Hypothesis testing versus goodness-of-fit:

A statistics "question arising" from ATLAS' arXiv:1307.1432

after the Higgs talk, Geoff asked a question about $0^-/0^+$ discrimination:

Hypothesis testing versus goodness-of-fit:

A statistics "question arising" from ATLAS' arXiv:1307.1432

after the Higgs talk, Geoff asked a question about $0^-/0^+$ discrimination:

- on p53, the upper plot shows distributions of a BDT discriminator for

Hypothesis testing versus goodness-of-fit:

A statistics "question arising" from ATLAS' arXiv:1307.1432

after the Higgs talk, Geoff asked a question about $0^-/0^+$ discrimination:

- on p53, the upper plot shows distributions of a BDT discriminator for
  - the 0^+ hypothesis

Hypothesis testing versus goodness-of-fit:

A statistics "question arising" from ATLAS' arXiv:1307.1432

after the Higgs talk, Geoff asked a question about $0^-/0^+$ discrimination:

- on p53, the upper plot shows distributions of a BDT discriminator for
  - the 0^+ hypothesis
  - the 0^- hypothesis

Hypothesis testing versus goodness-of-fit:

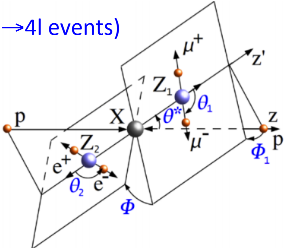
A statistics "question arising" from ATLAS' arXiv:1307.1432

after the Higgs talk, Geoff asked a question about $0^-/0^+$ discrimination:

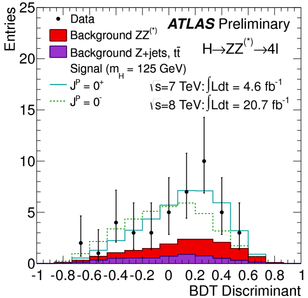
- on p53, the upper plot shows distributions of a BDT discriminator for
  - the 0^+ hypothesis
  - the 0^- hypothesis
  - the data

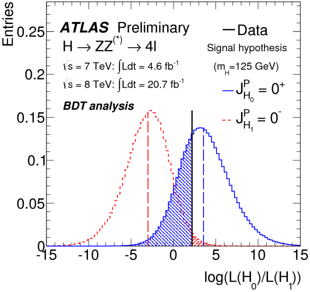
$J^P=0^-$ vs. $J^P=0^+$

(H→ZZ*→4l events)



- Sensitive variables:
  - Masses of the two Z bosons
  - Production angle θ^*
  - Four decay angles Φ_1, Φ, θ_1, θ_2





- Perform multivariate analysis (BDT)

$\boxed{\text{Exclude } J^P=0^- \text{ (vs. } 0^+\text{) with 97.8\% CL}}$

Hypothesis testing versus goodness-of-fit:

A statistics "question arising" from ATLAS' arXiv:1307.1432

after the Higgs talk, Geoff asked a question about $0^-/0^+$ discrimination:

- on p53, the upper plot shows distributions of a BDT discriminator for
  - the 0^+ hypothesis
  - the 0^- hypothesis
  - the data
- the data is "close" to the distributions for both hypotheses

Hypothesis testing versus goodness-of-fit:

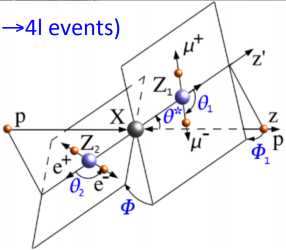
A statistics "question arising" from ATLAS' arXiv:1307.1432

after the Higgs talk, Geoff asked a question about $0^-/0^+$ discrimination:

- on p53, the upper plot shows distributions of a BDT discriminator for
  - the 0^+ hypothesis
  - the 0^- hypothesis
  - the data
- the data is "close" to the distributions for both hypotheses
- the lower plot shows the distributions of $\ln\left(\mathcal{L}(0^+)/\mathcal{L}(0^-)\right)$ for both hypotheses, and the value from data

