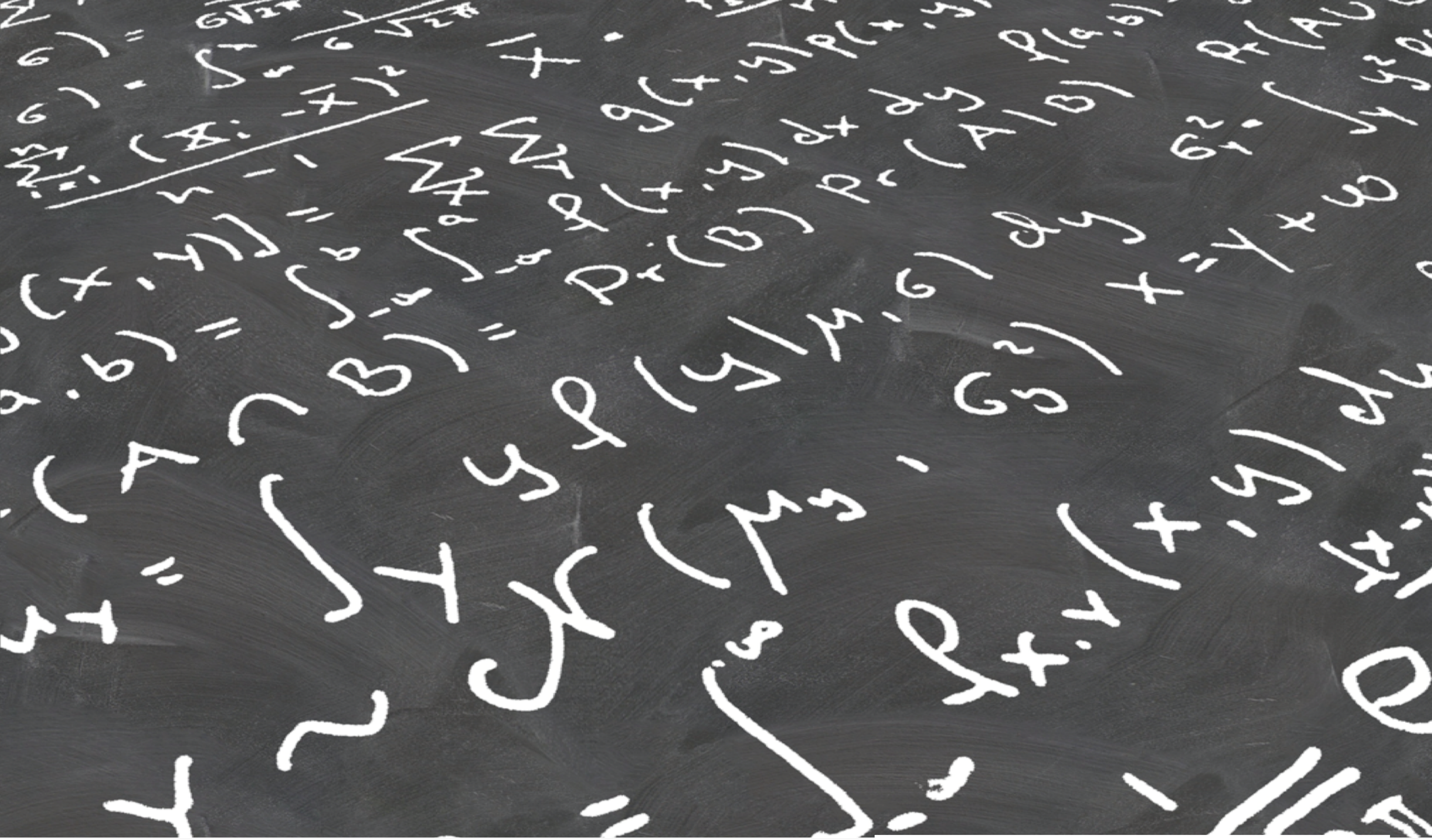


11. Chi-Square tests, the Binomial and Poision Distributions

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

Todays Content

- Repetition from last time
- Chi-Square Test
- The Binomial Distribution
  - Approximation to the Normal distribution
- The Poisson Distribution
  - Approximation to the Normal distribution

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

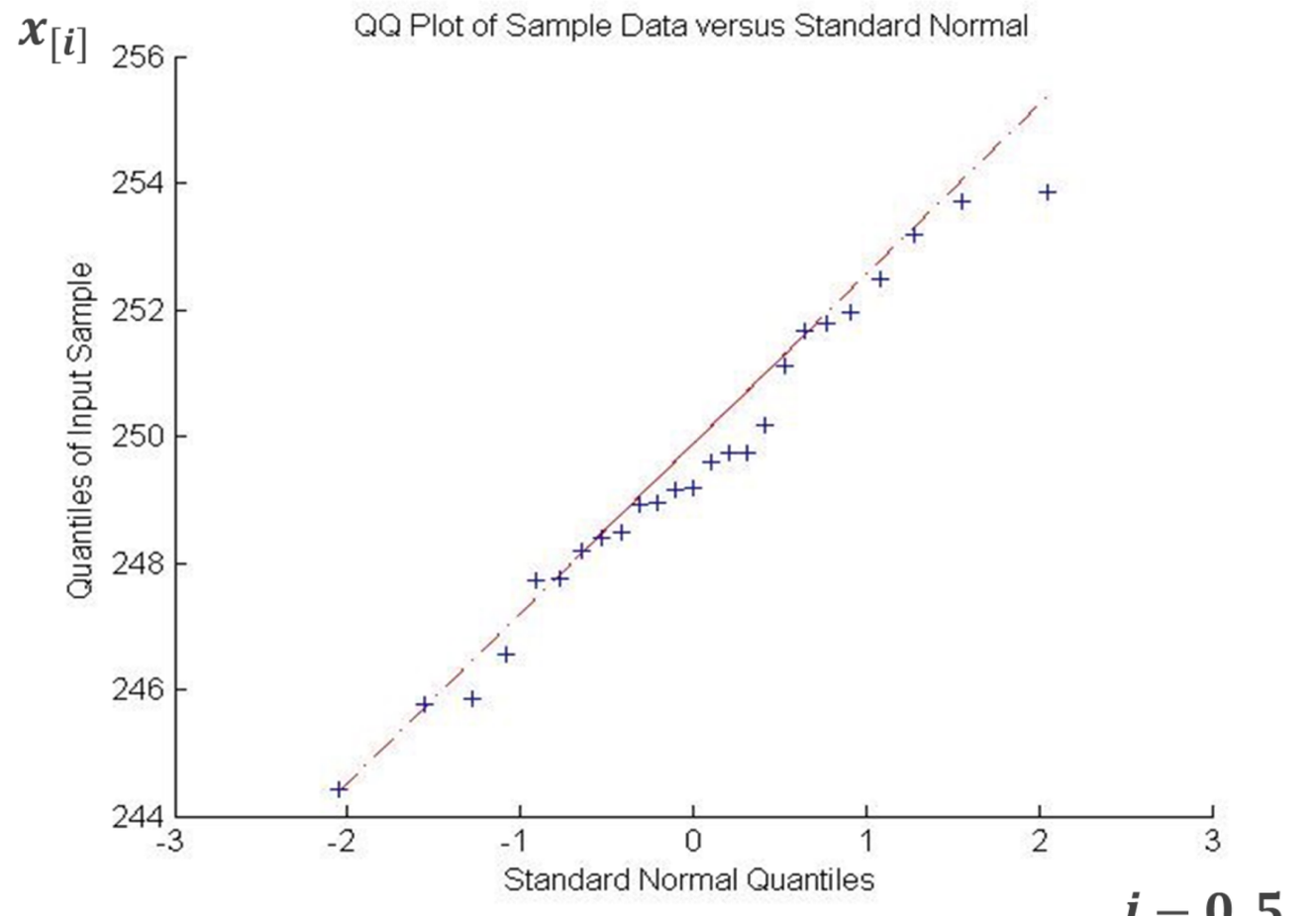
- **95% confidence interval:**
- $\mu_- = \bar{x} - t_0 \cdot s/\sqrt{n}$
- $\mu_+ = \bar{x} + t_0 \cdot s/\sqrt{n}$

