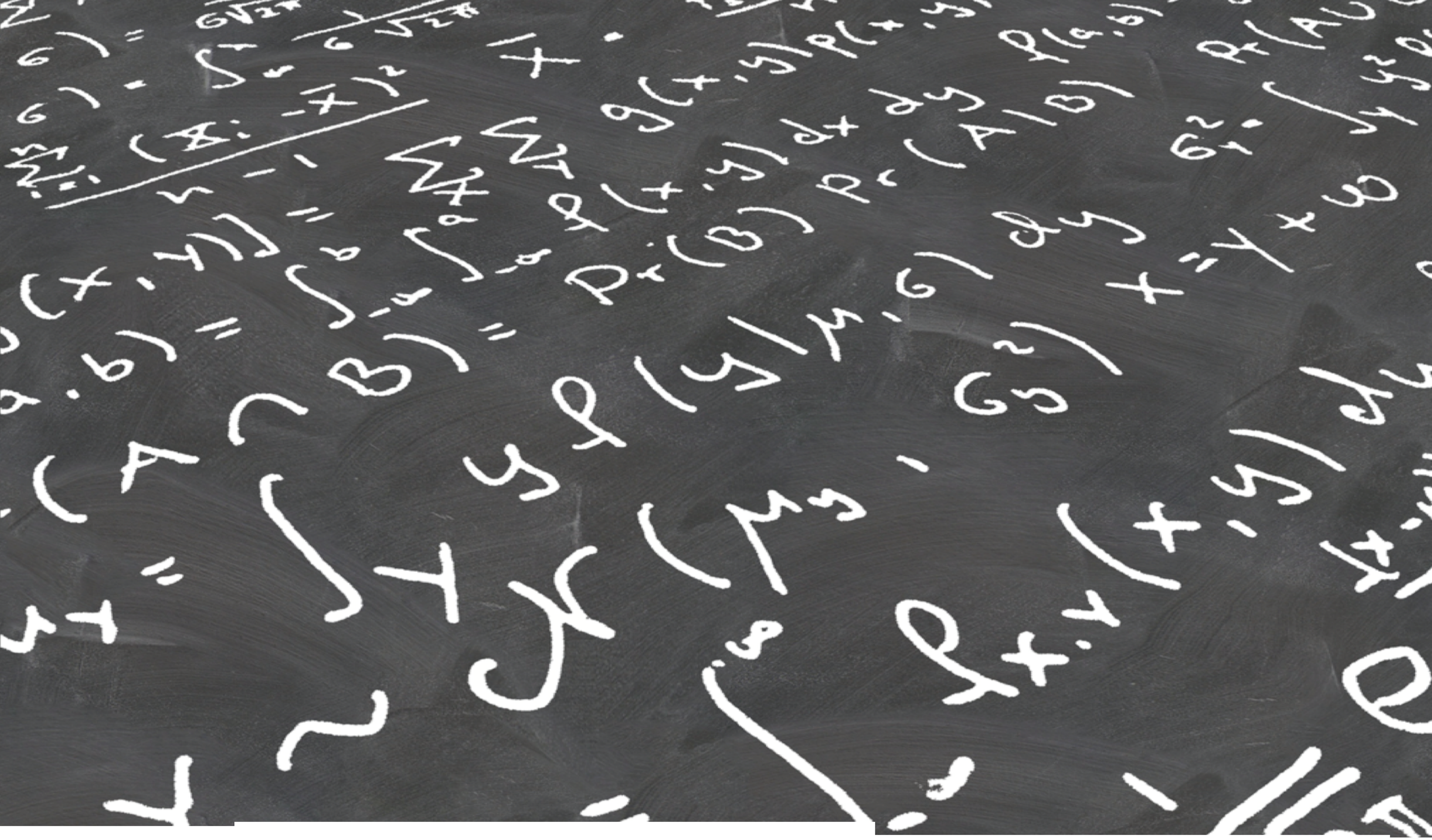


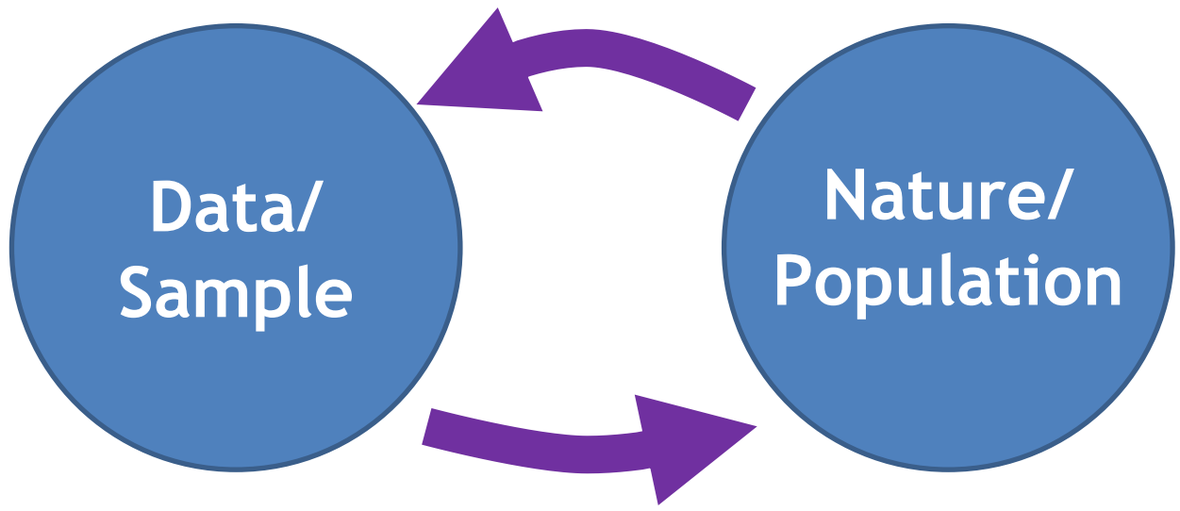
14. Review Statistics

Gunvor Elisabeth Kirkelund
Lars Mandrup
Slides and material provided in parts by
Henrik Pedersen

Introduction to Statistics

Probability theory

Given the cause (population), what should the data (sample) look like?



Statistics

Given the data (sample), what caused them (population)?

