Classifiers
- Data Preparation and Causal Inference
- Model Selection and Learning Theory
- Ensemble Methods and Clustering
- High-Dimensional Problems
- Association Rules and Recommenders
- Neural Networks

# Recommended Literature

- **Introduction to Econometrics**
  - Stock, James H., and Mark W. Watson
  - Chapter 2 – 7, 17, 18



- **The Elements of Statistical Learning**
  - Trevor Hastie, Robert Tibshirani, Jerome Friedman
  - [https://web.stanford.edu/~hastie/ElemStatLearn/](https://web.stanford.edu/~hastie/ElemStatLearn/)
  - Section 3.1-3.2: Linear Methods for Regression

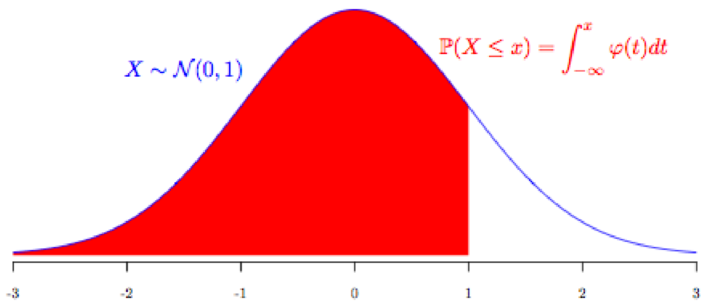

Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

- **Any Introduction to Statistics**
  (e.g.: Statistical Inference by George Casella, Roger L. Berger
  or online course http://onlinestatbook.com/)

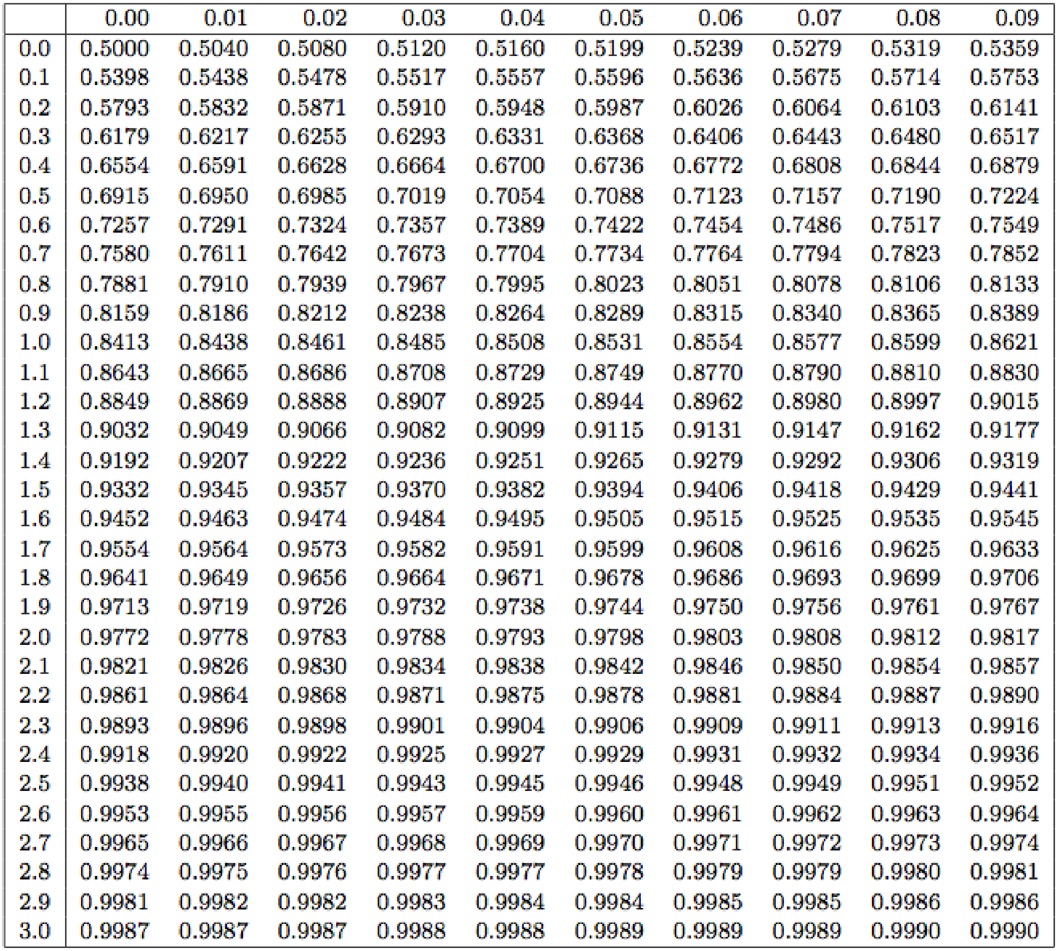
Today we revisit three important elements of statistical inference:

- Estimation, testing, regression

# Question



What is the probability that a sample of 100 randomly selected elements with a mean of 300 or more gets selected if the true population mean is 288 and the population standard deviation is 60?



| | 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 | 0.5160 | 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 | 0.7852 |
| 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 | 0.8810 | 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 | 0.9772 | 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 | 0.9864 | 0.9868 | 0.9871 | 0.9875 | 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 | 0.9925 | 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 | 0.9979 | 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 |
| 3.0 | 0.9987 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9989 | 0.9989 | 0.9989 | 0.9990 | 0.9990 |