- 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

- $\hat{\mu}$ → Gaussian Distributed $(\sum X_i)$ (CLT)
- $\hat{\sigma}^2$ → Not Gaussian Distributed $\left(\sum X_i^2\right) \rightarrow \chi^2$-distributed

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

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

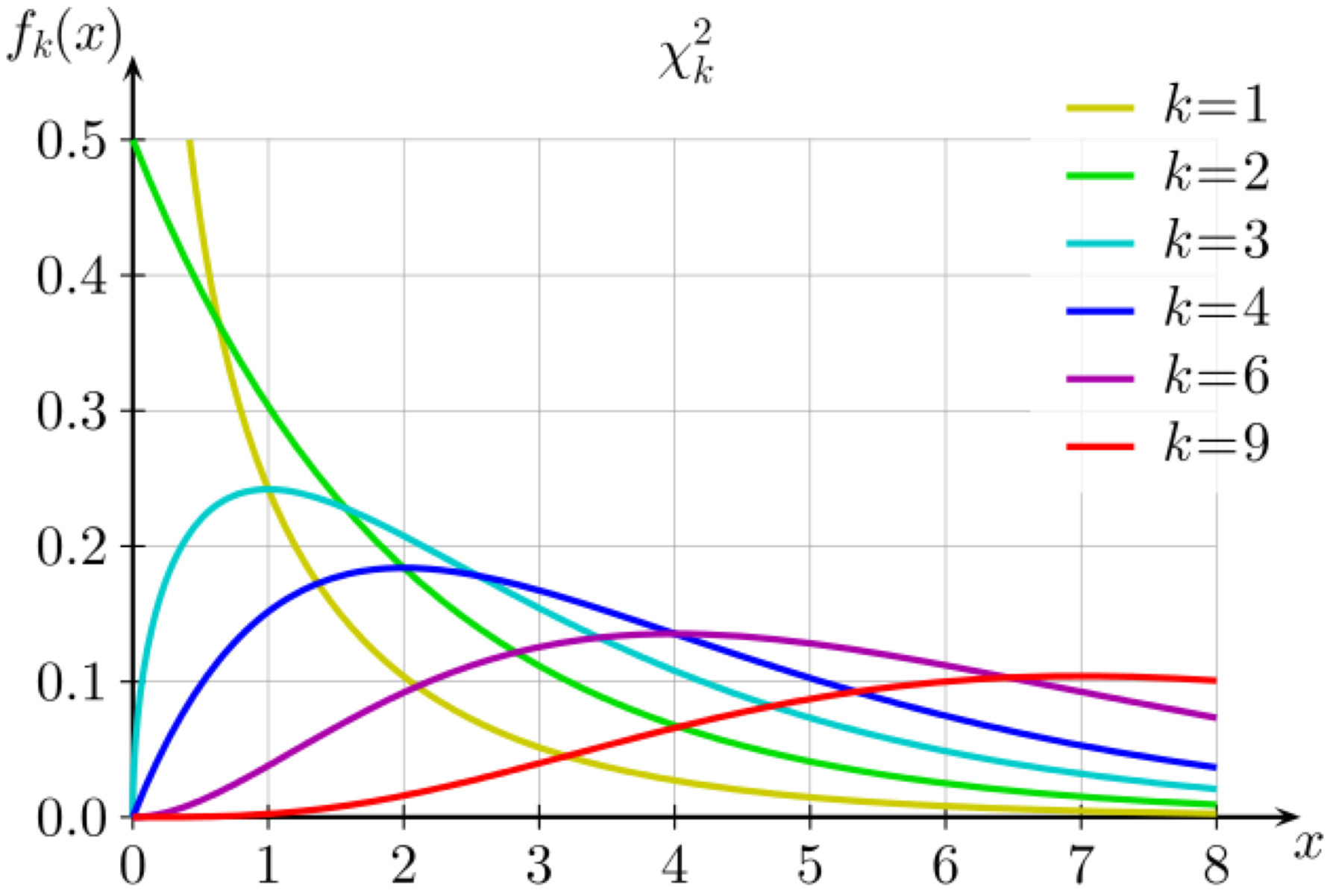
OBS: Standard (normalized) normal distibition

- Then have that: $Q = \sum_{i=1}^{n} X_i^2$
  is χ_k^2 distributed – k is the degree of freedom.

n-1

- Cdf: $f_{\chi_k^2}(x) = \frac{1}{2^{k/2} \cdot \Gamma(k/2)} \cdot x^{k/2-1} \cdot e^{-x/2}$ (Matlab: chi2cdf(x,k))

Chi-Square Distribution



$-\infty \geq \mathcal{N}(0,1) \geq \infty$

$\chi^2_k \geq 0$

Chi-Square Test for Independence

- We can compute the sample variance for an experiment with n observations:

$$X^2 = \sum_{i=1}^{n} \frac{(observed - expected)^2}{Expected}$$

- if X^2 is large, then the data is a poor fit, and thus the NULL hypothesis is rejected.