$J^P=0^-$ vs. $J^P=0^+$

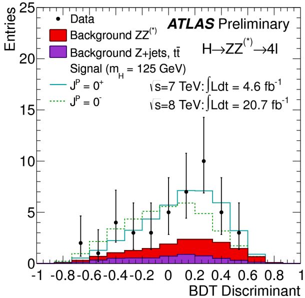
($H \rightarrow ZZ^* \rightarrow 4l$ events)

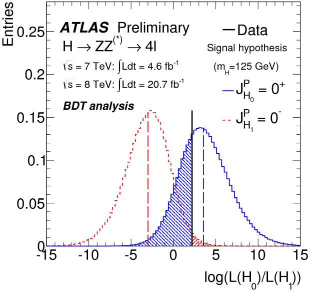


- Sensitive variables:
  - Masses of the two Z bosons
  - Production angle θ^*
  - Four decay angles Φ_1, Φ, θ_1, θ_2

- Perform multivariate analysis (BDT)

Exclude $J^P=0^-$ (vs. 0^+) with 97.8% CL





Hypothesis testing versus goodness-of-fit (1)

A statistics "question arising" from ATLAS' arXiv:1307.1432

after the Higgs talk, Geoff asked a question about $0^-/0^+$ discrimination:

- on p53, the upper plot shows distributions of a BDT discriminator for
  - the 0^+ hypothesis
  - the 0^- hypothesis
  - the data
- the data is "close" to the distributions for both hypotheses
- the lower plot shows the distributions of $\ln\left(\mathcal{L}(0^+)/\mathcal{L}(0^-)\right)$ for both hypotheses, and the value from data
- these show fairly strong discrimination between 0^+ and 0^-

Hypothesis testing versus goodness-of-fit (1)

A statistics "question arising" from ATLAS' arXiv:1307.1432

after the Higgs talk, Geoff asked a question about $0^-/0^+$ discrimination:

- on p53, the upper plot shows distributions of a BDT discriminator for
  - the 0^+ hypothesis
  - the 0^- hypothesis
  - the data
- the data is "close" to the distributions for both hypotheses
- the lower plot shows the distributions of $\ln\left(\mathcal{L}(0^+)/\mathcal{L}(0^-)\right)$ for both hypotheses, and the value from data
- these show fairly strong discrimination between 0^+ and 0^-
- there is an apparent contradiction between the two plots

Hypothesis testing versus goodness-of-fit (1)

A statistics "question arising" from ATLAS' arXiv:1307.1432

after the Higgs talk, Geoff asked a question about $0^-/0^+$ discrimination:

- on p53, the upper plot shows distributions of a BDT discriminator for
  - the 0^+ hypothesis
  - the 0^- hypothesis
  - the data
- the data is "close" to the distributions for both hypotheses
- the lower plot shows the distributions of $\ln\left(\mathcal{L}(0^+)/\mathcal{L}(0^-)\right)$ for both hypotheses, and the value from data
- these show fairly strong discrimination between 0^+ and 0^-
- there is an apparent contradiction between the two plots
- in fact, on closer inspection they appear to be consistent; the key is the difference between two related statistical tests:

Hypothesis testing versus goodness-of-fit (1)

A statistics "question arising" from ATLAS' arXiv:1307.1432

after the Higgs talk, Geoff asked a question about $0^-/0^+$ discrimination:

- on p53, the upper plot shows distributions of a BDT discriminator for
  - the 0^+ hypothesis
  - the 0^- hypothesis
  - the data
- the data is "close" to the distributions for both hypotheses
- the lower plot shows the distributions of $\ln\left(\mathcal{L}(0^+)/\mathcal{L}(0^-)\right)$ for both hypotheses, and the value from data
- these show fairly strong discrimination between 0^+ and 0^-
- there is an apparent contradiction between the two plots
- in fact, on closer inspection they appear to be consistent; the key is the difference between two related statistical tests:
  - *goodness-of-fit*, the thing we most often do by eye, and