- Testing a hypotesis
- Estimating means and variances

- If we don't know better: We assume data are normally distributed

Estimator

Estimator:

- An estimator $\hat{\theta}(X)$ is a statistic used to estimate the unknown parameter θ of a random sample X.
- An estimator is unbiased if $E[\hat{\theta}] = \theta$.

Unbiased estimators:

- The sample mean: $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Degrees of freedom

- The sample variance: $\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Degrees of freedom

Statistical Model

Statistical model:

- A random sample and its pdf, $f_X(x;\theta)$, where θ is the parameter(s) of the pdf.

- Because of the Central Limit Theorem (CLT) we often can use the normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2 as statistical model for the sample mean $\bar{X}$

Central Limit Theorem

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of a random variable *X* with mean μ and variance σ^2.
- Then, as $n \to \infty$, the sample mean ($\bar{X}$) becomes normally distributed with a mean that is equal to the population mean (μ) and a variance that is scaled by 1/*n*:

Sample mean $\bar{X} \sim N(\mu, \sigma^2/n)$

- Note that *X* can have any distribution, i.e., it is *not* required to be normally distributed.

- Although the exact number is subject of debate, it is common practice to require that the number of samples (*n*) should be 30 or larger in order to apply the CLT:

$$n \geq 30$$

...most important number in statistics

Hypothesis

- Definition – Null hypothesis (H_0)
  - The statement being tested in a test of statistical significance is called the **null hypothesis**. The test of significance is designed to assess the strength of the evidence against the null hypothesis.
  - Usually, the null hypothesis is a statement of 'no effect‘, 'no difference‘ or ‘no relation' between the phenomena whose relation is under investigation.
- Definition – Alternative hypothesis (H_1)
  - The statement that is hoped or expected to be true instead of the null hypothesis is the **alternative hypothesis**
  - The alternative hypothesis, as the name suggests, is the alternative to the null hypothesis: it states that there is some 'effect/difference’ or some ‘kind of relation'.

Important!

- One cannot “prove” a null hypothesis, one can only test how close it is to being true.
- Therefore, we never say that we *accept* the null hypothesis, but that we either ***reject it*** or ***fail to reject it***.

Test Statistics, p-value, significance level and confidence interval

- Test statistics:
  - A random variable that summarized a data-set by reducing the data to one value that can be used to perform the hypothesis test.

  - Known μ and σ^2: $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ z-statistic

  - Known μ and unknown σ^2: $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t(n-1)$ Students t

- p-value: $p - value = Pr(Worse\ result\ than\ X | H_0)$

- Significance level **α**: Limit for the p-value to reject the null hypothesis.
  Typical we use **α = 0,05 = 5%.**

- Confidence interval: $[\theta_-; \theta_+]$ such that $Pr(\theta_- \leq \theta \leq \theta_+) = 1 - \alpha$
  Typical the 95% confidence interval.

TEST CATALOG FOR THE MEAN (KNOWN VARIANCE)

- **Statistical model:**
- $X_1, X_2, \ldots, X_n$ are i.i.d. samples of a random variable *X* with mean μ and variance σ^2.
- Parameter estimate:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

- Where the observation is $\bar{x}$ = 'the average of *n* samples drawn from *X*'s distribution'.
- NOTE: The statistical model is only true if *n* is sufficiently large ($n \geq 30$) or if the samples are drawn from a normal population with mean μ and variance σ^2.

- **Hypothesis test (two-tailed):**
- $H_0: \mu = \mu_0$
- $H_1: \mu \neq \mu_0$
- Test size: $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

- **95% confidence interval:**
- $\mu_{\_} = \bar{x} - 1.96 \cdot \sigma/\sqrt{n}$
- $\mu_{+} = \bar{x} + 1.96 \cdot \sigma/\sqrt{n}$

TEST CATALOG FOR THE MEAN (UNKNOWN VARIANCE)

- **Statistical model:**
- $X_1, X_2, \ldots, X_n$ are i.i.d. samples of a random variable X with mean μ and variance σ^2.
- Parameter estimates:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

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

- Where the observation is $\bar{x}$ = 'the average of n samples drawn from X's distribution'.
- NOTE: The statistical model is only true if n is sufficiently large ($n \geq 30$) or if the samples are drawn from a normal population with mean μ and variance σ^2.

- **Hypothesis test (two-tailed):**
- $H_0: \mu = \mu_0$
- $H_1: \mu \neq \mu_0$
- Test size: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim t(n-1)$
- Approximate p-value: $2 \cdot \left|1 - t_{cdf}(|t|)\right|$

