Introduction to Statistical Inference

한양대학교 김도영

1. Statistical Inference

- Parameter estimation :

  Estimating the value of parameters based on measured data

- Hypothesis testing :

  Method to decide whether the data at hand sufficiently support a particular hypothesis

  (Hypothesis : a statement about the parameters)

2. Parameter estimation

Unknown Parameter θ	Statistic $\hat{\Theta}$	Point Estimate $\hat{\theta}$
μ	$\bar{X} = \frac{\Sigma X_i}{n}$	$\bar{x}$
σ^2	$S^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$	s^2
p	$\hat{P} = \frac{X}{n}$	$\hat{p}$
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2 = \frac{\Sigma X_{1i}}{n_1} - \frac{\Sigma X_{2i}}{n_2}$	$\bar{x}_1 - \bar{x}_2$
$p_1 - p_2$	$\hat{P}_1 - \hat{P}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$	$\hat{p}_1 - \hat{p}_2$

- Parameter of interest θ
- Sample statistic $\widehat{\Theta}$
- Numerical value of the sample statistic $\hat{\theta}$
- Several choices are exist → Consider MSE

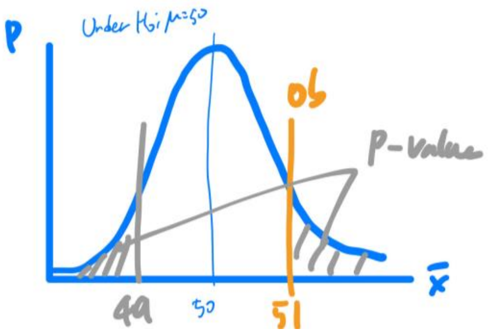
  Estimate the mean of a population

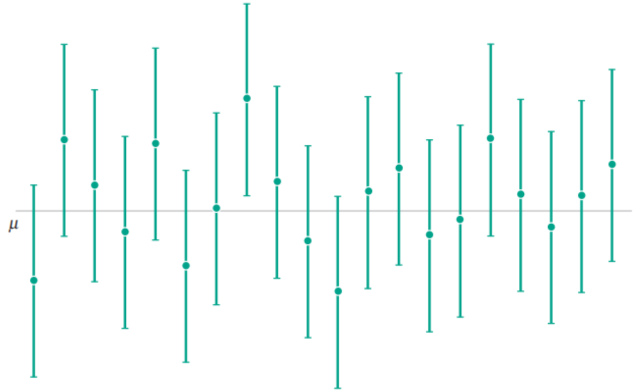
  → sample mean, sample median …

- $MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right] = Var(\hat{\theta}) + bias^2$

  $(\mathrm{Var}(\hat{\theta}) = E(\hat{\theta}^2) - [E(\hat{\theta})]^2, bias = E(\hat{\theta}) - \theta)$

3. Hypothesis testing





- Null hypothesis H_0 vs Alternative hypothesis H_1

  Deciding whether or not the mean burning rate is 50cm/s

  → H_0 : μ = 50 vs H_1 : μ ≠ 50.

- Evidence collection → Use P-value or Confidence Interval
- P-value :

  Probability computed under the condition that the **null hypothesis is true**, of the test statistic being at least as **extreme as** the value of the test statistic that was **actually observed**.

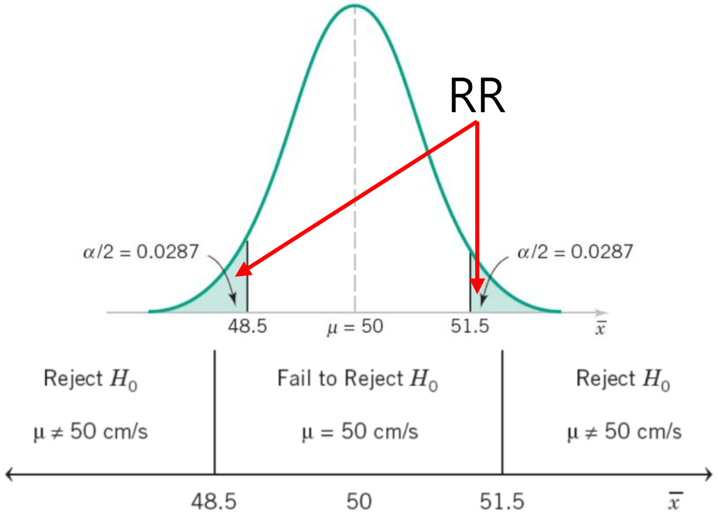
- Confidence Interval (CI) :