Observed = Expected

k=4, α=0.05: H_0 rejected if X^2>9.4 (see tabel 3 in "Random Signals")

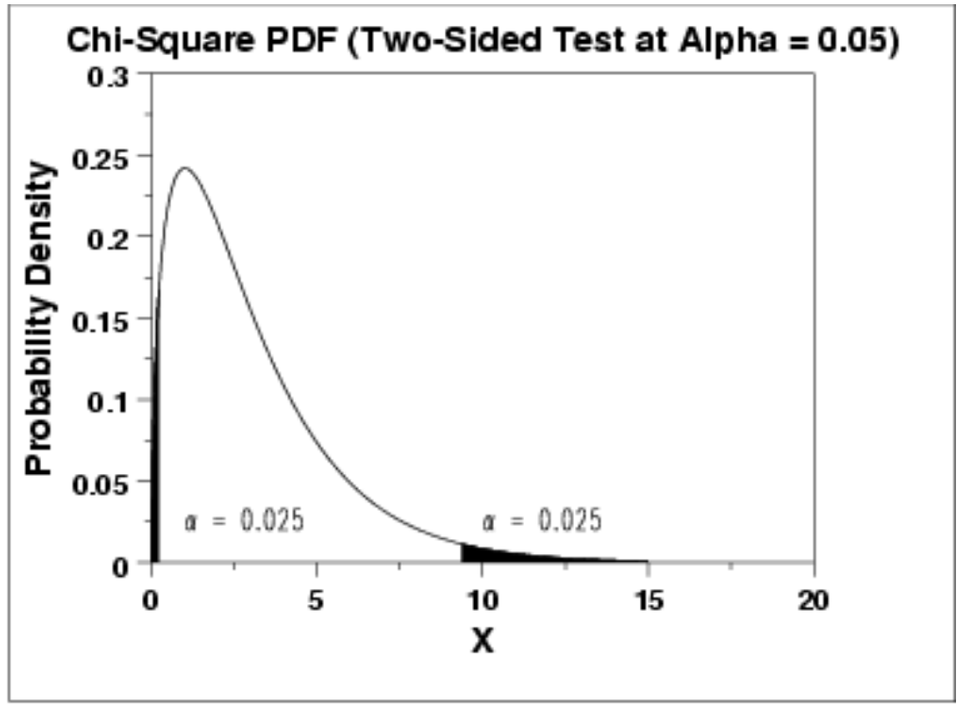
- Fx: How well fits data with an expected function (curve)?

Chi-Square Test for Variance

- Hypothesis: $H_0 : \sigma^2 = \sigma_0^2$

  $H_1 : \sigma^2 \neq \sigma_0^2$

- Test statistics: $T = (N-1) \cdot \frac{s^2}{\sigma_0^2}$
- N = sample size (should be large).



From "Engineering Statistics Handbook"

- For a two tailed test we fail to reject NULL if:

$$\chi^2_{N-1,\alpha/2} < T < \chi^2_{N-1,1-\alpha/2}$$

- where $\chi^2_{N-1,*}$ are the lower and upper critical values of the Chi-Square distribution with N-1 degrees of freedom

Bernoulli Trial

- Two possible outcomes

$$B = \{0,1\}$$

- Probabilities

$$\Pr(B = 1) = p \qquad \text{(success)}$$
$$\Pr(B = 0) = 1 - p \qquad \text{(failure)}$$

- Notation

$$B \sim bernoulli(p)$$

The Binomial Distribution

- Let $B_1, B_2, \ldots, B_n$ be independent random variables, where

$$B_i \sim bernoulli(p)$$

- Then the number of successes

$$X = \sum_{i=1}^{n} B_i$$

- is a binomially distributed random variable with parameters *n* and *p*.

- Notation

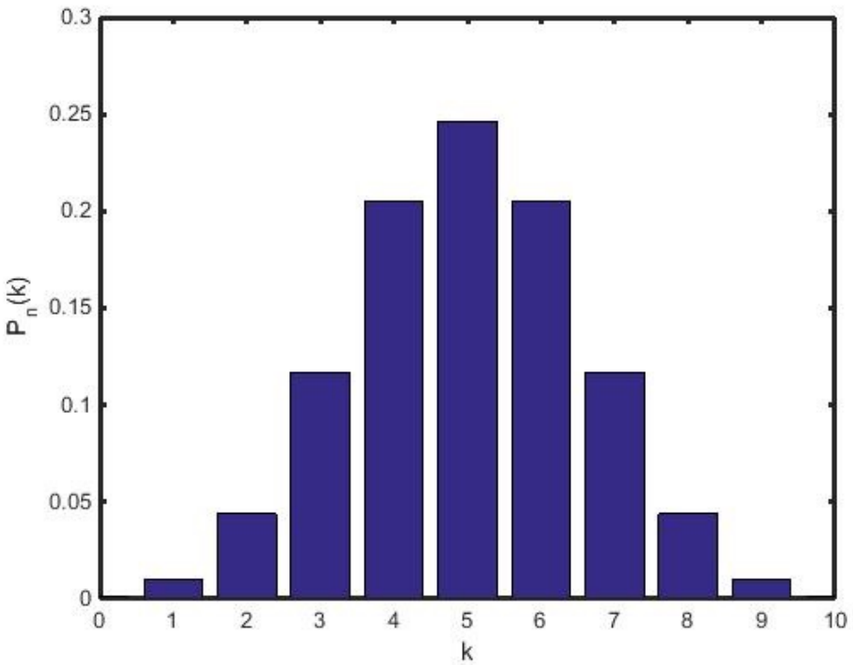
$$X \sim binomial(n, p)$$

The Binomial Distribution

- We have n repeated trials.
- Each trial has two possible outcomes
  - **Success** — probability p
  - **Failure** — probability 1-p
- We write the mass function as:

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

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



Mendel's Experiment

- Pea seed color is controlled by two alleles
  - 'A' the dominant (green)
  - 'a' the recessive (yellow)
- Genotypes
  - 'AA', 'Aa', 'aA' should produce a green pea.
  - 'aa' should produce a yellow pea.
- Mendel's hypothesis
  - Crossing 'Aa' genotypes should result in equally many pea plants of each genotype.
- More formally
  - $\Pr(yellow\ plant) = \frac{1}{4}$
  - $\Pr(green\ plant) = \frac{3}{4}$

Mendel's Experiment

- Mendel looked at the colors of 580 offspring plants.
- Result
  - 152 yellow plants
  - 428 green plants

Two possible outcomes → Bernoulli Trial

- In an idealized experiment
  - 145 yellow plants (580/4)
  - 435 green plants (580*3/4)
- Could the deviation be explained by random variation?
- Or is Mendel's hypothesis incorrect?