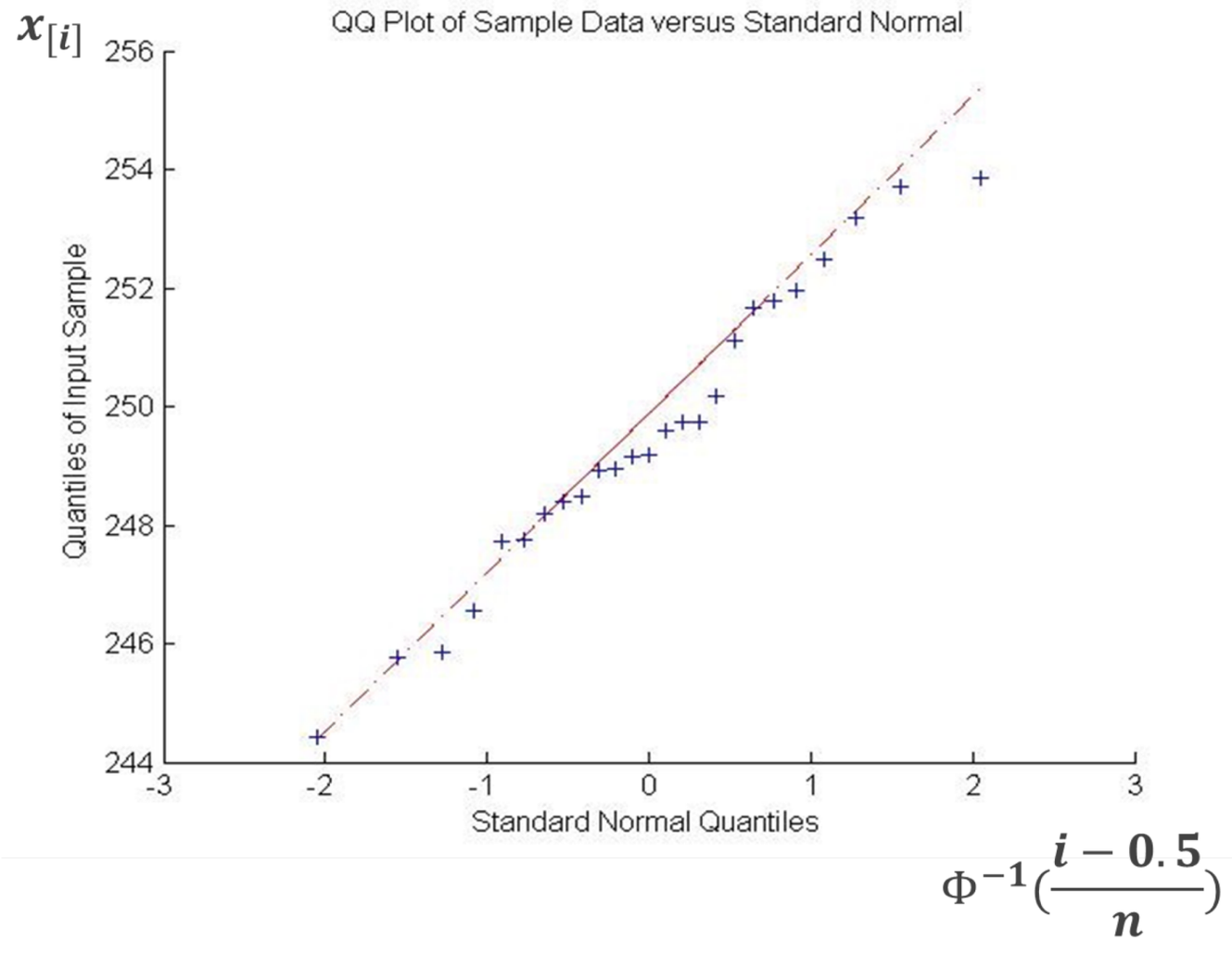
- **95% confidence interval:**
- $\mu_- = \bar{x} - t_0 \cdot s/\sqrt{n}$
- $\mu_+ = \bar{x} + t_0 \cdot s/\sqrt{n}$

- where t0 = tinv(1-0.05/2,n-1)

Q-Q plot

- Q-Q plot – a method to check whether the data are normally distributed

- Sort the samples in ascending order:
  $x_1, x_2, \dots, x_n$
- Plot $x_{[i]}$ vs. $\Phi^{-1}(\frac{i-0.5}{n})$
- If the data are consistent with a sample from a normal distribution it should result in a straight line.



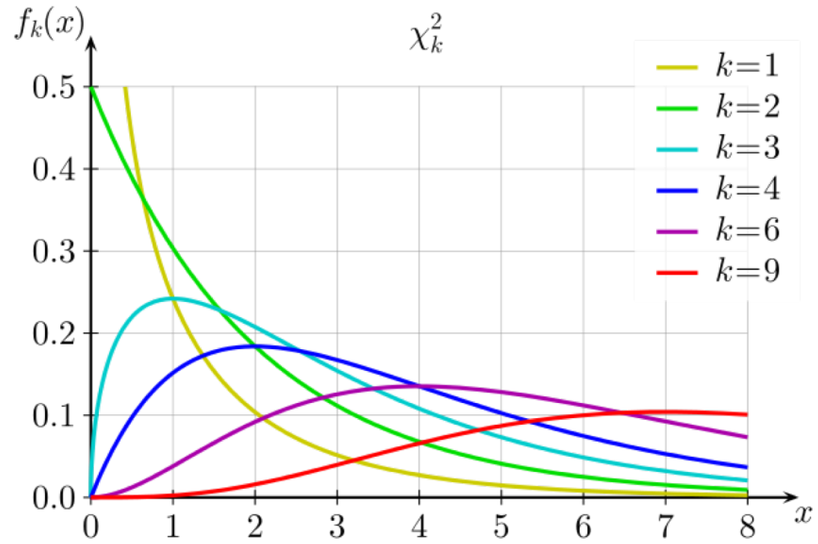
Chi-Square Distribution

- If we have a set of i.i.d. data $X_1, X_2, \ldots, X_n$ distributed according to:

$$X_i \sim \mathcal{N}(0,1)$$

- Then we have that: $\mathrm{Q} = \sum_{i=1}^{n} X_i^2$

  is χ_k^2 distributed with k degrees of freedom.



- χ^2 -test for independence: $X^2 = \sum_{i=1}^{n} \frac{(observed - expected)^2}{Expected}$

  How well does the observed data fits the expected values

- χ^2 -test for variance: Test statistics $T = \frac{N-1}{s/\sigma_0^2}$

  Hypothesis test of the variance σ_0^2

The Binomial Distribution

- We have n repeated trials.
- Each trial has two possible outcomes *Bernoulli Event*
  - **Success** — probability p
  - **Failure** — probability 1-p

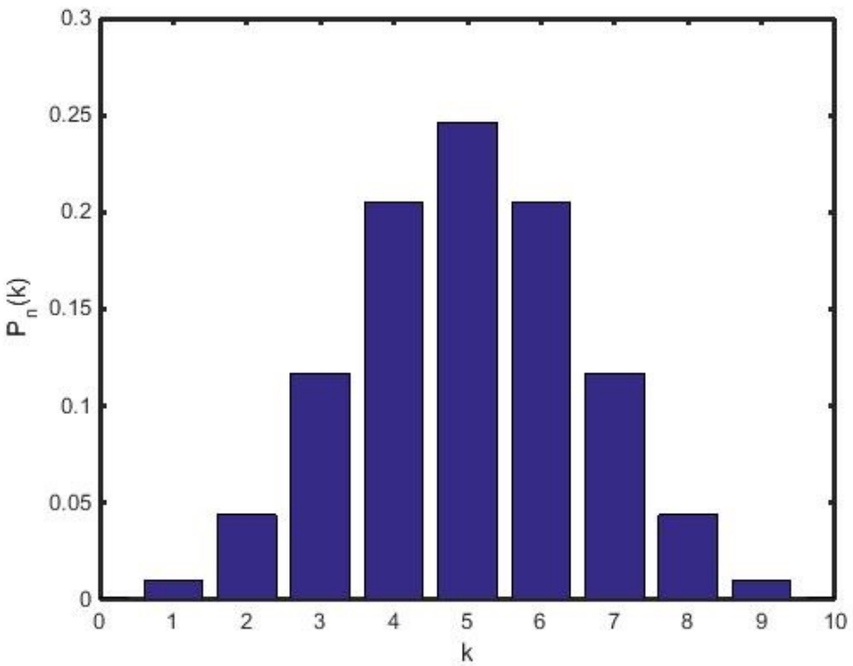
- We write the mass function as:

  X =Number of successes in n trials

$$Pr(X = k) = f(k|n,p)$$

$$f(k|n,p) = \frac{n!}{k!\,(n-k)!}p^k(1-p)^{n-k}$$

$$= \binom{n}{k}p^k(1-p)^{n-k}$$



Test catalog for the Binomial Distribution

- **Statistical model:**
- $X \sim binomial(n, p)$
- Parameter estimate: $\hat{p} = x/n$
- Where the observation is *x* = 'number of successes out of *n* trials'

- **Hypothesis test (two-tailed):**
- $H_0: p = p_0$
- $H_1: p \neq p_0$
- Test size: $z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} \sim N(0,1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

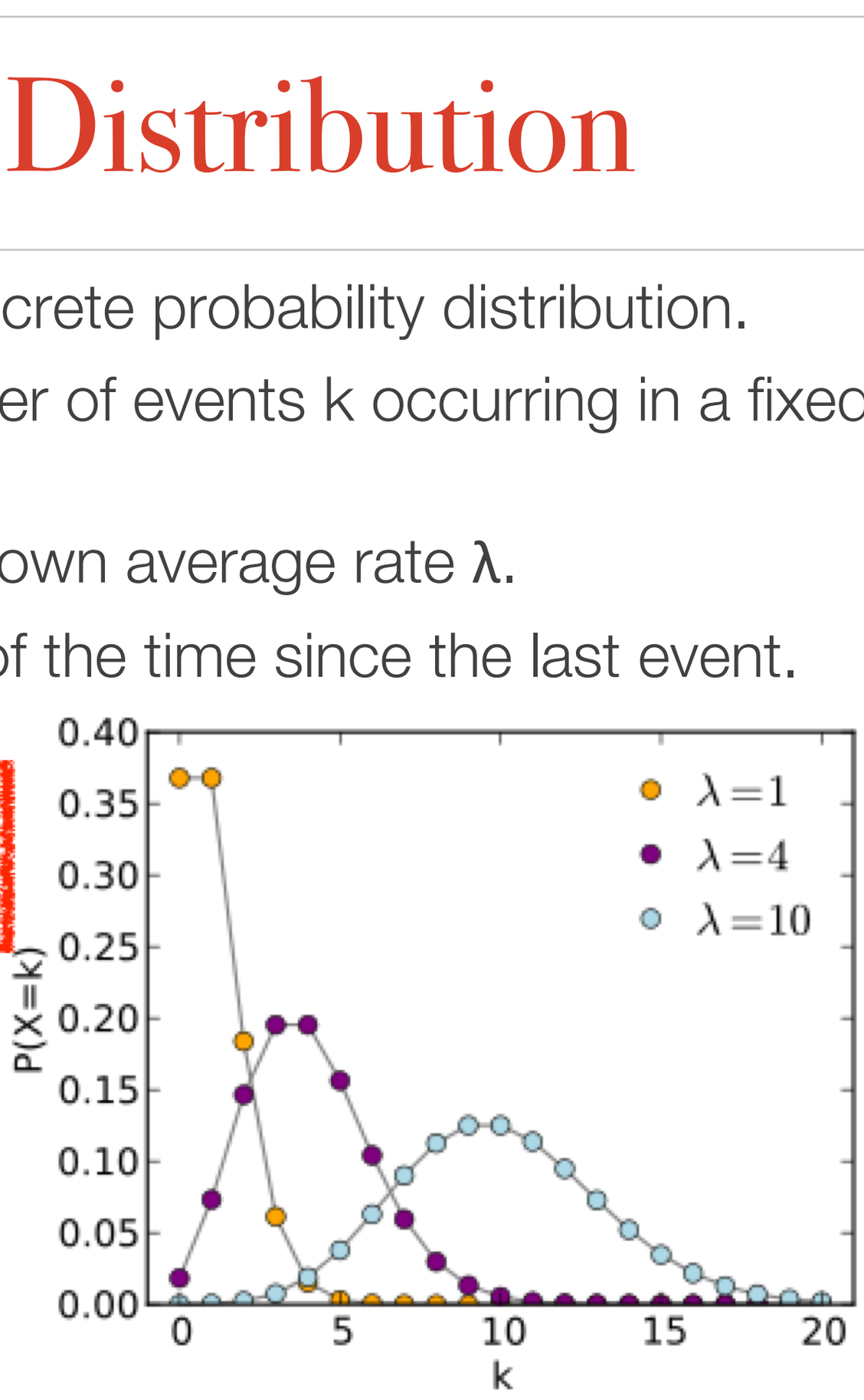
- **95% confidence interval:**
- $p_- = \frac{1}{n+1.96^2}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1.96^2}{4}}\right]$
- $p_+ = \frac{1}{n+1.96^2}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1.96^2}{4}}\right]$

The Poisson Distribution

- The Poisson distribution is a discrete probability distribution.
- The probability of a given number of events k occurring in a fixed interval of time t, when
  - these events occur with a known average rate **λ**.
  - the events are independent of the time since the last event.

- $Pr(X = k) = \frac{(t \cdot \lambda)^k}{k!} e^{-t \cdot \lambda} = \frac{\gamma^k}{k!} e^{-\gamma}$

- where $\gamma = t \cdot \lambda$ are the expected number of events in time interval t.