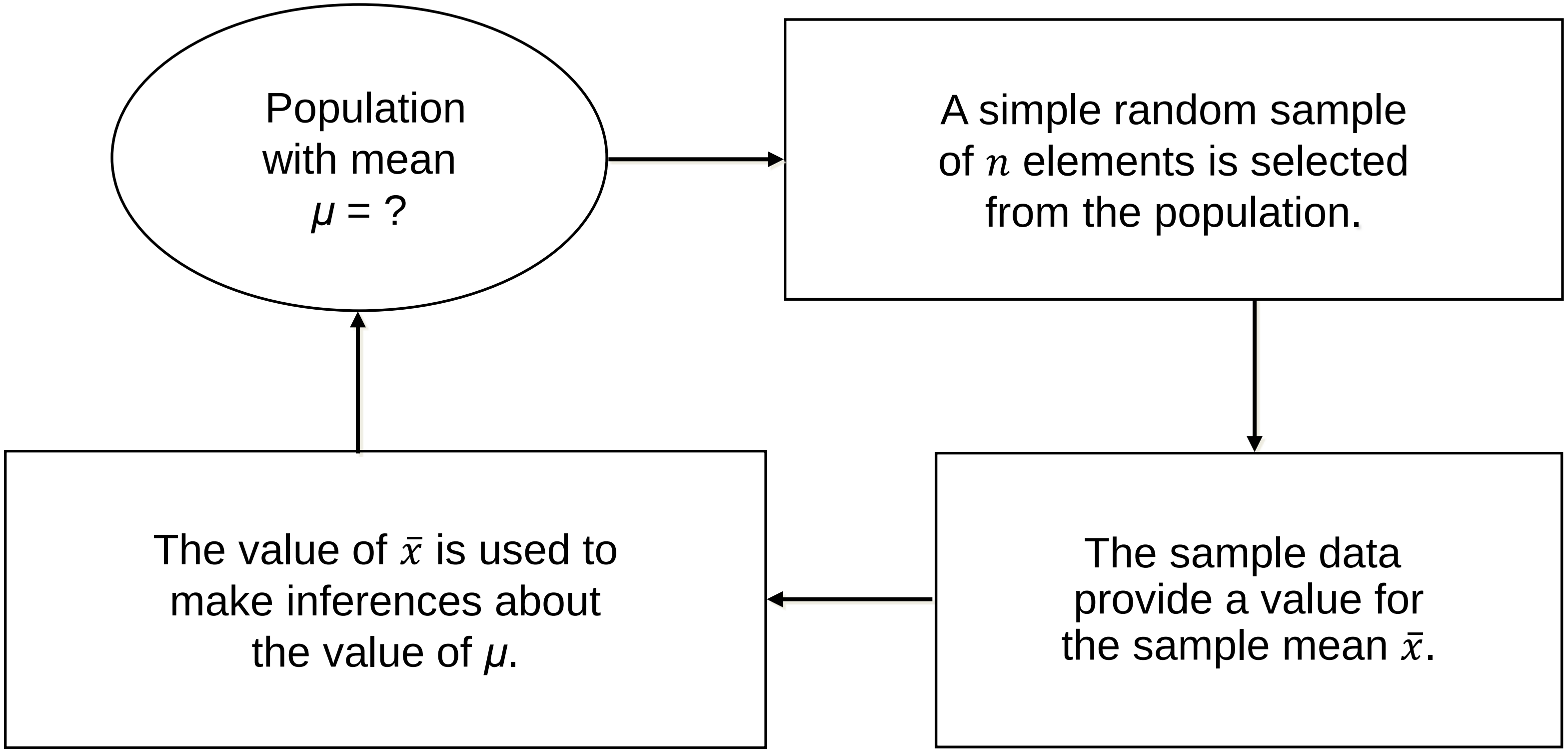
# Question

- What is the probability that a sample of 100 randomly selected elements with a mean of 300 or more get selected if the true population mean is 288 and the population standard deviation is 60?
  - The sample was randomly selected and we draw on the Central Limit Theorem.
  - We need to take the standard dev. of the sample mean.

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{300 - 288}{60 / \sqrt{100}} = 2$$

  - Check out the table of the standard normal distribution.
  - There is a 2.28% chance of selecting a sample with a mean > 300.

# Statistical Estimation



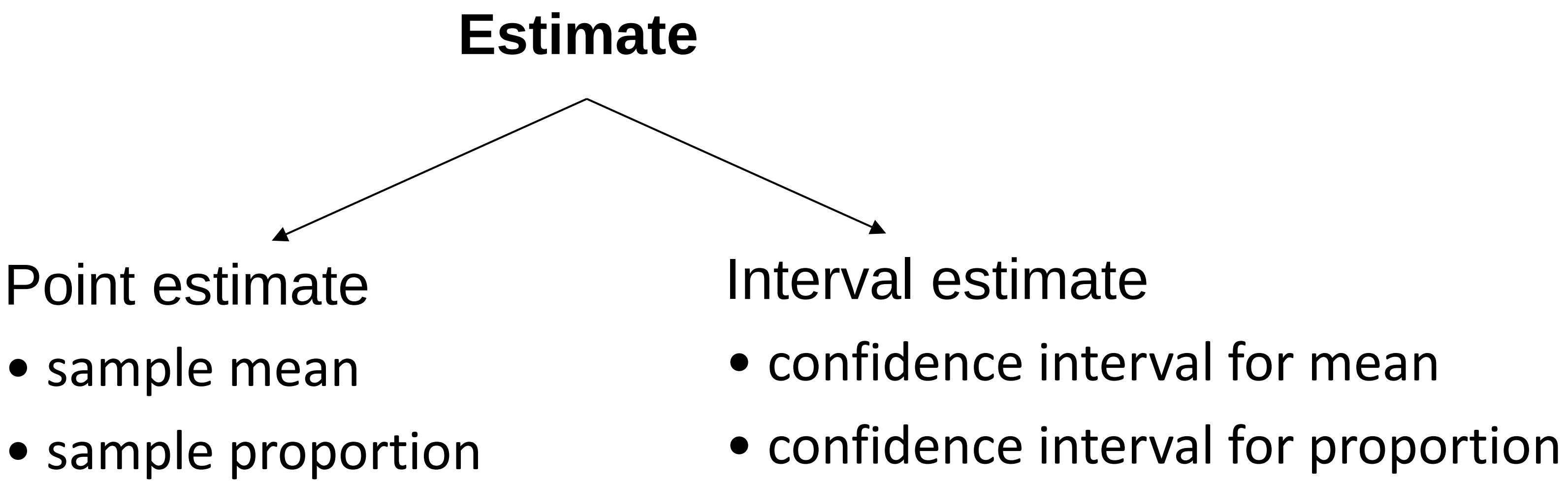
Population with mean $\mu$ = ?

A simple random sample of $n$ elements is selected from the population.

The sample data provide a value for the sample mean $\bar{x}$.

The value of $\bar{x}$ is used to make inferences about the value of $\mu$.