Mendel's Experiment

- We denote by X the number of yellow plants.
- Then the statistical model is

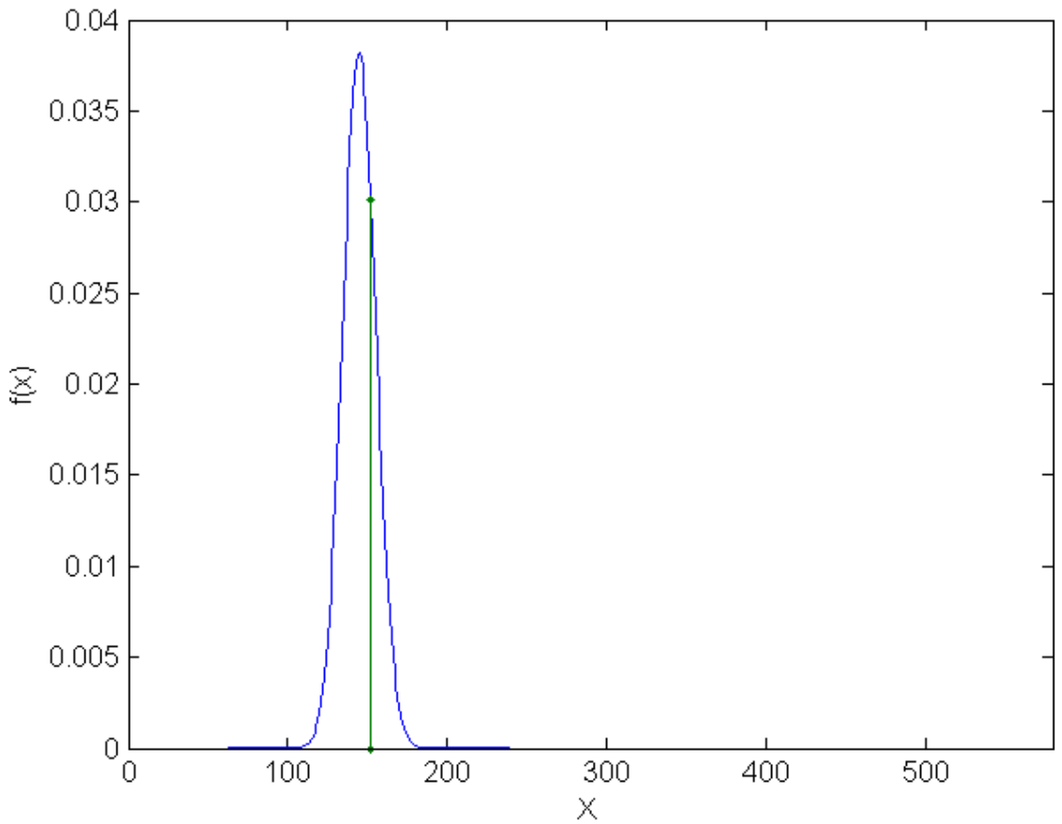
$$X \sim binomial(n = 580, p)$$

- The hypothesis concerns the unknown parameter, p

$$H_0: p = 1/4$$
$$H_1: p \neq 1/4$$

Mendel's Experiment

- The pdf for the underlying Mendel's experiment (n=580, p=1/4) along with the observed value, x=152.




```
n      = 580;  % Number of trials
p      = 0.25; % Probability of success
x      = 0:n;  % x-values for plotting the PDF
fx     = binopdf(x,n,p); % PDF
xobs   = 152;  % Observed x (=number of
                                  successes)

plot(x,fx,…
     …[xobs xobs],[0 binopdf(xobs,n,p)],'.-')
axis([0 580 0 0.04])
xlabel('X')

ylabel('f(x)')
```

p-value for Mendel's Experiment

- The p-value is the probability of observing a value of the random variable X that is more extreme than 152.
- By extreme we mean with respect to the value of X that we would observe in an idealized experiment, given that the null hypothesis is true:

$$p \cdot n = \frac{1}{4} \cdot 580 = 145$$

- Hence, for a two-tailed test, we need to consider the events $\{X \geq 152\}$ and $\{X \leq 138\}$, both of which deviate by 7 from the theoretical value of 145.

p-value for Mendel's Experiment

- Calculation of p-value:

$$pval = 2 \cdot min\{Pr(X \geq x_{max}), Pr(X \leq x_{min})\}$$

$$= 2 \cdot min\{Pr(X \geq 152), Pr(X \leq 138)\}$$

$$= 2 \cdot min\{1 - Pr(X < 152), Pr(X \leq 138)\}$$

$$= 2 \cdot min\left\{1 - F_{bino}\left(151; n = 580, p = \frac{1}{4}\right), F_{bino}\left(138; n = 580, p = \frac{1}{4}\right)\right\}$$

$$= 2 \cdot min\{1 - 0.7350, 0.2682\} = 2 \cdot min\{0.2650, 0.2682\}$$

$$= 0.5300 > 0.05 = \alpha$$

- where $F_{bino}(x;n,p)$ denotes the cdf of a binomial distribution with parameters n and p.
- Since pval>α=0.05, we fail to reject the null hypothesis (Mendel's hypothesis)

Binomial Distribution in Matlab

- Calculating the probabilities $\Pr(X = x)$ and $\Pr(X \leq x)$ of a binomially distributed random variable

$$X \sim binomial(n, p)$$

- $\Pr(X = x)$ = binopdf(x,n,p)
- $\Pr(X \leq x)$ = binocdf(x,n,p)

- x must be an integer value and can in general be a vector or array.