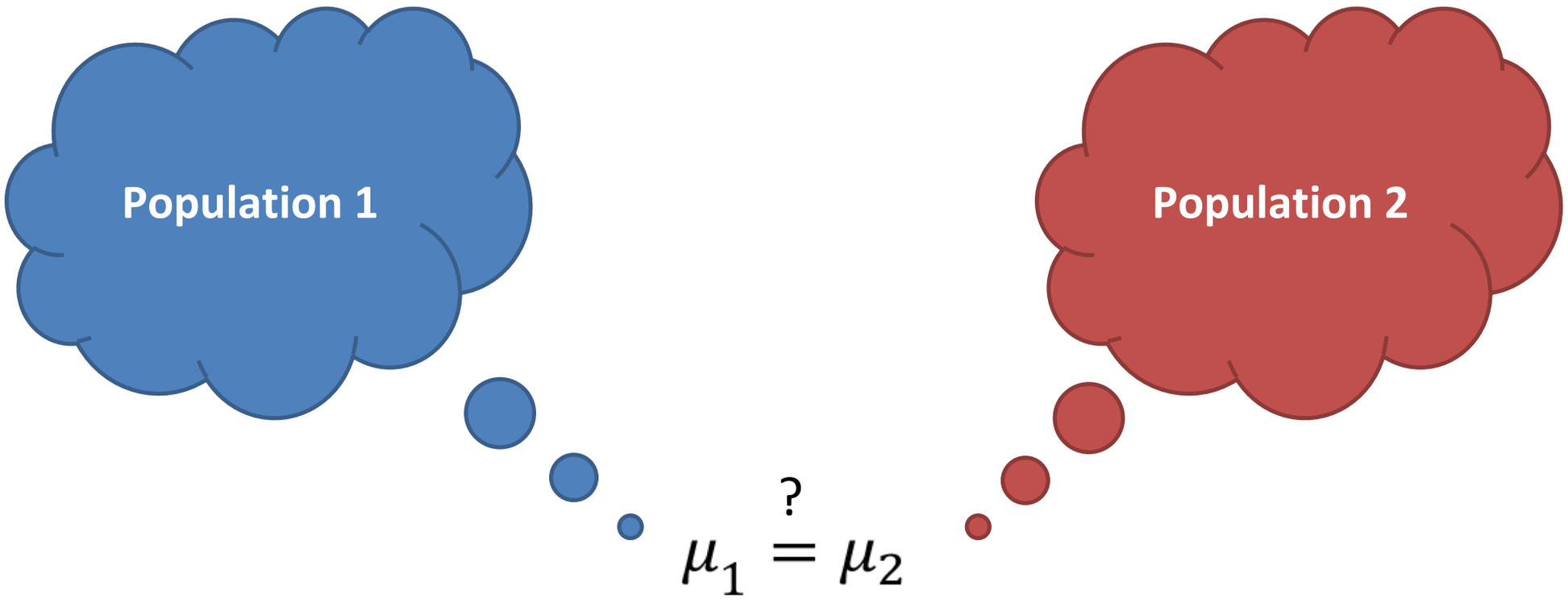
Test Catalog for the Poisson Distribution

- **Statistical model:**
- $X \sim poisson(\lambda \cdot t)$
- Parameter estimate: $\hat{\lambda} = x/t$
- Where the observation is *x* = 'number of arrivals/events observed over a period of time *t*'

- **Hypothesis test (two-tailed):**
- $H_0: \lambda = \lambda_0$
- $H_1: \lambda \neq \lambda_0$
- Test size: $z = \frac{x - \lambda \cdot t}{\sqrt{\lambda \cdot t}} \sim N(0,1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

- **95% confidence interval:**
- $\lambda_- = \frac{1}{t}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{x + \frac{1.96^2}{4}}\right]$
- $\lambda_+ = \frac{1}{t}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{x + \frac{1.96^2}{4}}\right]$

Comparing two population means



- Fx. The height of people from Funen (μ_1) and Jutland (μ_2)

Test catalog for Comparing Two Means (known variance)

Statistical model:

- $X_{1i} \sim N\left(\mu_1, \sigma_1^2\right), i = 1,2, \dots, n_1$ and $X_{2i} \sim N\left(\mu_2, \sigma_2^2\right)\ i = 1,2, \dots, n_2$
- Parameter estimate: $\hat{\delta} = \bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2\right)$
- Where the observation is $\bar{x}_1 - \bar{x}_2$ = 'the difference between two sample means'.

Hypothesis test (two-tailed):

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$
- Test size: $z = \frac{(\bar{x_1} - \bar{x_2})}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0,1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

95% confidence interval:

- $\delta_- = (\bar{x}_1 - \bar{x}_2) - 1.96 \cdot \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- $\delta_+ = (\bar{x}_1 - \bar{x}_2) + 1.96 \cdot \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Test Catalog for Comparing Two Means (unknown variance)

Statistical model:

- $X_{1i} \sim N\left(\mu_1, \sigma_1^2\right), i = 1,2,\dots,n_1$ and $X_{2i} \sim N\left(\mu_2, \sigma_2^2\right)\ i = 1,2,\dots,n_2$
- Parameter estimate:

$$\hat{\delta} = \bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

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

- Where the observation is $\dot{x_1} - \dot{x_2}$ = 'the difference between two sample means'.

Hypothesis test (two-tailed):

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$
- Test size: $t = \frac{(\dot{x_1}-\dot{x_2})}{s\sqrt{1/n_1+1/n_2}} = \sim t(n_1+n_2-2)$
- Approximate p-value: $2 \cdot \left(1 - t_{cdf}(|t|, n_1+n_2-2)\right)$

95% confidence interval:

- $\delta_- = (\dot{x_1} - \dot{x_2}) - t_0 \cdot s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
- $\delta_+ = (\dot{x_1} - \dot{x_2}) + t_0 \cdot s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
- where t0 = tinv(1-0.05/2,n1+n2-2)

OBS:
t-test (compared with Z-test)

- *Less knowledge*
- *Larger uncertaincy*
- *Confidence interval larger*
- *More difficult to reject H_o*

Test Catalog for Paired Data

Statistical model:

- $d_i = X_{1i} - X_{2i}$, where $d_i \sim N(\delta, \sigma^2), i = 1,2,\dots,n$
- Parameter estimate:

$$\hat{\delta} = \overline{d} = \frac{1}{n}\sum_{i=1}^{n} X_{1i} - X_{2i}$$

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2$$

- Where the observation is $\overline{d}$ = 'the average of the differences between paired samples'.