# Statistical Estimation



**Estimate**

Point estimate

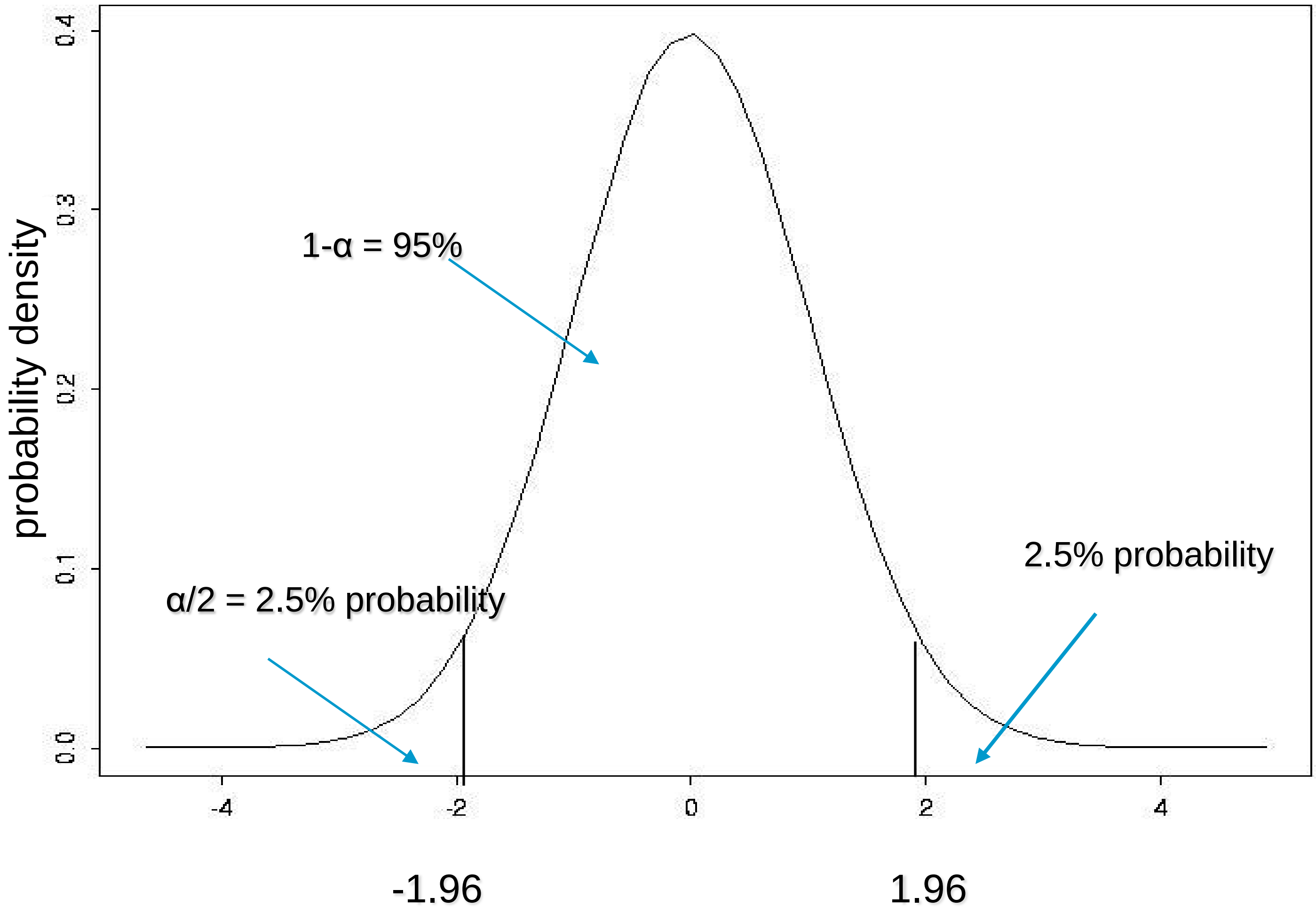
- sample mean
- sample proportion

Interval estimate

- confidence interval for mean
- confidence interval for proportion

**Point estimate is always within the interval estimate**

# Confidence Interval



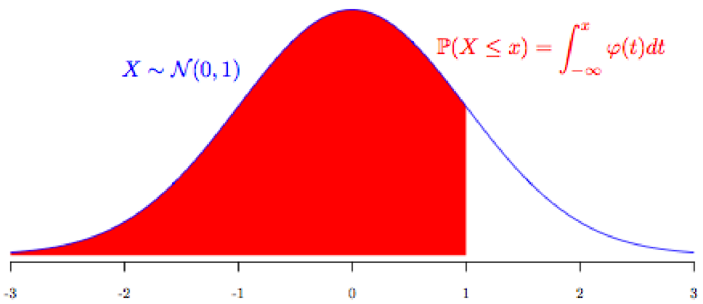
# Confidence Interval (CI)

Suppose the samples are drawn from a normal distribution. The CI provide us with a range of values that we believe, with a given level of confidence, contains a population parameter:

$$\mathrm{Pr}\ (\bar{X} - z_{(1-\frac{\alpha}{2})}\ \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{(1-\frac{\alpha}{2})} \frac{\sigma}{\sqrt{n}})$$

There is a 95% chance that your interval contains $\mu$.

$$\mathrm{Pr}(\bar{X} - 1.96\ SD < \mu < \bar{X} + 1.96\ SD) = 0.95$$



| | 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 | 0.5160 | 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 | 0.7852 |
| 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 | 0.8810 | 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 | $\boxed{0.9750}$ | 0.9756 | 0.9761 | 0.9767 |
| 2.0 | 0.9772 | 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 | 0.9864 | 0.9868 | 0.9871 | 0.9875 | 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 | 0.9925 | 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 | 0.9979 | 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 |
| 3.0 | 0.9987 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9989 | 0.9989 | 0.9989 | 0.9990 | 0.9990 |

# Example

Suppose we have a sample of $n$ = 100 persons

mean = 215, standard deviation = 20

95% CI = $\bar{X} \pm 1.96 * \sigma/\sqrt{n}$

Lower Limit: 215 – 1.96*20/10

Upper Limit: 215 + 1.96*20/10

= (211, 219)

“We are 95% confident that the interval 211-219 contains $\mu$.”

If the population standard deviation $\sigma$ is unknown, use the sample standard deviation $s$ and the *t*-distribution. If $n$ is large enough, you might also use $s$ and the standard Normal distribution.

# Effect of Sample Size

Suppose we had only 10 observations

What happens to the confidence interval?

$$\bar{X} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

For $n = 100$, $215 \pm 1.96 * (20)/\sqrt{100} \approx (211{,}219)$

For $n = 10$, $215 \pm 1.96 * (20)/\sqrt{10} \approx (203{,}227)$

Larger sample size = smaller interval

# Effect of Confidence Level

Suppose we use a 90% confidence level
What happens to the confidence interval?

$$\overline{X} \pm 1.645 * s / \sqrt{n}$$

90%: $215 \pm 1.645 * (20) / \sqrt{100} \approx (212,218)$

Lower confidence level = smaller interval
(A 99% interval would use 2.58 as multiplier and the interval would be larger)

# Effect of Standard Deviation

Suppose we had a *s* of 40 (instead of 20)

What happens to the confidence interval?

$$\bar{X} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

$$215 \pm 1.96 * (40) / \sqrt{100} \approx (207{,}223)$$

More variation = larger interval

# Estimation for Population Mean $\mu$

Point estimate: $\bar{X} = \frac{\sum X}{n}$

Estimate of variability in population $s = \sqrt{\frac{1}{n-1} \sum_i (X_i - \bar{X})^2}$

(if $\sigma$ is unknown, use $s$)

True standard deviation of sample mean $SD = \sigma / \sqrt{n}$

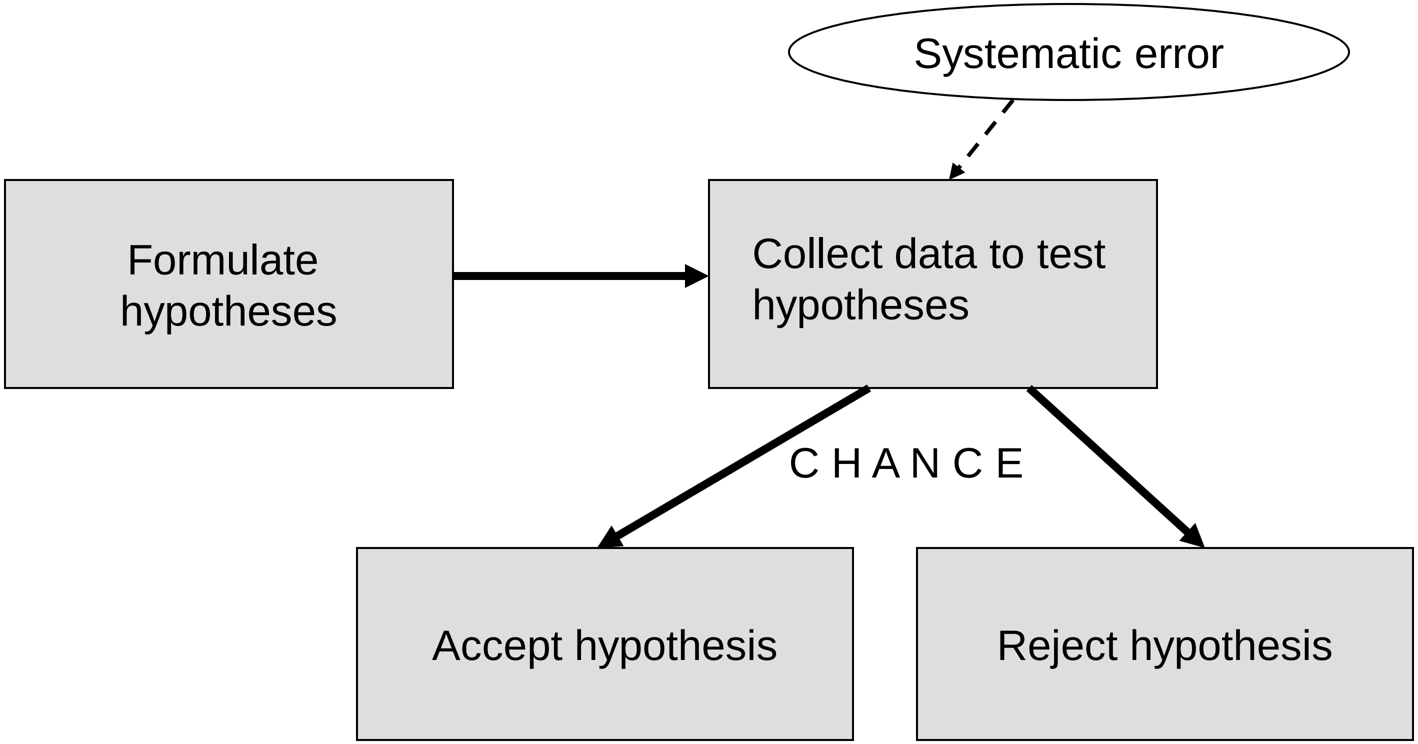
Standard error of sample mean $SE = s / \sqrt{n}$