Normal Approximation to the Binomial Distribution

- First, consider

$$B \sim bernoulli(p)$$

- Mean

$$E[B] = \sum_{b=\{0,1\}} b \cdot \Pr(B = b) = 0 \cdot \Pr(B = 0) + 1 \cdot \Pr(B = 1)$$

$$= 0 \cdot (1 - \mathrm{p}) + 1 \cdot \mathrm{p} = \mathrm{p}$$

- Variance

$$Var(B) = \sum_{b=\{0,1\}} (b - p)^2 \cdot \Pr(B = b)$$

$$= (0 - p)^2 \cdot \Pr(B = 0) + (1 - p)^2 \cdot \Pr(B = 1)$$

$$= p^2(1 - p) + (1 - p)^2 p = p(1 - p)$$

Normal Approximation to the Binomial Distribution

- Now, define

$$X = \sum_{i=1}^{n} B_i$$

- where $B_i \sim bernoulli(p)$, and $B_i's$ are independent.
- Mean

$$E[X] = E\left[\sum_{i=1}^{n} B_i\right] = \sum_{i=1}^{n} E[B_i] = \sum_{i=1}^{n} p = np$$

- Variance

$$Var(X) = Var\left(\sum_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} Var(B_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

Normal Approximation to the Binomial Distribution

- Now define the standardized random variable

$$Z = \frac{X - E[X]}{\sqrt{Var(X)}} = \frac{X - np}{\sqrt{np(1-p)}}, \text{ where } X \sim binomial(n, p)$$

approximately

- Then if $np > 5$ and $n(1-p) > 5$, Z is standard normally distributed

$$Z \sim N(0,1)$$

- This fact follows from the central limit theorem:

$$\boldsymbol{X = \sum_{i=1}^{n} B_i \sim N(n \cdot E[B], n \cdot Var(B)) = N(np, np(1-p))}$$

Approximate p-value for Mendel's Experiment

- Standardizing the observation, x=152, we get

$$z = \frac{x - np}{\sqrt{np(1-p)}} = \frac{152 - 580 \cdot 1/4}{\sqrt{580 \cdot 1/4 \cdot (1 - 1/4)}} = \frac{152 - 145}{\sqrt{145 \cdot 3/4}} = 0.6712$$

- and the two-tailed p-value is

$$\begin{aligned} pval &= 2 \cdot min\{Pr(Z \geq z), Pr(Z \leq z)\} \\ &= 2 \cdot min\{Pr(Z \geq 0.6712), Pr(Z \leq 0.6712)\} \\ &= 2 \cdot min\{1 - Pr(z < 0.6712), Pr(X \leq 0.6712)\} \\ &= 2 \cdot min\{1 - \Phi(0.6712), \Phi(0.6712)\} \\ &= 2 \cdot min\{1 - 0.7490, 0.7490\} = 2 \cdot min\{0.2510, 0.7490\} \\ &= 0.5021 > 0.05 = \alpha \end{aligned}$$

- which is slightly smaller than the exact p-value calculated earlier, but leads to the same result: failure to reject the null hypothesis.

➢ OBS: If $p \approx \alpha$ – the normal distribution approximation should not be used

Estimation of p in Binomially Distributed Data

- The estimator of p is

$$\hat{p} = x/n$$

- Unbiased

$$E[\hat{p}] = E[x/n] = \frac{1}{n}E[x] = \frac{1}{n}np = p$$

- Variance

$$Var(\hat{p}) = Var\left(\frac{x}{n}\right) = \frac{1}{n^2}Var(x) = \frac{1}{n^2}np(1-p) = \frac{1}{n}p(1-p)$$

- Notice that the variance of the estimate decreases with $1/n$.

Approximate 95% Confidence Interval

- To find the 95% confidence interval for the parameter p, we must find the limits p_- and p_+, such that the true parameter p lies in the interval $[p_-; p_+]$ with probability 0.95:

$$Pr(p_- \leq p \leq p_+) = 0.95$$

- Assuming that we can use the normal approximation, this condition is equivalent to:

$$Pr(-1.96 \leq z \leq 1.96) = 0.95$$

- where z is the standarized random variable defined earlier: $z = \frac{x-np}{\sqrt{np(1-p)}}$

- Inserting we get:

$$Pr\left(-1.96 \leq \frac{x-np}{\sqrt{np(1-p)}} \leq 1.96\right) = 0.95 \quad \Rightarrow$$

$$Pr\left(\frac{1}{n+1.96^2}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1.96^2}{4}}\right] \leq p \leq \frac{1}{n+1.96^2}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1.96^2}{4}}\right]\right) = 0.95$$

Approximate 95% Confidence Interval

- And therefore we get for the 95% confidence interval $[p_-; p_+]$:

$$p_- = \frac{1}{n + 1.96^2}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1.96^2}{4}}\right]$$

$$p_+ = \frac{1}{n + 1.96^2}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1.96^2}{4}}\right]$$

- In generel, the limits of the 1-α confidence interval for p are:

$$p_- = \frac{1}{n + u^2}\left[x + \frac{u^2}{2} - u\sqrt{\frac{x(n-x)}{n} + \frac{u^2}{4}}\right]$$

$$p_+ = \frac{1}{n + u^2}\left[x + \frac{u^2}{2} + u\sqrt{\frac{x(n-x)}{n} + \frac{u^2}{4}}\right]$$

- where $u = \Phi^{-1}(1 - \frac{\alpha}{2})$

Estimation of p and 95% Confidence Interval Mendel's Experiment

- Mendel's Experiment:
  - Number of trials (plants): n = 580
  - Number of successses (yellow): x = 152

  - Estimated parameter: $\hat{p} = \frac{x}{n} = \frac{152}{580} = 0.2621$
  - Estimated variance: $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n} = \frac{0.2621 \cdot 0.7379}{580} = 0.000333$

  - 95% confidence interval:

$$p_{-} = \frac{1}{580 + 1.96^2}\left[152 + \frac{1.96^2}{2} - 1.96\sqrt{\frac{152(580-152)}{580} + \frac{1.96^2}{4}}\right] = 0.2279$$

$$p_{+} = \frac{1}{580 + 1.96^2}\left[152 + \frac{1.96^2}{2} + 1.96\sqrt{\frac{152(580-152)}{580} + \frac{1.96^2}{4}}\right] = 0.2993$$

- Since $p = 0.25$ lies within the 95% confidence interval, the null hypothesis can't be rejected.