Hypothesis test (two-tailed):

- $H_0: \delta = \delta_0$
- $H_1: \delta \neq \delta_0$
- Test size: $t = \frac{\overline{d} - \delta_0}{s_d/\sqrt{\mathrm{n}}} = \sim t(\mathrm{n}-1)$
- Approximate p-value: $2 \cdot \left(1 - t_{cdf}(|t|, \mathrm{n}-1)\right)$

95% confidence interval:

- $\delta_- = \overline{d} - t_0 \cdot \frac{s_d}{\sqrt{n}}$
- $\delta_+ = \overline{d} + t_0 \cdot \frac{s_d}{\sqrt{n}}$

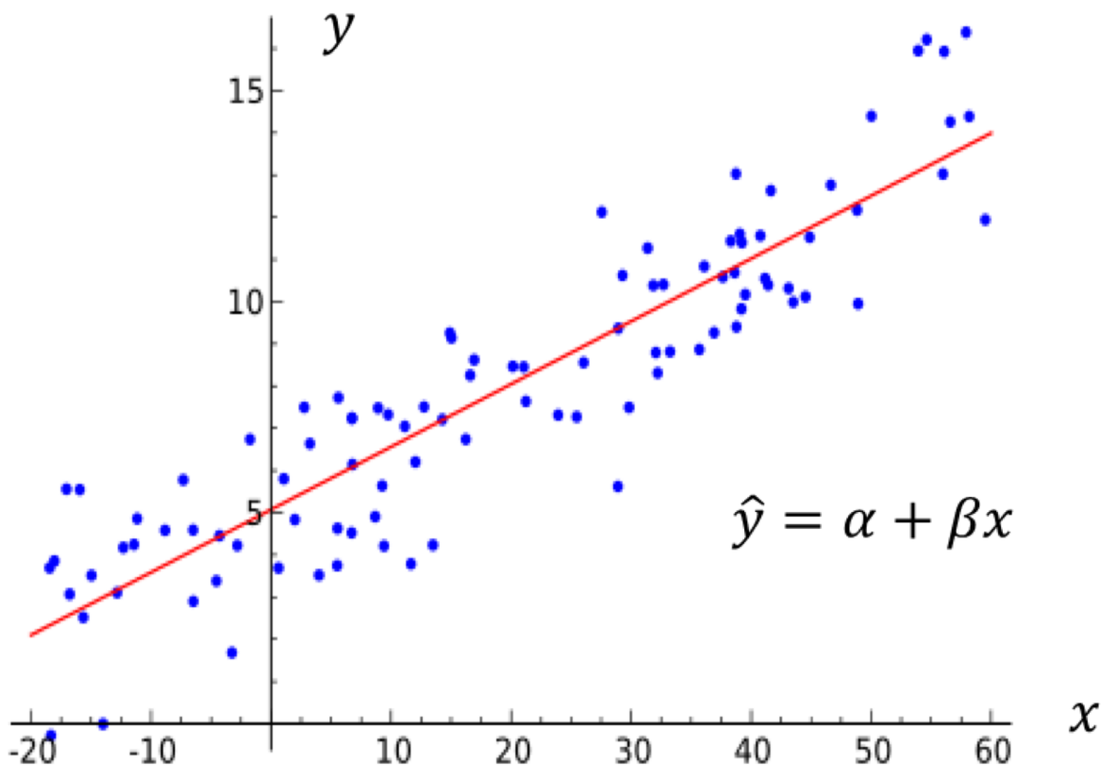
- where t0 = tinv(1-0.05/2,n-1)

Paired test (vs. unpaired test):

- *A one-to-one correspondance between X_1 and X_2 data*
- *Sample size n_1 and n_2 equal*
- *Elimination of factors not related to the test*
- *Reducing uncertaincy*
- *Easier to reject the H_o hopothesis*

Linear Regression

- Fits a straight line through the set of n points (x_i, y_i)
- Make the sum of squared residuals ($\epsilon^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \sim \chi^2$ - the vertical distances between til points of data and the fitted line) of the model as small as possible
- Statistical model: $y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$



Model Fitting

- The goal of linear regression is to determine the choice of slope (β) and intercept (α) that minimizes the sum of squared residuals of the model.

$$R(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 ,$$

ϵ_i^2

- The parameter estimates that minimize R are

$$\widehat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{n} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}})}{\sum_{i=1}^{n} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})^2} = \frac{\boldsymbol{s_{xy}}^2}{\boldsymbol{s_x}^2} = \boldsymbol{r}\frac{\boldsymbol{s_x}}{\boldsymbol{s_y}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}$$

- where $\bar{x}$ is the average of $x_1, x_2, \ldots, x_n$ and $\bar{y}$ is the average of $y_1, y_2, \ldots, y_n$.

Statistical Inference on the Regression Slope

- In general, the null hypothesis about the slope that we wish to test takes the following form

$$H_0: \beta = \beta_0$$

- It can be shown that the estimator of the slope is normally distributed with mean β and variance

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

- where σ^2 is the variance used in the statistical model, $y_i \sim N(\alpha + \beta x_i, \sigma^2)$.
- Using the estimated variance, s_r^2, instead of the population variance, the appropriate test statistic for $\widehat{\beta}$ is

$$t = \frac{\widehat{\beta} - \beta_0}{s_r \sqrt{1/\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim t(n-2)$$

- The p-value is

$$2 \cdot \left(1 - t_{cdf}(|t|, n-2)\right)$$

Statistical Inference on the Regression Intercept

- In general, the null hypothesis that we wish to test takes the following form

$$H_0: \alpha = \alpha_0$$

- It can be shown that the estimator of the intercept is normally distributed with mean α and variance

$$\sigma^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

- where σ^2 is the variance used in the statistical model, $y_i \sim N(\alpha + \beta x_i, \sigma^2)$.
- The appropriate test statistic for $\widehat{\alpha}$ is

$$t = \frac{\widehat{\alpha} - \alpha_0}{s_r \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}} \sim t(n-2)$$