95% confidence interval $\bar{X} \pm 1.96\, SD$

, or $\bar{X} \pm 1.96\, SE$

How does the size of the random sample impact the size of a confindence interval?

# Statistical Tests



Random error (chance) can be controlled by statistical significance or by confidence interval

# Hypothesis Testing

- State null and alternative hypothesis ($\mathrm{H}_0$ and $\mathrm{H}_1$)
  - $\mathrm{H}_0$ usually a statement no difference between groups
- Choose α level (related to confidence level)
  - Probability of falsely rejecting $\mathrm{H}_0$ (Type I error), typically 0.05 or 0.01
- Calculate test statistic, find $p$-value ($p$)
  - Measures how far data are from what you expect under null hypothesis
- State conclusion:

$p \leq \alpha$, reject $\mathrm{H}_0$

$p > \alpha$, insufficient evidence to reject $\mathrm{H}_0$

# Hypothesis Testing

Hypothesis: A statement about parameters of population or of a model ($\mu = 200$ ?)

Test: Does the data agree with the hypothesis? (sample mean $220$)
Simple random sample from a normal population
(or $n$ large enough for CLT)

$\mathrm{H}_o$: $\mu = \mu_o$
$\mathrm{H}_1$ : $\mu \neq \mu_o$ , pick $\alpha$

# Z-Test

**Problem of interest**:

- Population mean $\mu$ and population standard deviation $\sigma$ are known

Z-confidence interval: $\bar{X} \pm \mathrm{z}_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}$

Z-test: $\mathrm{z} = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$

| H$_1$ | Rejection region |
|---|---|
| $\mu\mu \neq \mu_0$ | $\lvert \mathrm{z} \rvert \geq \mathrm{z}_{1-\alpha/2}$ |
| $\mu\mu > \mu_0$ | $\mathrm{z} \geq \mathrm{z}_{1-\alpha}$ |
| $\mu\mu < \mu_0$ | $\mathrm{z} \leq \mathrm{z}_{\alpha} \quad = -\mathrm{z}_{1-\alpha}$ |

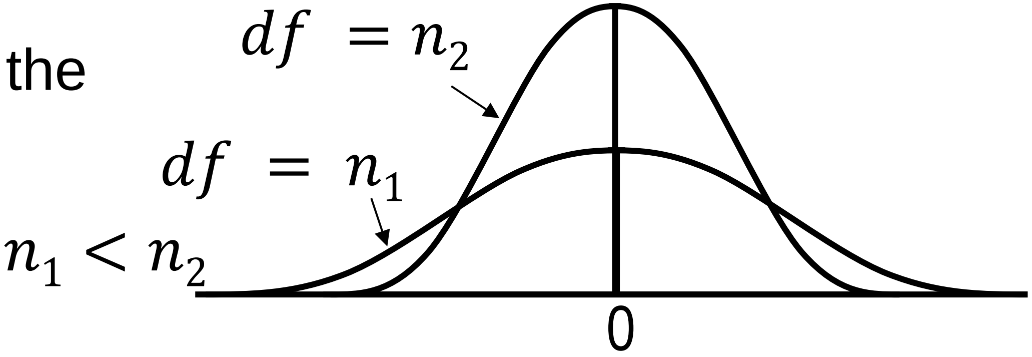
# Student t-Distribution: Test Statistic for a mean μ with unknown σ

$$t(df = n - 1) = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

**When the population is normally distributed, the statistic $t$ is *Student $t$* distributed.**

The "degrees of freedom ($df$)", a function of the sample size, determines how spread the distribution is (compared to the normal distribution)

The $t$ distribution is bell-shaped, and symmetric around zero.

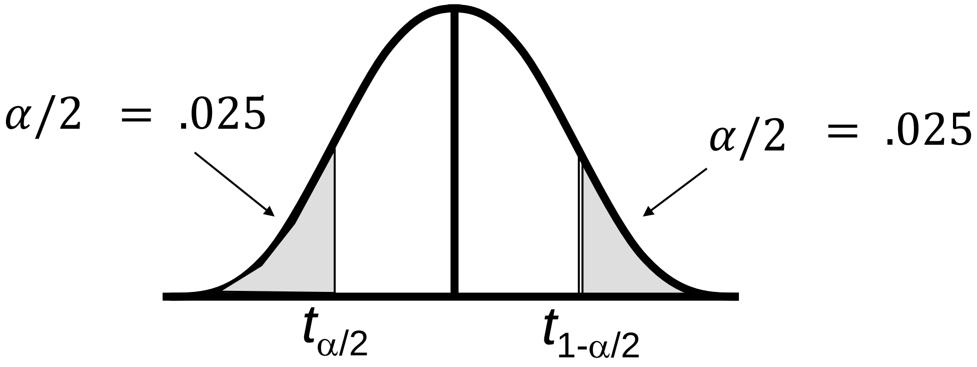


# CI and 2-Sided Tests

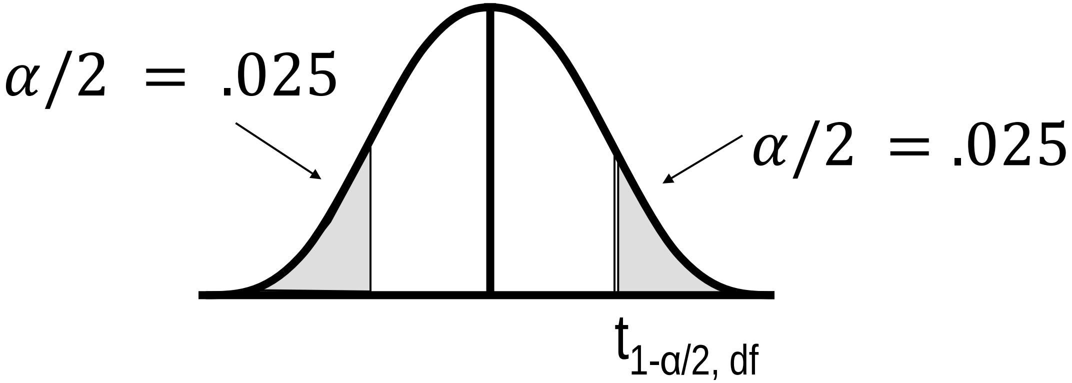
- A level $\alpha$ 2-sided test rejects $\mathrm{H}_0: \mu = \mu_0$ exactly when the value $\mu_0$ falls outside a level $1 - \alpha$ confidence interval for $\mu$.
- Calculate $1 - \alpha$ level confidence interval, then
  - if $\mu_0$ within the interval, do not reject the null hypothesis,

$$|\mathrm{t}| < \mathrm{t}_{1-\alpha/2}$$

  - otherwise, $|\mathrm{t}| \geq \mathrm{t}_{1-\alpha/2}$ => reject the null hypothesis.