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'

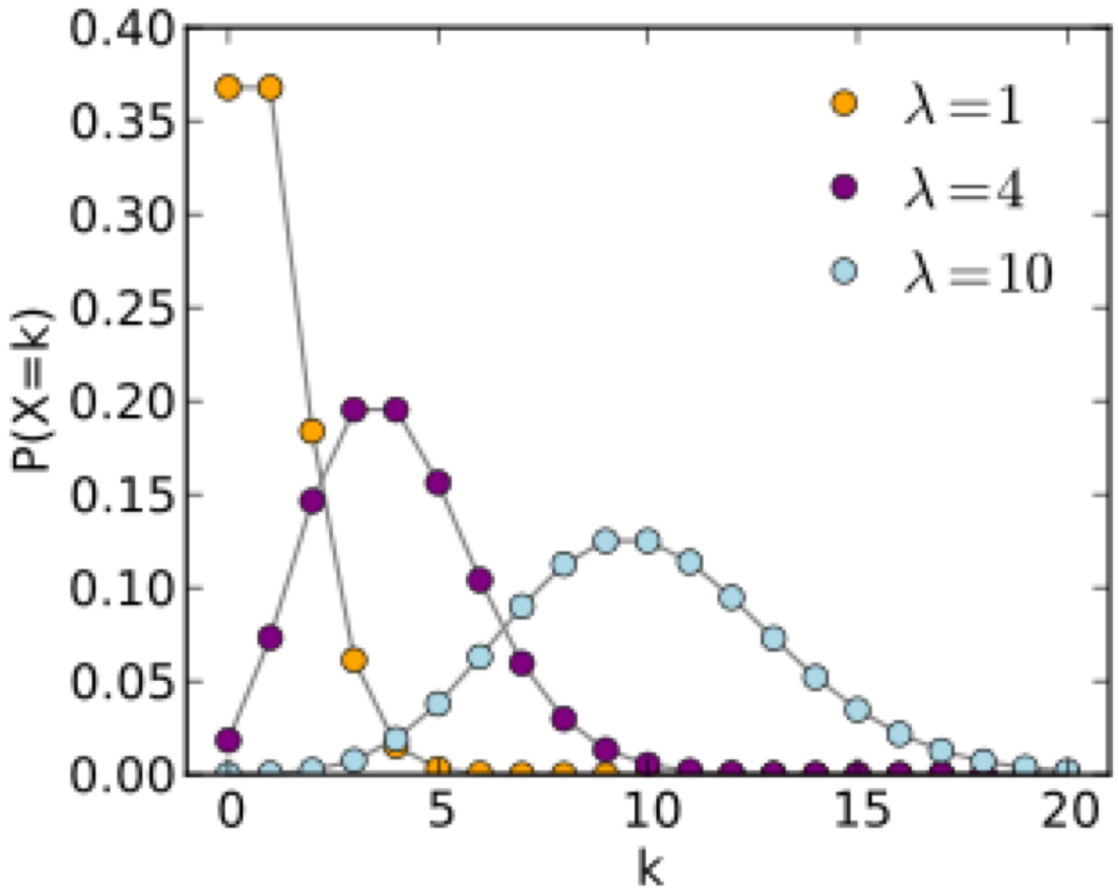
- **Hypothesis test (two-tailed):**
- $H_0: p = p_0$
- $H_1: p \neq p_0$
- Test size: $z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} \sim N(0,1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

- **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, when:
  - these events occur with a known average rate **λ**
  - the events are independently of the time since the last event



The Poisson Distribution

- Assume that we have a time axis that is divided into *N* intervals of length Δt.
- For each interval there is one Bernoulli distributed random variable, denoted B_i for the *i*'th interval, denoting the number of arrivals/events in that interval.
- Recalling that B_i={0,1}, there can be either 0 or 1 arrival in each interval.
- Denoting by λ the known average rate of the arrivals/event, we have

$$B_i \sim bernoulli(\lambda \cdot \Delta t)$$

Δt so small that λ· Δt <1

- That is, the probability of observing an event in the *i*'th interval is proportional to the length (Δt) of the interval.

The Poisson Distribution

- We assume that the observations $B_1, B_2, \ldots, B_N$ are independent.
- Then, the probability of observing $X = x$ events over the entire period of duration $t = N \cdot \Delta t$ is binomially distributed:

$$X \sim binomial(N, \lambda \cdot \Delta t)$$

- Observe that

Expected number of events in time t

$$N \cdot (\lambda \cdot \Delta t) = constant = \frac{t}{\Delta t} \cdot (\lambda \cdot \Delta t) = t \cdot \lambda = \gamma$$

- In the limit, as $N \to \infty$ (or $\Delta t \to 0$), it can be shown that

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

Poisson Distribution in Matlab

- Calculating the probabilities $\Pr(X = x)$ and $\Pr(X \leq x)$ of a Poisson distributed random variable

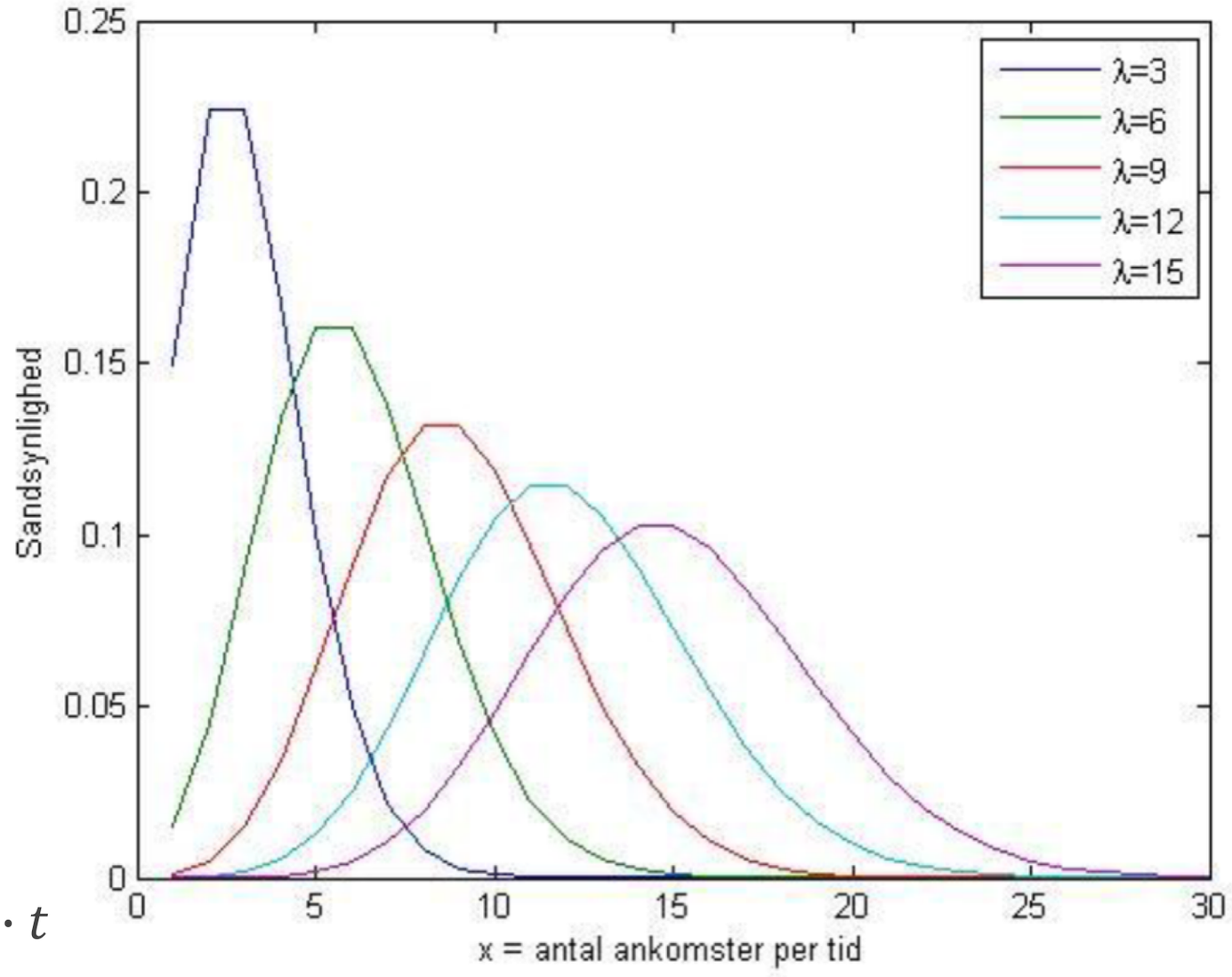
$$X \sim poisson(t \cdot \lambda = \gamma)$$