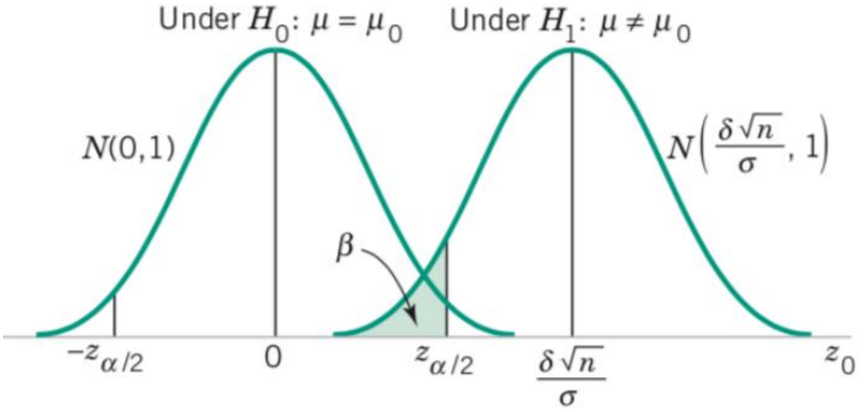
  If an infinite number of samples are collected, then $\mathbf{100(1-\alpha)\%}$ of **intervals (CI)** will **contain the true value** of the parameter.

  (α : significance level, $1-\alpha$: confidence coefficient)

3-1 Hypothesis testing using P-value

- Type I error α : Rejecting the null hypothesis H_0 when it is true
- Type II error β : Failing to reject the null hypothesis H_0 when it is not true
- If P-value is less than the significance level α, we would reject the null hypothesis
- Or use Rejection region RR, which is a set of values for the test statistic for which the null hypothesis is rejected





3-2 Steps of Hypothesis testing

- P-value

1. Establish H_0 and H_1
2. Calculate the test statistic
3. Compute P-value or Rejection Region(RR)
4. Compare P-value to α (if P>α, fail to reject H_0)

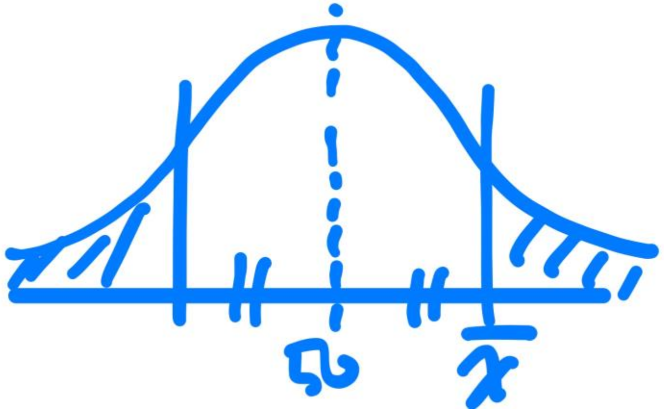
- Confidence Inverval

2. Choose confidence coefficient $1 - \alpha$
3. Construct $100(1 - \alpha)\%$ CI = [L,U] (upper and lover confidence bound for parameter)
4. If CI contain true population mean, the we fail to reject H_0

4. Example

- We want to **estimate** the mean burning rate $\boldsymbol{\mu}$.
- We **know** that the distribution of it is **normal** and $\boldsymbol{\sigma} = 2cm/s$.
- We selects a random sample of n=25 and **decided** to specify a type I error $\boldsymbol{\alpha} = 0.05$.
- We obtains a sample average $\bar{x} = 51.3cm/s$.
- **We want to know if mean burning rate is $\mathbf{50}\boldsymbol{cm/s}$ or not**

4. Example



Parameter of interest : mean burning rate μ

1. Hypothesis : $H_0 \colon \mu = 50$ vs $H_1 \colon \mu \neq 50$
2. Test statistic : $\bar{x}$ ($\bar{X} \sim N\left(50, \frac{2^2}{25}\right)$ under H_0) or $z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
3. P-value : 2*P($\bar{X} > \bar{x}$)=2*P($\mathrm{Z} > \frac{51.3-50}{2/\sqrt{25}}$)=0.0012
4. P-value = 0.0012 < $\alpha = 0.05$ → Reject H_0 at $\alpha = 0.05$