# Student t-Distribution for α=0.05



| **Degrees of Freedom** | $t_{.9}$ | $t_{.95}$ | $t_{.975}$ | $t_{.99}$ | $t_{.995}$ |
|---|---|---|---|---|---|
| **1** | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| **2** | 1.886 | 2.92 | 4.303 | 6.965 | 9.925 |
| **.** | . | . | . | . | . |
| **.** | . | . | . | . | . |
| **24** | . | 1.711 | 2.064 | 2.492 | . |
| **.** | . | . | . | . | . |
| **.** | . | . | . | . | . |
| **200** | 1.286 | 1.653 | 1.972 | 2.345 | 2.601 |
| **∞** | 1.282 | 1.645 | 1.96 | 2.326 | 2.576 |

$t$-distribution critical values

# Possible Results of Tests

| | | What we decide | |
|---|---|---|---|
| | | Reject null | Fail to reject null |
| Reality | Null true | Type I Error ($\alpha$) (false positive) | Correct |
| | Null false | Correct | Type II Error ($\beta$) (false negative) |

**Type I error** - You reject the null hypothesis when the null hypothesis is actually true.

**Type II error** - You fail to reject the null hypothesis when the the alternative hypothesis is true.

# $t$-Tests

Formula is slightly different for each:

- *Single sample:*
  - tests whether a sample mean is significantly different from a pre-existing value
- *Paired samples:*
  - tests the relationship between 2 linked samples, e.g. means obtained in 2 conditions by a single group of participants
- *Independent samples:*
  - tests the relationship between 2 independent populations

# The Paired $t$ -Test with 2 Paired Samples

Null hypothesis: $H_0: \mu_d = \mu_1 - \mu_2 = \Delta_0$

Test statistic: $t = \frac{\bar{d} - \Delta_0}{s/\sqrt{n}}$

| $H_1$ | Rejection region |
|---|---|
| $\mu_d \neq \Delta_0$ | $\lvert t \rvert \geq t_{1-\alpha/2,\, n-1}$ |
| $\mu_d > \Delta_0$ | $t \geq t_{1-\alpha,\, n-1}$ |
| $\mu_d < \Delta_0$ | $t \leq t_{\alpha,\, n-1} = -t_{1-\alpha,\, n-1}$ |

Observations are dependent, e.g., pre and post test, left and right eyes, brother-sister pairs

# The Paired $t$ -Test with 2 Paired Samples

Subjects: random sample of 25 students from TUM

Mean grades of the students on two subsequent exams $A$ and $B$

Is there a significant difference between the two exams?

Null Hypothesis: $E(A) = E(B)$

Answer can be given based on significance testing

| *No.* | *A* | *B* | $d$=*A*-*B* |
|---|---|---|---|
| 1 | 3.7 | 3.5 | 0.2 |
| 2 | 2.2 | 2.3 | -0.1 |
| ... | | | |
| 25 | 4.8 | 4.4 | 0.4 |

$\bar{d} = 0.093$

$s = 0.150$

$n = 25$

$s/\sqrt{n} = 0.03$

$t_{0.975;24} = 2.064$

$$t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{0.093}{0.03} = 3.1$$

$$p = \Pr\{|\mathrm{t}| > 3.1|DF = 24\} = 0.005$$

# The $p$-Value

The $p$-value describes the probability of having $t = 3.1$ (or larger), given the null hypothesis. The smaller the $p$-value, the more unlikely it is to observe the corresponding sample value (or more extreme) by chance under $\mathrm{H}_0$.

```
> # R code
> x = c(3, 0, 5, 2, 5, 5, 5, 4, 4, 5)
> y = c(2, 1, 4, 1, 4, 3, 3, 2, 3, 5)
> t.test(x,y,alt="two.sided", paired=TRUE)

           Paired t-test

data:  x and y
t = 3.3541, df = 9, p-value = 0.008468
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.325555 1.674445
sample estimates:
mean of the differences
```

# Independent Samples

2 independent samples (possibly different size and variance):

Does the amount of credit card debt differ between households in rural areas compared to households in urban areas?

Population 1
All Rural Households
$m_1$

Population 2
All Urban Households
$m_2$

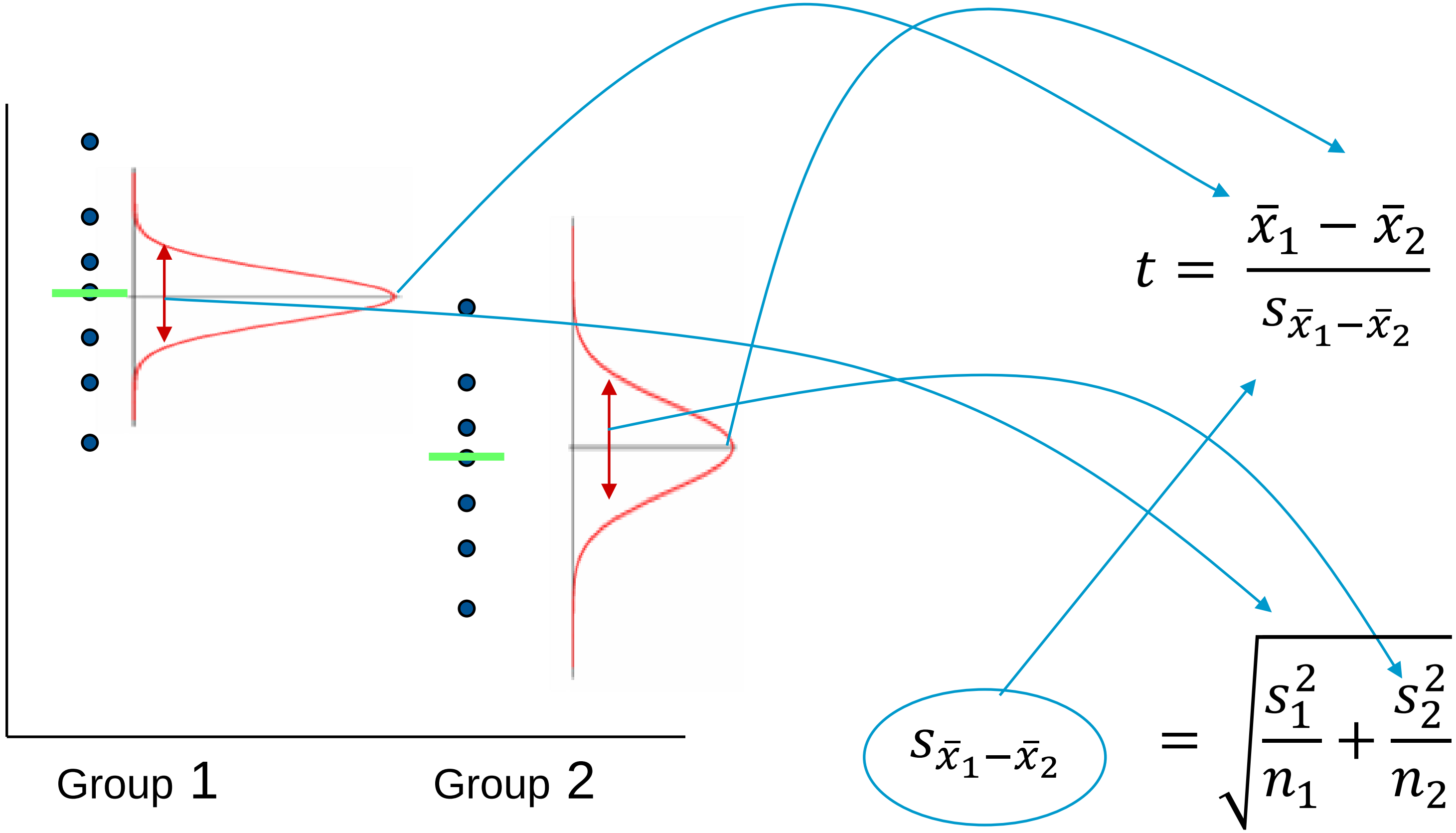
Null Hypothesis: $\mathrm{H}_0 : m_1 = m_2$