- $\Pr(X = x) = poisspdf(x, \gamma)$
- $\Pr(X \leq x) = poisscdf(x, \gamma)$

OBS:

x hændelser i tiden t

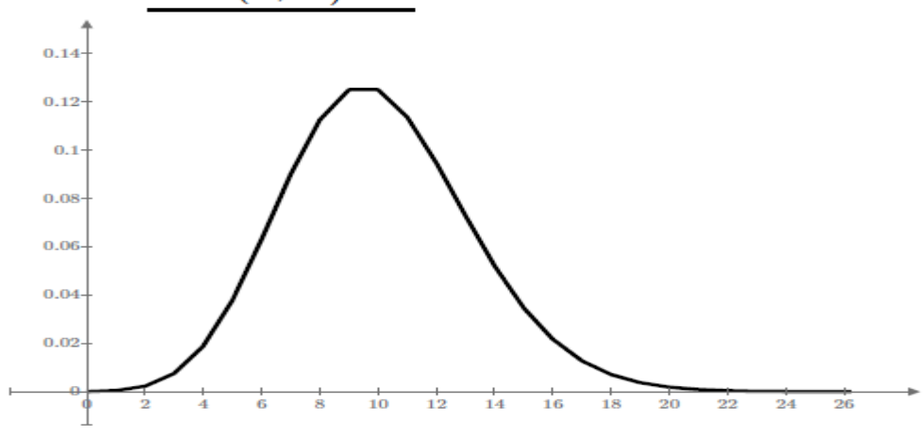
$\gamma = \lambda \cdot t$



The Poisson Distribution

- A custom-handling system receive in average 2 orders pr. minute: $\lambda = 2/min$

- What is the probability that within the next 5 minutes the system should handle x orders:
  - $Pr(X = x) = \frac{\gamma^x}{x!} e^{-\gamma} = \frac{10^x}{x!} e^{-10}$ $(\gamma = \lambda \cdot t = 10)$
    $Pois(x, 10)$
  - 0 orders? $Pr(0) = 0.000045$
  - 8 orders? $Pr(8) = 0.113$
  - 10 orders? $Pr(10) = 0.125$
  - 15 orders? $Pr(15) = 0.035$



- If 99% of all orders should be handled within 5 minutes, how many orders shall the system be designed to handle:
  - $\Pr(X > x) < 0.01 \Rightarrow x \geq poissinv(0.99,10) = 18$ → poisscdf(x, γ)=0.99

The Poisson Distribution

- A store claim to have 150 customers pr. hour. $(\lambda = 150/time)$
- What is the probability that within the next 2 hours the store will have x customers:
  - $Pr(X = x) = \frac{\gamma^x}{x!} e^{-\gamma} = \frac{300^x}{x!} e^{-300}$ $(\gamma = \lambda \cdot t = 300)$

Very large

Very small

- 200 customers? $Pr(200) = 3.01 \cdot 10^{-20}$
- 300 customers? $Pr(300) = 2.30 \cdot 10^{-2}$
- 400 customers? $Pr(400) = 5.67 \cdot 10^{-9}$
- 500 customers? $Pr(500) = 1.53 \cdot 10^{-26}$

Very small numbers

The Poisson Distribution

- A store claim to have 150 customers pr. hour ($\lambda = 150/time$).

- What is the probability that within the next 2 hours the store will have:
  - $Pr(X \leq x) = poisscdf(x, \gamma)$ $\quad (\gamma = \lambda \cdot t = 300)$

- <250 customers? ➢ $poisscdf(250{,}300) = 0.002$
- 275-325 customers? ➢ $poisscdf(325{,}300) - poisscdf(275{,}300) = 0.851$
- >325 customers? ➢ $1 - poisscdf(325{,}300) = 0.072$

Normal Approximation to the Poisson Distribution

- Defining the Poisson distributed random variable

$$X \sim poisson(t \cdot \lambda = \gamma)$$

- it can be shown that

$$E[X] = t \cdot \lambda = \gamma$$
$$Var(X) = t \cdot \lambda = \gamma$$

- Now define the standardized random variable

$$Z = \frac{X - E[X]}{\sqrt{Var(X)}} = \frac{X - t \cdot \lambda}{\sqrt{t \cdot \lambda}} = \frac{X - \gamma}{\sqrt{\gamma}}$$

- Then, if $t \cdot \lambda = \gamma > 5$, Z is approximately standard normally distributed

$$Z \sim N(0,1)$$

Approximate p-value

- A store claims: average rate of 150 customers pr. hour.
  $\lambda = \gamma/t = 150$.

- Observe: $x = 280$ customers for 2 hours.