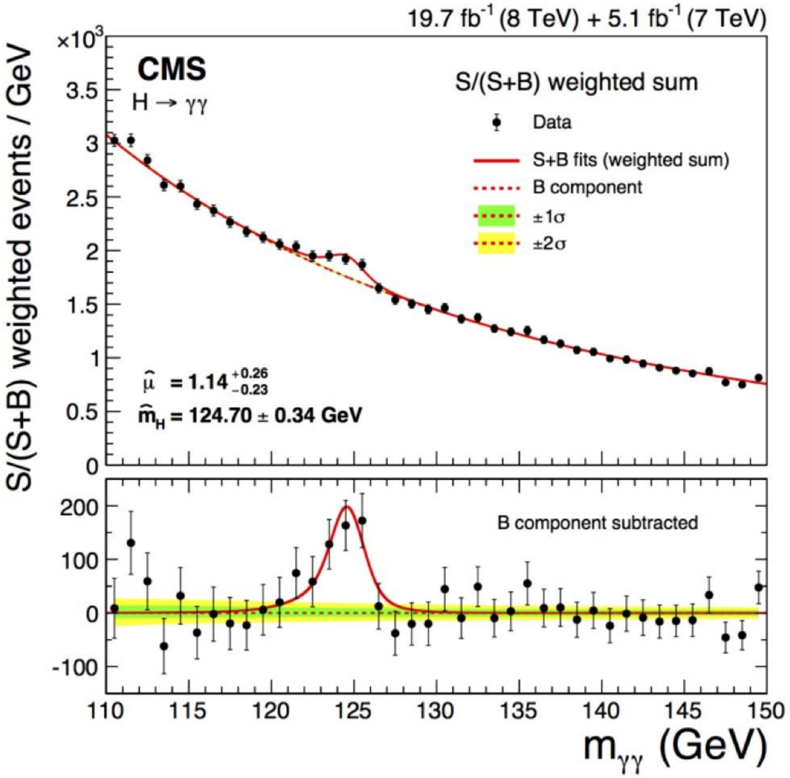
OR using Confidence interval for μ

1. a $100(1-\alpha)\%$ CI for μ is given by $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
2. For $\alpha = 0.05$, $z_{\alpha/2} = 1.96$, so 95% CI is $[L, U]$ = [50.52,52.08]
3. The value we observed(51.3) is not in CI → Reject H_0 at $\alpha = 0.05$

추정	Variance 알 때 μ	Variance 모를 때 μ	Normal population의 σ	비/, $\hat{P} = \frac{X}{n} \sim N(p, \frac{p(1-p)}{n})$	Goodness of fit: k 카테고리
Hypothesis	$H_0: \mu = \mu_0$ vs $H_1: \mu \neq >< \mu_0$ (From normal or n>=30)	= (무조건 정규)	$H_0: \sigma^2 = \sigma_0^2$ vs $H_1: \sigma^2 \neq \sigma_0^2$	H_0: $p=p_0$ vs H_1: $p \neq p_0$ (np>5, n(1-p)>5 or exact)	H_0 : $p_1=p_{10}$, ... $p_k=p_{k0}$ vs H_1 : not H_0
Test statistic	$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$	$X_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$	$Z_0 = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$	$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-p-1}$
분포	Standard normal	t-distribution, df=n-1	Chi-square, df=n-1, $\boldsymbol{\chi^2_{n-1}}$	Standard normal	Chi-square, df=n-k-1
P-value (2side/Upper/Lower)	$P(Z \geq \|z_0\|) + P(Z \leq \|z_0\|)$ $P(Z >< \|z_0\|)$ if $H_1 ><$	$2P(T \geq \|t_0\|)$ $P(T >< t_0)$ if $H_1 ><$	RR사용, $X_0^2 > \chi^2_{\frac{\alpha}{2},n-1}$ or $X_0^2 < \chi^2_{1-\alpha/2,n-1}$	$2P(Z \geq \|z_0\|)$ $P(Z >< z_0)$ if $H_1: p >< p_0$,	$P(\chi^2_{k-p-1} > \chi_0^2)$
100(1-α)% CI	$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	μ: $\bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$	$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$	p: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	
Upper/Lower Confidence Bound	$\boldsymbol{\mu < \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}}$, $\boldsymbol{\mu > \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}}$	μ<~, μ>~ 이때 α/2->α		p<~, p>~ 이때 α/2->α	