Alternate Hypothesis: $\mathrm{H}_1 : m_1 \neq m_2$

# Independent Two-Sample $t$ -Test (Welch’s $t$ -Test)

Two-sample unpaired $t$-test with (un)equal sample sizes, assuming unequal variance

Under $\mathrm{H_o}$ *t* follows a t-distribution with $\dfrac{(s_1^{\,2}/n_1 + s_2^{\,2}/n_2)^2}{(s_1^{\,2}/n_1)^2/(n_1-1) + (s_2^{\,2}/n_2)^2/(n_2-1)}$ degrees of freedom (df)



$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

# Independent Two-Sample $t$ –Test: Example

| Group 1 | Group 2 |
|---|---|
| 21 | 22 |
| 19 | 25 |
| 18 | 27 |
| 18 | 24 |
| 23 | 26 |
| 17 | 24 |
| 19 | 28 |
| 16 | 26 |
| 21 | 30 |
| 18 | 28 |
| $\bar{x}_1 = 19$ | $\bar{x}_2 = 26$ |
| $s_1 = \sqrt{40/9}$ | $s_2 = \sqrt{50/9}$ |

$df \; = \; 18$ (rounded to integer)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{19 - 26}{1} = -7$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{40/9}{10} + \frac{50/9}{10}}$$
$$= 1$$

$$t_{(0.975,18)} = 2.101$$

$$|t| \geq t_{(0.975,18)}$$

→ Reject $\mathrm{H}_0$ $(\mu_1 - \mu_2 = 0)$

# Selected Statistical Tests

- **Parametric Tests**
  - The family of $t$-tests
    - Compares two sample means or tests a single sample mean
  - F-test
    - Compares the equivalence of variances of two samples

- **Non-parametric Tests**
  - Wilcoxon signed-rank test for 2 *paired* i.i.d samples.
  - Mann-Whitney-U test is used for 2 *independent* i.i.d samples
  - Kruskal-Wallis-Test for several i.i.d non-normally distributed samples

- **Tests of the Probability Distribution**
  - Kolmogorov-Smirnov and Chi-square test
    - used to determine whether two underlying probability distributions differ, or whether an underlying probability distribution differs from a hypothesized distribution

Please explain the role of confidence intervals in a single-sample t-test.

# Linear Regression

- **Regressions identify relationships between dependent and independent variables**
  - Is there an association between the two variables
  - Estimation of impact of an independent variable
  - Formulation of the relation in a functional form
  - Used for numerical prediction and time series forecasting

- **Regression as an established statistical technique:**
  - Sir Francis Galton (1822-1911) studied the relationship between a father’s height and the son’s height

# Terminology

- Data streams $X$ and $Y$, forming the measurement tuples $(x_1, y_1), \dots\,.\,, (x_n, y_n)$
- $x_i$ is the predictor (regressor, covariate, feature, independent variable)
- $y_i$ is the response (dependent variable, outcome)
- Denote the *regression function* by: $\eta(x) = E\,(Y\,|x)$
- The linear regression model assumes a specific linear form

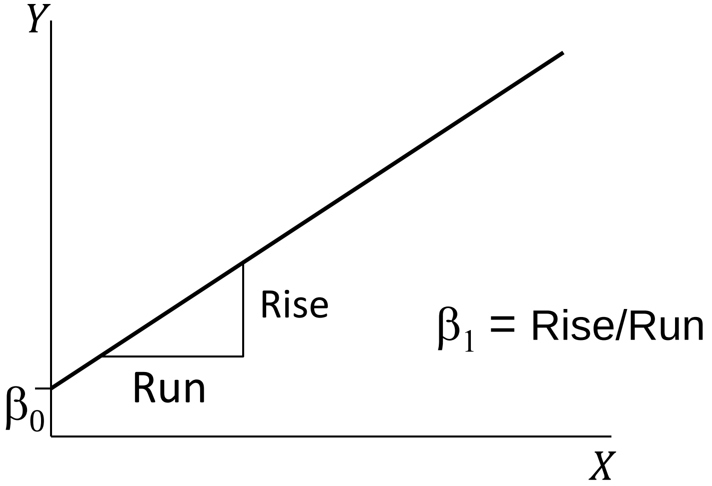
# The Simple Linear Regression Model

- Linear regression is a statistical tool for numerical predictions
- The first order linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$\beta_0$ and $\beta_1$ are unknown, therefore, are estimated from the data

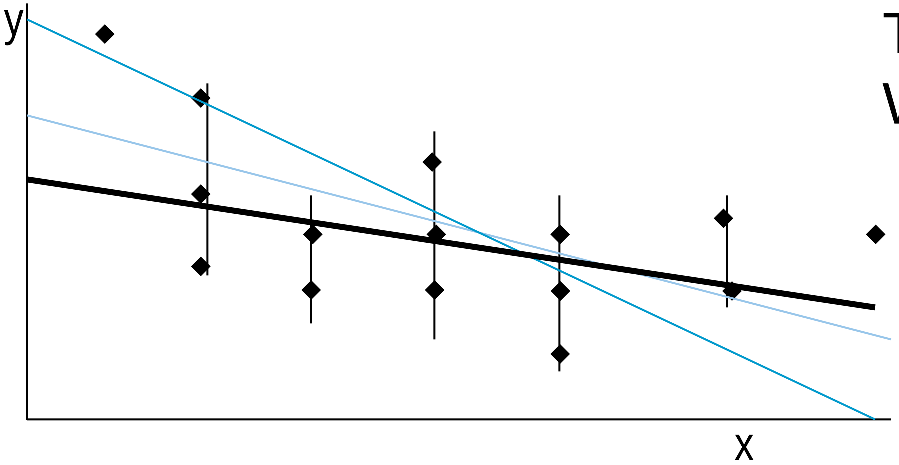
$Y$ = response variable
$X$ = predictor variable
$\beta_0$= y-axis intercept
$\beta_1$ = slope of the line
$\varepsilon$ = random error term (residual)



$\beta_1$ = Rise/Run

# Estimating the Coefficients

- Coefficients are random variables
- (Ordinary Least Squares) estimates are determined by
  - drawing a sample from the population of interest
  - calculating sample statistics
  - producing a straight line that cuts into the data



The question is:
Which straight line fits best?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

# OLS Estimators

- Ordinary Least Squares (OLS) approach:
  - Minimize the sum of squared residuals (aka. loss function)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