- The p-value is

$$2 \cdot \left(1 - t_{cdf}(|t|, n-2)\right)$$

Confidence intervals

- The 95% confidence interval for the slope β is:

$$\beta_- = \hat{\beta} - t_0 \cdot \frac{s_r}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \hat{\beta} - t_0 \cdot \frac{s_r}{s_x} \cdot \frac{1}{\sqrt{n-1}}$$

$$\beta_+ = \hat{\beta} + t_0 \cdot \frac{s_r}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \hat{\beta} + t_0 \cdot \frac{s_r}{s_x} \cdot \frac{1}{\sqrt{n-1}}$$

- The 95% confidence interval for the intercept α is:

$$\alpha_- = \hat{\alpha} - t_0 \cdot s_r \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$$\alpha_+ = \hat{\alpha} + t_0 \cdot s_r \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

- where

$$t_0 = tinv\left(1 - \frac{0.05}{2}, n-2\right) = tinv(0.975, n-2)$$

Checking for Normality

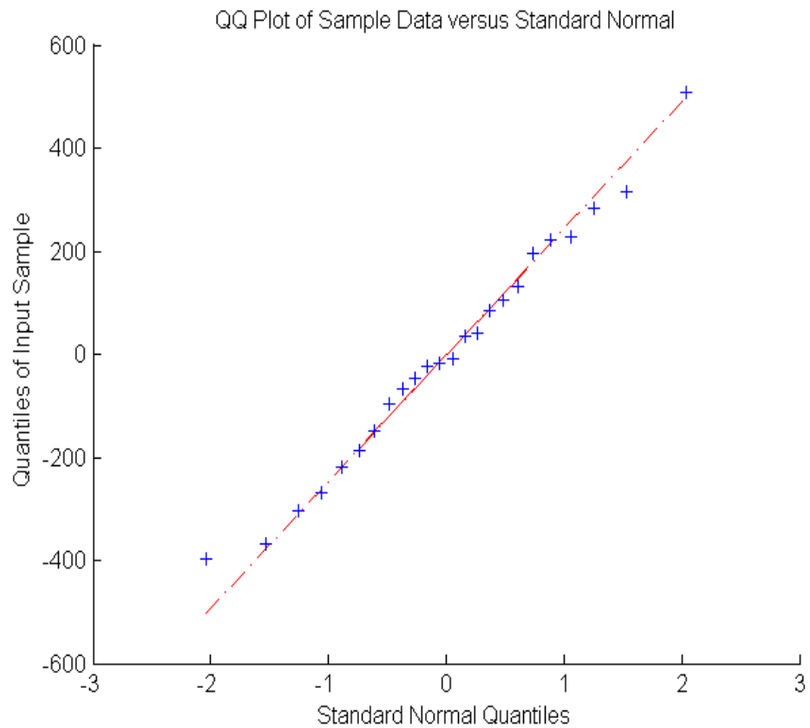
- Recalling that the statistical model underlying linear regression is

$$y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$$

- the residual of the i'th sample should be normally distributed with zero mean and variance σ^2

$$\epsilon_i = y_i - (\alpha + \beta x_i) \sim \mathcal{N}(0, \sigma^2)$$

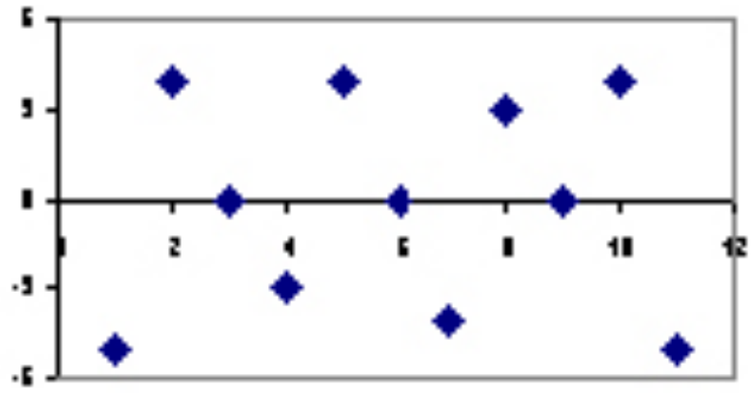
- Hence, a good way to check whether the assumption of linearity between x and y holds is to first fit the linear model and subsequently check that the residuals ϵ_i are normally distributed using a Q-Q plot.

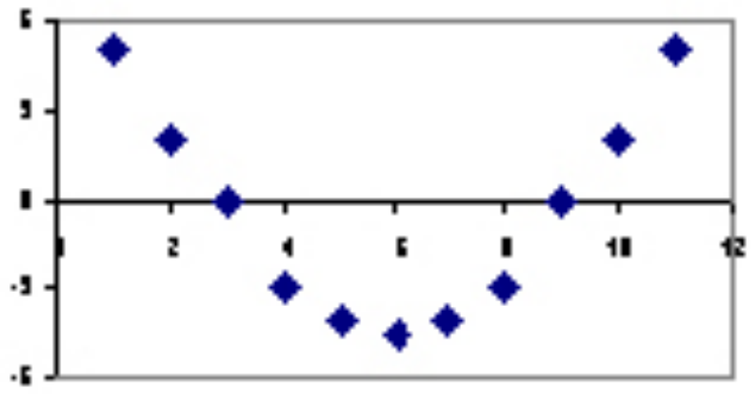


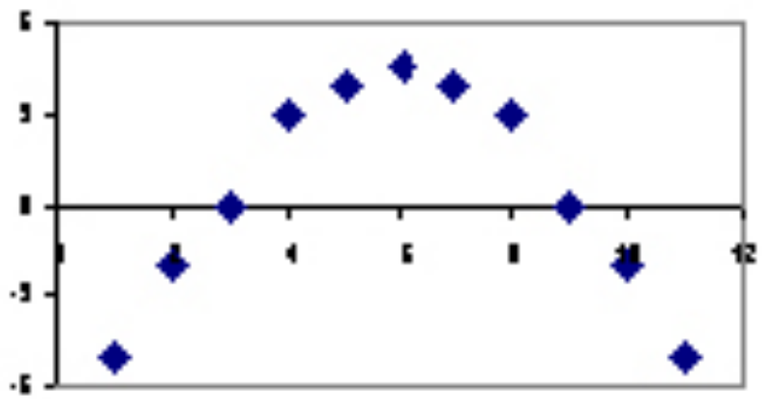
Q-Q plot

Residual Plots

- Another way to check the normality assumption is to make a so-called *residual plot*.
- A residual plot is a graph that shows the residuals on the vertical axis and the independent variable (x) on the horizontal axis.
- If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.