- Formulation of null hypothesis:

$$H_0: \quad \lambda = 150$$

- Standardising the observation, test statistics:

$$z = \frac{x - t \cdot \lambda}{\sqrt{t \cdot \lambda}} = \frac{280 - 2 \cdot 150}{\sqrt{t \cdot 150}} = -1.1547$$

- Two-tailed p-value:

$$2 \cdot |1 - \Phi(|z|) = 2 \cdot |1 - \Phi(1.1547)| = 2 \cdot |1 - 0.8759| = 0.2482$$

- We Fail to reject til null hypothesis.

Estimation of the Average Rate Parameter

- In general, the average rate parameter $\lambda = \gamma/t$ is unknown and has to be estimated from observed data.
- Given the observation *x* = 'number of arrivals/events" over a time period of duration *t*, the maximum-likelihood estimator is

$$\hat{\lambda} = \frac{x}{t}$$

- This is an unbiased estimator, because the expected value of $\hat{\lambda}$ is the true parameter

$$E[\hat{\lambda}] = \lambda$$

Approximate 95% Confidence Interval

- To find the 95% confidence interval for the parameter λ, we must find the limits λ_- and λ_+, such that the true parameter λ lies in the interval $[\lambda_-; \lambda_+]$ with probability 0.95:

$$Pr(\lambda_- \leq \lambda \leq \lambda_+) = 0.95$$

- Assuming that we can use the normal approximation, this condition is equivalent to:

$$Pr(-1.96 \leq z \leq 1.96) = 0.95$$

- where z is the standarized random variable defined earlier: $z = \frac{x - t\cdot\lambda}{\sqrt{t\cdot\lambda}}$

- Inserting we get:

$$Pr\left(-1.96 \leq \frac{x - t\cdot\lambda}{\sqrt{t\cdot\lambda}} \leq 1.96\right) = 0.95 \quad \Rightarrow$$

$$Pr\left(\frac{1}{t}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{x + \frac{1.96^2}{4}}\right] \leq \lambda \leq \frac{1}{t}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{x + \frac{1.96^2}{4}}\right]\right) = 0.95$$

Approximate 95% Confidence Interval

- And therefore we get for the 95% confidence interval $[\lambda_-; \lambda_+]$:

$$\lambda_- = \frac{1}{t}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{x + \frac{1.96^2}{4}}\right]$$

$$\lambda_+ = \frac{1}{t}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{x + \frac{1.96^2}{4}}\right]$$

- In generel, the limits of the 1-α confidence interval for p are:

$$\lambda_- = \frac{1}{t}\left[x + \frac{u^2}{2} - u\sqrt{x + \frac{u^2}{4}}\right]$$

$$\lambda_+ = \frac{1}{t}\left[x + \frac{u^2}{2} + u\sqrt{x + \frac{u^2}{4}}\right]$$

- where $u = \Phi^{-1}(1 - \frac{\alpha}{2})$

Approximate 95% Confidence Interval

- The parameter estimate is

$$\hat{\lambda} = \frac{x}{t} = \frac{280}{2} = 140$$

- Since $t \cdot \lambda = 2 \cdot 150 = 300 > 5$ we can use the normal approximation of the confidencde interval:

$$\lambda_{-} = \frac{1}{t} \cdot \left[x + \frac{1.96^2}{2} - 1.96 \cdot \sqrt{x + \frac{1.96^2}{4}} \right] = \frac{1}{2} \cdot \left[280 + \frac{1.96^2}{2} - 1.96 \cdot \sqrt{280 + \frac{1.96^2}{4}} \right] = 124.5$$

$$\lambda_{+} = \frac{1}{t} \cdot \left[x + \frac{1.96^2}{2} + 1.96 \cdot \sqrt{x + \frac{1.96^2}{4}} \right] = \frac{1}{2} \cdot \left[280 + \frac{1.96^2}{2} + 1.96 \cdot \sqrt{280 + \frac{1.96^2}{4}} \right] = 157.4$$

- Note that since the hypothesized parameter ($\lambda = 150$) lies within til 95% confidence interval, we accept the NULL hypothesis.

Geiger–Marsden Experiment

- In one of their experiments, Geiger and Marsden detected $x = 11571$ alpha particles (using a Geiger counter) over a time period of $t = 187776$ seconds. 52t 9min 36sek
- For illustration purposes only, let us assume that we hypothesize $\lambda = 0.060$.
- **Statistical model:**
- x = number of alpha particles detected = 11571
- $X \sim poisson(t \cdot \lambda)$, where t = 187776 seconds

Parameter estimate: $\hat{\lambda} = \frac{11571}{187776} = 0.06162$

Expected number of alpha particles: $\gamma = t \cdot \lambda = 11266$

Geiger–Marsden Experiment

Hypothesis test:

- H_0: $\lambda = 0.060$
- H_1: $\lambda \neq 0.060$

- Test size:

$$z = \frac{x - t \cdot \lambda}{\sqrt{t \cdot \lambda}} = \frac{11571 - 187776 \cdot 0.060}{\sqrt{187776 \cdot 0.060}} = 2.8682 \sim \mathcal{N}(0,1)$$

- Approximative p-value:

$$p = 2 \cdot |1 - \Phi(|z|)| = 2 \cdot |1 - \Phi(2.8682)| = 2 \cdot |1 - 0.9979| = 0.0041$$

- Since $p < 0.05$ we reject the null hypothesis and conclude that it is very unlikely that the true parameter is $\lambda = 0.060$.

Geiger–Marsden Experiment

- **95% confidence interval:**

$$\lambda_{-} = \frac{1}{t}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{x + \frac{1.96^2}{4}}\right] = \frac{1}{187776}\left[11571 + \frac{1.96^2}{2} - 1.96\sqrt{11571 + \frac{1.96^2}{4}}\right] = 0.0605$$

$$\lambda_{+} = \frac{1}{t}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{x + \frac{1.96^2}{4}}\right] = \frac{1}{187776}\left[11571 + \frac{1.96^2}{2} + 1.96\sqrt{11571 + \frac{1.96^2}{4}}\right] = 0.0628$$

- Note that since the hypothesized parameter (λ = 0.06) does <u>not</u> lie within the 95% confidence interval, we reject the null hypothesis.

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

Words and Concepts to Know

Chi-Square Distribution

Bernoulli Trial

Normal approximation

χ_k^2

Average rate

Poisson Distribution

Binominal Distribution

Critical values

Chi-Square Test