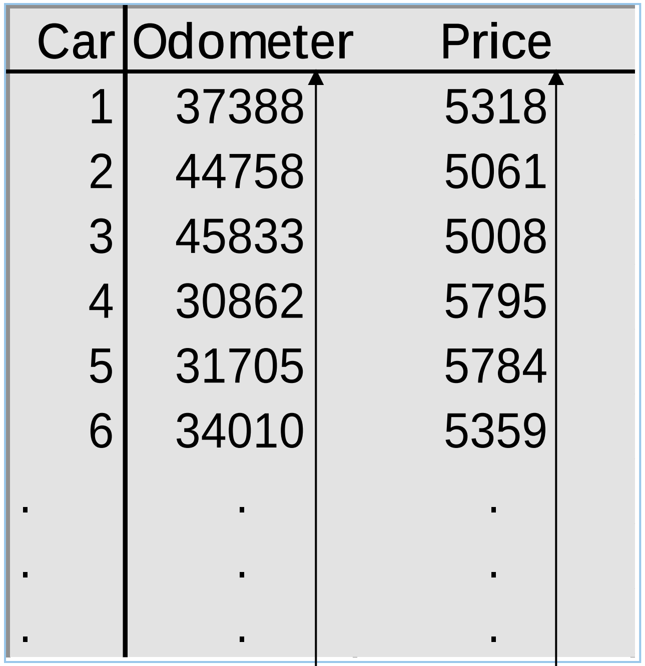
$$\min \sum_{\mathrm{i}} e_{\mathrm{i}}^2 = \min \sum_{\mathrm{i}} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \frac{Cov(x, y)}{Var(x)}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

# Example

- A car dealer wants to find the relationship between the odometer reading and the selling price of used cars.
- A random sample of 100 cars is selected, and the data recorded.
- Find the regression line.



| Car | Odometer | Price |
|---|---|---|
| 1 | 37388 | 5318 |
| 2 | 44758 | 5061 |
| 3 | 45833 | 5008 |
| 4 | 30862 | 5795 |
| 5 | 31705 | 5784 |
| 6 | 34010 | 5359 |
| . | . | . |
| . | . | . |
| . | . | . |

Independent/predictor variable $x$

Dependent/respond variable $y$

# Solving a Simple Regression

- To calculate $\beta_0$ and $\beta_1$ we can calculate several statistics first:

$$\bar{x} = 36009.45; \qquad s_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = 43{,}528{,}688$$

$$\bar{y} = 5411.41; \qquad \operatorname{cov}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = -1{,}356{,}256$$

where $n = 100$:

$$\hat{\beta}_1 = \frac{\operatorname{cov}(X,Y)}{s_x^2} = \frac{-1{,}356{,}256}{43{,}528{,}688} = -.0312$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 5411.41 - (-.0312)(36{,}009.45) = 6{,}533$$

$$\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 6{,}533 - 0.0312x}$$

# Residual Sum of Squares (RSS)

- This is the sum of squared differences between the points and the regression line
- It can serve as a measure of how well the line fits the data (fits well, if statistic is small)
- An unbiased estimator of the RSS of the population is given by

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

# Total Deviation

- The Total Sum of Squares (TSS) is the sum of the Explained Sum of Squares (ESS) and the RSS.

$$\boxed{\sum(y-\bar{y})^2 \qquad = \sum(\hat{y}-\bar{y})^2 \qquad + \sum(y-\hat{y})^2}$$

TSS = ESS + RSS

Total deviation = explained deviation + unexplained deviation

# Coefficient of Determination

- $R^2$ measures the proportion of the variation in $y$ that is explained by the variation in $x$

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \ , TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = ESS + RSS$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$

- $R^2$ takes on any value between zero and one
  - $R^2 = 1$: Perfect match between the line and data points
  - $R^2 = 0$: There is no linear relationship between x and y

# Testing the Coefficients

- Test the significance of the linear relationship

  $\mathrm{H}_0: \beta_1 = 0$

  $\mathrm{H}_1: \beta_1 \neq 0$
- The test statistic is

$$t = \frac{\hat{\beta}_1}{S\mathrm{E}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{\frac{\mathrm{RSS}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \frac{1}{n-2}}}$$

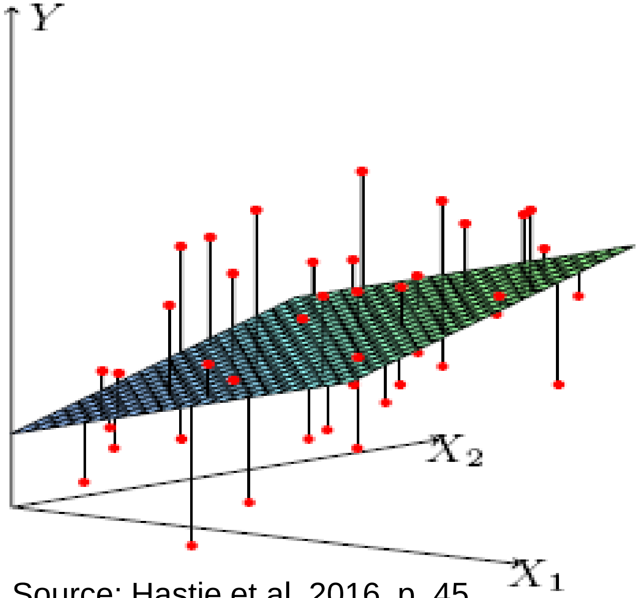
Variance of $\hat{\beta}_1$

- If $SE(\hat{\beta}_1)$ is large, then $\hat{\beta}_1$must be large to reject $\mathrm{H}_0$
- $SE(\hat{\beta}_1)$ is smaller, if the $x_i$ are more spread out
- If the error variable is normally distributed, the statistic is a Student $t$ –distribution with $n - 2$ degrees of freedom (if $n$ is large, draw on the CLT)
- Reject $\mathrm{H}_0$, if: $t < t_{\alpha/2}$ or $t > t_{1-\alpha/2}$

# The Multiple Linear Regression Model

- A $p$-variable regression model can be expressed as a series of equations
- Equations condensed into a matrix form, give the general linear model
- $\beta$ coefficients are known as partial regression coefficients
- $X_1, X_2$, for example,
  - $X_1$='years of experience'
  - $X_2$='age'
  - Y='salary'
- Estimated equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 = \mathbf{X}\hat{\beta}$$



Source: Hastie et al. 2016, p. 45

# Matrix Notation

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

| $y$ | $\boldsymbol{X}$ | β | + ε |
|---|---|---|---|
| $(n \times 1)$ | $(n \times (p+1))$ | $((p+1) \times 1)$ | $(n \times 1)$ |

# OLS Estimation

- Sample-based counter part to population regression model:

$$y = \mathbf{X}\beta + \varepsilon$$
$$y = \mathbf{X}\hat{\beta} + e$$

- OLS requires choosing values of the estimated coefficients, such that Residual Sum of Squares (RSS) is as small as possible for the sample

$$RSS = e^T e = (y - \mathbf{X}\hat{\beta})^T (y - \mathbf{X}\hat{\beta})$$

- Need to differentiate with respect to the unknown coefficients

# Least Squares Estimation

$\mathbf{X}$ is $n{\times}(p+1)$, $y$ is the vector of outputs

RSS(β) = $(y - \mathbf{X}\beta)^T\,(y - \mathbf{X}\beta)$

If $\mathbf{X}$ is full rank, then $\mathbf{X}^T\mathbf{X}$ is positive definite

$$RSS = (y^T y - 2\beta^T\,\mathbf{X}^T\mathrm{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta)$$