추정	Variance 알 때 두 평균 차이 $\mu_D = \mu_1 - \mu_2$ 정규 or n>=30	Variance 모를 때 두 평균μ 차이($\sigma_1^2 = \sigma_2^2$) 무조건 정규	Variance 모를 때 두 평균μ 차이($\sigma_1^2 \neq \sigma_2^2$)	Variance 비율(두 편차 같은지) σ_1^2/σ_2^2	두binomial에서 proportion p_1, p_2, $p_1 - p_2$ Estimator: $\hat{p}_1 = \frac{X_1}{n_1}, \hat{p}_2 = \frac{X_2}{n_2}$
Hypothesis	H_0 : $\mu_1-\mu_2=\Delta_0$ vs H_1: $\mu_1-\mu_2\neq\Delta_0$	H_0 : $\mu_1 - \mu_2 = \Delta_0$ vs H_1 : $\mu_1 - \mu_2 \neq \Delta_0$	=	$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$	H_0 : $p_1=p_2$
Test statistic	$Z_0 = \frac{\overline{X_1} - \overline{X_2} - \Delta_0}{\sqrt{(\sigma_1^2)/n_1 + (\sigma_2^2)/n_2}}$	$T_0 = \frac{\overline{X_1} - \overline{X_2} - \Delta_0}{S_p\sqrt{1/n_1 + 1/n_2}}$	$T_0 = \frac{\overline{X_1} - \overline{X_2} - \Delta_0}{\sqrt{(S_1^2)/n_1 + (S_2^2)/n_2}}$	$F_0 = \frac{S_1^2}{S_2^2} = F_{n_1-1,n_2-1}$	$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
분포	Standard normal	t-ditribution df= $n_1 + n_2 - 2$	t-ditribution, df= ν(못외움)	F-분포, df 2개	Standard
P-value (2side/Upper/Lower)	$2\,P(Z > \|z_0\|)$ $P(Z >< z_0)$ for $H_1: \mu_D >< \Delta_0$	$2\,P(T > \|t_0\|)$ $P(T >< t_0)$ for $H_1: \mu_D >< \Delta_0$		RR 사용, $f_0 > f_{\alpha/2,,n_1-1,n_2-1}$ or $f_0 < f_{1-\alpha/2,,n_1-1,n_2-1}$ And $f_0 >< f_{\alpha,,n_1-1,n_2-1}$	$2\,P(Z > \|z_0\|)$ $P(Z >< z_0)$ for $H_1: p_1 >< p_2$
100(1-α)% CI	$\mu_1 - \mu_2: \overline{x_1} - \overline{x_2}$ $\pm z_{\alpha/2}\sqrt{(\sigma_1^2)/n_1 + (\sigma_2^2)/n_2}$	$\mu_1 - \mu_2: \overline{x_1} - \overline{x_2}$ $\pm t_{\alpha/2,n_1+n_2-2}s_p\sqrt{1/n_1 + 1/n_2}$	$\mu_1 - \mu_2: \overline{x_1} - \overline{x_2}$ $\pm t_{\alpha/2,v}\sqrt{(s_1^2)/n_1 + (s_2^2)/n_2}$	$\frac{S_1^2}{S_2^2} f_{1-\alpha/2,n_2-1,n_1-1} < \frac{\sigma_1^2}{\sigma_2^2}$ $< \frac{S_1^2}{S_2^2} f_{\alpha/2,n_2-1,n_1-1}$	$\hat{p}_1 - \hat{p}_2$ $\pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
Upper/Lower Confidence Bound	$\mu_1 - \mu_2 < \sim, \mu_1 - \mu_2 > \sim$, 이때 α/2->α	$\mu_1 - \mu_2 < \sim, \mu_1 - \mu_2 > \sim$, 이때 α/2->α	μ<~, μ>~ 이때 α/2->α	σ<~, σ>~, 이때 α/2-> α 참고: $f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,u,v}}$	참고: $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$ pooled sample proportion)