Sample Correlation Coefficient
Coefficient of determination

- If we wish to quantify the strength of a linear relation, we can use the **sample correlation coefficient** ($-1 \leq r \leq 1$):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{{s_{xy}}^2}{s_x \cdot s_y} = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}}$$

- or the **coefficient of determination**: $R^2 = r^2$
- where s_x and s_y are the empitical standard deviations and ${s_{xy}}^2$ the empirical covariance of x and y.

- $0 \leq R^2 \leq 1$ indicates how well the data fit the linear model
- $R^2 \sim 1$ suggesting a strong linear relationsship (a good linear fit)
- $R^2 \sim 0$ suggesting no linear relationsship (a poor linear fit)

Outliers

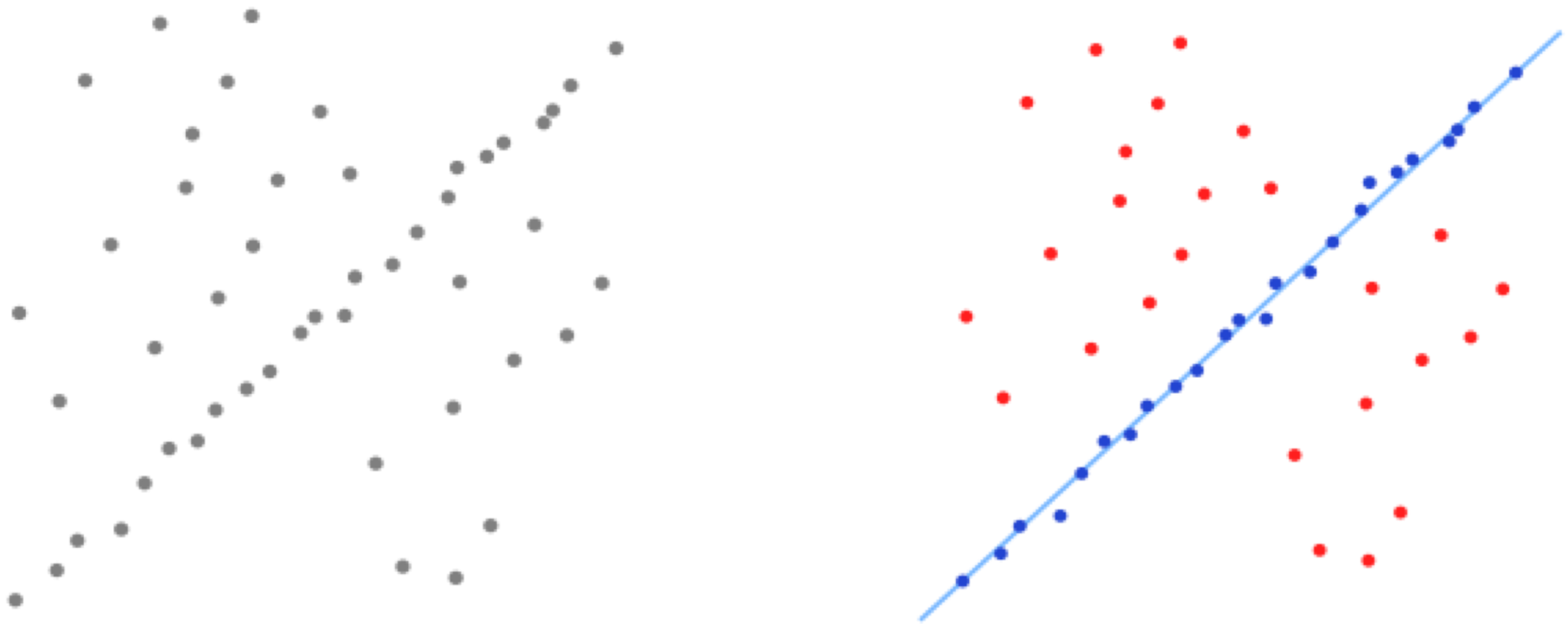
- Outliers are data points that are separated from the rest of the data and potentially influential for the regression analysis.
- Outliers can have a dramatic on the sample correlation coefficient (and therefore the slope).
- Recalling the definition of the sample correlation coefficient,

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot \left(y_i - \bar{y}\right)}{(n-1) \cdot s_x \cdot s_y},$$

- an outlier is a point $\left(x_i, y_i\right)$, such that either $(x_i - \bar{x})$ or $\left(y_i - \bar{y}\right)$, or both, is large.
- The extent of influence of any point can be judged in part by computing the correlation coefficient with and without that point.

RANSAC

Random Sample Consensus



RANSAC is an iterative method to estimate parameters of a mathematical model from a set of observed data which contains outliers. It is a non-deterministic algorithm in the sense that it produces a reasonable result only with a certain probability, with this probability increasing as more iterations are allowed.

Words and Concepts to Know

Descriptive statistics
Statistic
Bernoulli Trial
Estimator
Biased/Unbiased
Linear Regression
Heavy
Random Sample Consensus
Left-tailed
Chi-Square Distribution
Paired test
Central Limit Theorem
Population
Null hypothesis
Linear Model
Quantiles
Test catalog
Slope parameter
Sample variance
Sample Correlation Coefficient
Normal approximation
Intercept parameter
Significance Level
Regression Intercept
Inferential statistics
Q-Q plot
Statistical model
Alternative hypothesis
Two-sided
t-score
Average rate
Extrapolation
χ_k^2
Confidence Level
Outliers
Test statistics
True mean
Sample mean
Predicted data
Maximum-likelihood
Model Fitting
Slope parameter
Binominal Distribution
Right-tailed
Pooled variance
Students t-distribution
Unpaired test
Residual
Empirical Variance
Comparing two population means
Standard Normal Distribution
Poisson Distribution
Residual plot
z-statistic
Reject
p-value
Inliers
Sample
Chi-Square Test
Two-tailed
Hypothesis test
One-sided
Degrees of freedom
Confidence Interval
Measured data
RANSAC
Sample Size
Critical values
Coefficient of Determinaton
Fail to reject