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T\mathrm{y} + 2\mathbf{X}^T\mathbf{X}\beta = 0$$ First-order condition

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

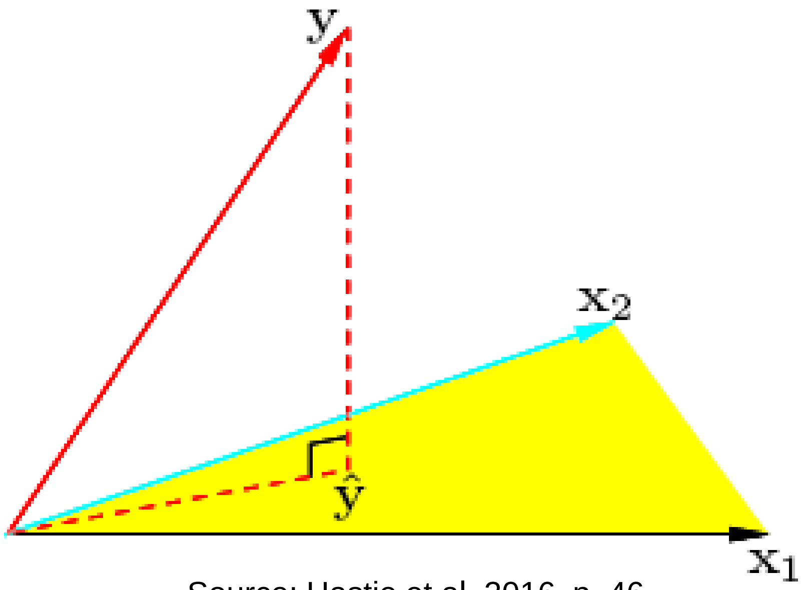
$$\hat{y} = \underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{\text{“Hat” or projection matrix } H} y$$

# Geometrical Representation

- Least square estimates in $\mathbb{R}^n$
- Minimize RSS(β)=$||y - \mathbf{X}\beta||^2$, s.t. residual vector $y$ – $\hat{y}$ is orthogonal to this subspace.

**Definition (Projection):**
The set $C \subset \mathbb{R}^n$ is non-empty, closed and convex. For a fixed $y \in \mathbb{R}^n$ we search a point $\hat{y} \in C$, with the smallest distance to $y$ (wrt. the Euclidean norm), i.e. we solve the minimization problem

$$P_C(y) = \min_{\hat{y} \in C} ||y - \hat{y}||^2$$



Source: Hastie et al. 2016, p. 46

# Example

| | | | | | |
|---|---|---|---|---|---|
| $y$: | 2.6 | 1.6 | 4.0 | 3.0 | 4.9 |
| $x$: | 1.2 | 3.0 | 4.5 | 5.8 | 7.2 |

$$y = \mathbf{X}\hat{\beta} + e$$

$$\begin{pmatrix} 2.6 \\ 1.6 \\ 4.0 \\ 3.0 \\ 4.9 \end{pmatrix} = \begin{pmatrix} 1 & 1.2 \\ 1 & 3.0 \\ 1 & 4.5 \\ 1 & 5.8 \\ 1 & 7.2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{pmatrix}$$

$$\hat{\beta} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathrm{y}$$

$$\left( \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1.2 & 3.0 & 4.5 & 5.8 & 7.2 \end{pmatrix} \begin{pmatrix} 1 & 1.2 \\ 1 & 3.0 \\ 1 & 4.5 \\ 1 & 5.8 \\ 1 & 7.2 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1.2 & 3.0 & 4.5 & 5.8 & 7.2 \end{pmatrix} \begin{pmatrix} 2.6 \\ 1.6 \\ 4.0 \\ 3.0 \\ 4.9 \end{pmatrix} =$$

$$\begin{pmatrix} 5 & 21.7 \\ 21.7 & 116.17 \end{pmatrix}^{-1} \begin{pmatrix} 16.1 \\ 78.6 \end{pmatrix} = \begin{pmatrix} 1.0565 & -0.1973 \\ -0.1973 & 0.0455 \end{pmatrix} \begin{pmatrix} 16.1 \\ 78.6 \end{pmatrix} = \begin{pmatrix} 1.498 \\ 0.397 \end{pmatrix}$$

# Check Results in R

```
> y <- c(2.6, 1.6, 4.0, 3.0, 4.9)
> x <- c(1.2, 3.0, 4.5, 5.8, 7.2)
> mod <- lm(y ~ x)
> summary(mod)

Call:
lm(formula = y ~ x)

Residuals:
      1       2       3       4       5
 0.6259 -1.0883  0.7165 -0.7993  0.5452

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4980     1.0322   1.451    0.243
x             0.3968     0.2142   1.853    0.161

Residual standard error: 1.004 on 3 degrees of freedom
Multiple R-Squared: 0.5336,     Adjusted R-squared: 0.3782
F-statistic: 3.433 on 1 and 3 DF,  p-value: 0.1610
```

1. check coefficients
2. check significance
3. check coefficient of determination

# Selected Statistics

**Adjusted $R^2$**

- It represents the proportion of variability of $y$ explained by $X$
  $R^2$ is adjusted so that models with a different number of variables can be compared

$$\bar{R}^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

**The F-test**

- Significant F indicates a linear relationship between $y$ and at least one of the xs: $\mathrm{H}_0\!: \ \beta_1 = \beta_2 \ldots \beta_p = 0$

**The *t*-test of each partial regression coefficient**

- Significant t indicates that the variable in question influences the response variable while controlling for other explanatory variables

# Model Specification

In regression analysis the specification is the process of developing a regression model.

- This process consists of selecting an appropriate functional form for the model and choosing which variables to include.
- The model might include irrelevant variables or omit relevant variables

Non-linear models are challenging, but some nonlinear regression problems can be linearized.

- Dummy variables for discrete variables (e.g. 0/1 for gender)
- Quadratic models: $y = \beta_0 + \beta_1 x_1 + \beta_2 {x_2}^2 + \varepsilon$ use $z_2 = {x_2}^2$
- Models with interaction terms $y = \beta_0 + \beta_1 x_1 x_2$ use $z_1 = x_1 x_2$
- Exponential terms $y = \alpha x^\beta \varepsilon$ can be transformed using the logarithm to

$$\ln(y) = \ln(\alpha) + \beta \ln(x) + \ln(\varepsilon)$$

# Subset Selection

- Setting: Possibly a large set of predictor variables, some irrelevant
- Goal: Fit a parsimonious model that explains variation in $Y$ with a small set of predictors
  - Aka. subset selection or feature selection problem
- Automated procedures:
  - Best subset (among all exponentially many, computationally expensive)
  - Backward elimination (top down approach)
  - Forward selection (bottom up approach)
  - Stepwise regression (combines forward/backward)

- More in the context of the class on dimensionality reduction
  - Subset selection vs. shrinkage methods

# Example: Backward Elimination

- Select a significance level to stay in the model (generally 0.05 is too low, causing too many variables to be removed)
- Fit the full model with all possible predictors
- Consider the predictor with lowest $t$-statistic (highest $p$-value).
  - If $p >$ sign. level, remove the predictor and fit model without this variable (must re-fit model here because partial regression coefficients change)
  - If $p \leq$ sign. level, stop and keep current model
- Continue until all predictors have $p$-values below sign. level

- Forward selection is similar: predictors with lowest $p$-value are added until none is left with $p >$ sign. level.

Please explain the term „ordinary least squares“ estimator and how it solves a convex optimization problem.