5. Apply to Particle Physics

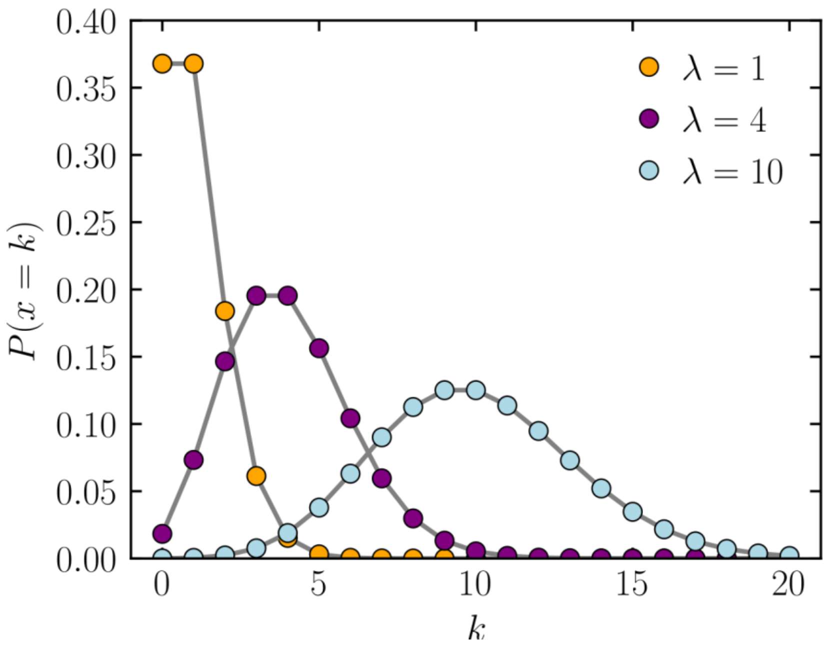


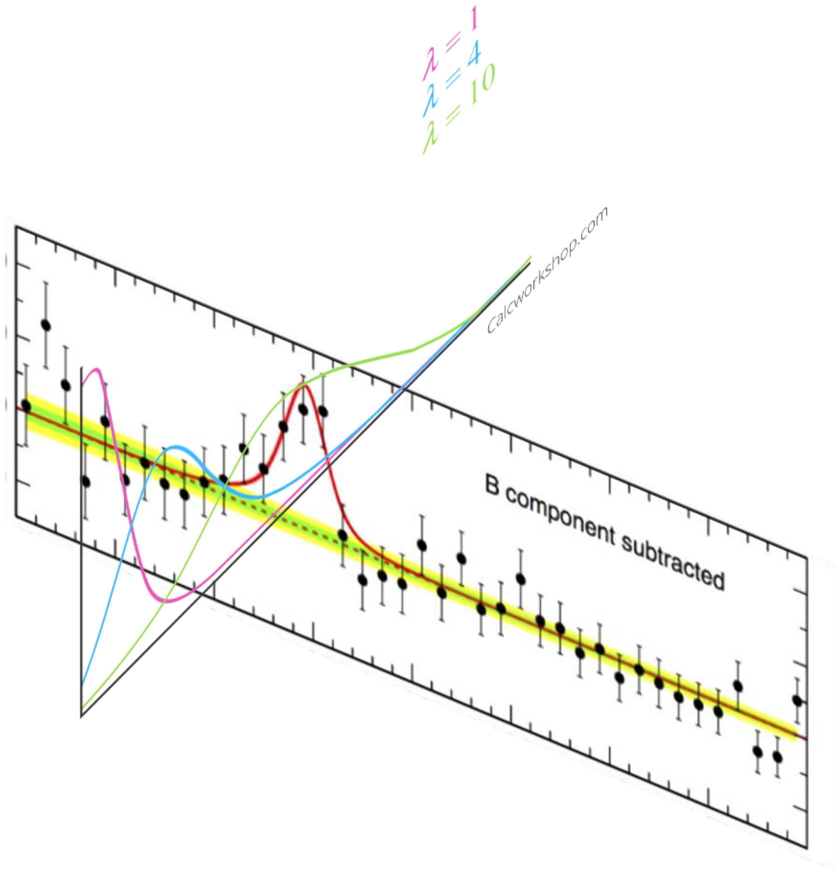
- **Data** has both **background**(b) and **signal**(s)
  → Data is given by **d=b+s** (the mean count)
- Goal : **Find** the mean Higgs boson event count **s**.

- Let's analyze the summary results of the measurement of the Higgs boson in the 4-lepton final states. (H→ZZ)

⇒ **N=25**(**observed 4-lepton events**) with **background estimate** of $\boldsymbol{B \pm \delta B = 9.4 \pm 0.5}$

- $\boldsymbol{H_0}$: **background-only** (**no signal**) vs $\boldsymbol{H_1}$: background **plus signal**


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
k
$P(x = k)$
$\lambda = 1$
$\lambda = 4$
$\lambda = 10$


$\lambda = 1$
$\lambda = 4$
$\lambda = 10$
Calcworkshop.com
B component subtracted

5. Apply to Particle Physics

- Distribution of d (events)

1. **Each collision** between protons is a **Bernoulli trial** (Higgs boson is created or not)
2. The collection of the Bernoulli trial can be represented by a **Binomial distribution** (pmf: $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$)
3. If $\lambda = np$ is fixed, then $\lim_{n \to \infty} f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$=**Poisson distribution**
4. Full model : $\boldsymbol{p(n, m|s, b) = Poisson(n, s+b) Poisson(m, kb)}$

Data: $p(n|s,b) = Poisson(n, s+b) = \frac{(s+b)^n e^{-(s+b)}}{n!}$, background: $p(m|kb) = Poissson(m, kb)$

5. Apply to Particle Physics

- Average of Poisson distribution = λ ($f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$)
- Variance of Poisson distribution = λ
- Let **N** is the **total number of observation** and **M** is the **number of background** observation(unknown) → M will be the average value
- $B \pm \delta B = 9.4 \pm 0.5$ and background = $p(M|kb) = Poissson(M, kb)$

$\rightarrow B = E(b) = \frac{M}{k}, \delta B^2 = Var(b) = \frac{1}{k^2} Var(kb) = \frac{M}{k^2}$

$\rightarrow B = \frac{M}{k}, \delta B = \frac{\sqrt{M}}{k} \rightarrow M = \left(\frac{B}{\delta B}\right)^2 = 353.4, k = \frac{B}{\delta B^2} = 37.6$

- Full likelihood : $p(D|s, b) = \frac{(s+b)^N e^{-(s+b)}}{N!} \frac{(kb)^M e^{-kb}}{\Gamma(M+1)} \equiv L(s, b)$ (D=N,M)

5. Apply to Particle Physics

- Construct Confidence interval of s

1. Use maximum likelihood estimates(MLE)

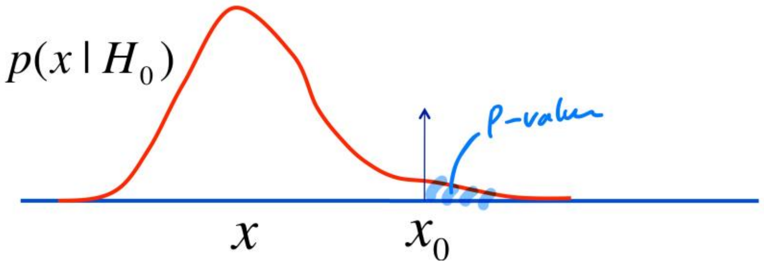
$\frac{\partial \ln p(D|s,b)}{\partial b} = 0 \rightarrow \hat{s} = N - b$, $\frac{\partial \ln p(D|s,b)}{\partial b} = 0 \rightarrow \hat{b} = \frac{N+M-(1+k)s+\sqrt{(N+M-(1+k)s)^2+4(1+k)Ms}}{2(1+k)}$ →We can use $L(s,b) = L(s)$

2. Let $\lambda(s) = \frac{L(s)}{L(\hat{s})}$ and $t(s) = -\mathbf{2\ln\lambda(s)} = t(\hat{s}+s-\hat{s}) \approx t(\hat{s}) + t'(\hat{s})(s-\hat{s}) + t''(\hat{s})(s-\hat{s})^2/2 \approx (s-\hat{s})^2/\sigma^2 \approx \boldsymbol{\chi_1^2}$ **(chi-square distribution)**

3. We know the distribution of s, so we can calculate the confidence interval using the table of the chi-square distribution

4. **For 68%CI = [10.9,21.0]** but N=25 is not in the CI, so we **reject the null hypothesis**. (=we observed the particle)

5. Apply to Particle Physics



- Use P-value

1. Compute P-value

$$\mathrm{p-value} = \sum_{k=N=25}^{\infty} Poisson(k, 9.4) = 1.76 * 10^{-5}$$

2. Use Z-value → $\mathrm{Z} = \sqrt{2}\,\mathrm{erf}^{-1}\big(1 - (\mathrm{p-value})\big) = \mathbf{4.14\sigma}$

3. If the p-value is **judged to be small enough**, the null hypothesis is rejected. (=we observed the particle)

References

- Montgomery, D. C., Runger, G. C., & Hubele, N. F. (2010). *Engineering Statistics*. Wiley.
- Prosper, H. B. (2019). *Practical Statistics for Particle Physicists.* CERN.