Hypothesis testing versus goodness-of-fit (1)

A statistics "question arising" from ATLAS' arXiv:1307.1432

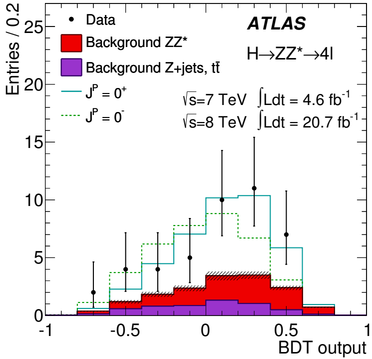
after the Higgs talk, Geoff asked a question about $0^-/0^+$ discrimination:

- on p53, the upper plot shows distributions of a BDT discriminator for
  - the 0^+ hypothesis
  - the 0^- hypothesis
  - the data
- the data is "close" to the distributions for both hypotheses
- the lower plot shows the distributions of $\ln\left(\mathcal{L}(0^+)/\mathcal{L}(0^-)\right)$ for both hypotheses, and the value from data
- these show fairly strong discrimination between 0^+ and 0^-
- there is an apparent contradiction between the two plots
- in fact, on closer inspection they appear to be consistent; the key is the difference between two related statistical tests:
  - *goodness-of-fit*, the thing we most often do by eye, and
  - *hypothesis testing*, which the ATLAS analysis is doing

Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

there are seven bins: after normalisation, possible datasets fill a 6D space

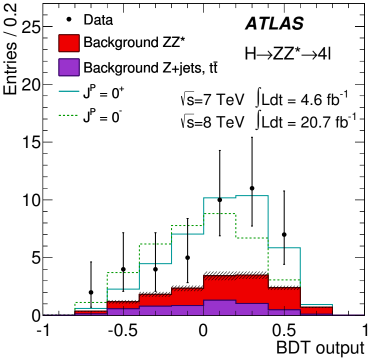


Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

there are seven bins: after normalisation, possible datasets fill a 6D space

> any test statistic corresponds to an *ordering principle* that assigns a number q to every point in the 6D space, collapsing it onto a line



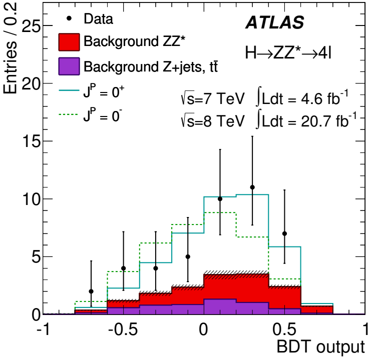
Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

there are seven bins: after normalisation, possible datasets fill a 6D space

> any test statistic corresponds to an *ordering principle* that assigns a number q to every point in the 6D space, collapsing it onto a line

- *goodness-of-fit* puts likely fluctuations up one end ("good"), and unlikely fluctuations at the other ("bad")



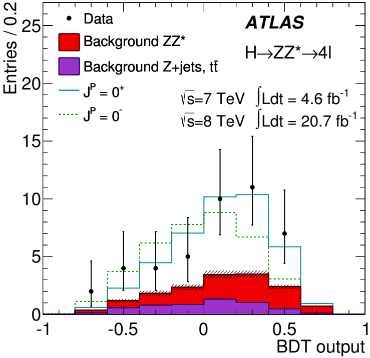
Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

there are seven bins: after normalisation, possible datasets fill a 6D space

> any test statistic corresponds to an *ordering principle* that assigns a number q to every point in the 6D space, collapsing it onto a line

- *goodness-of-fit* puts likely fluctuations up one end ("good"), and unlikely fluctuations at the other ("bad")
- g.o.f.(0^-) is done w/o reference to 0^+



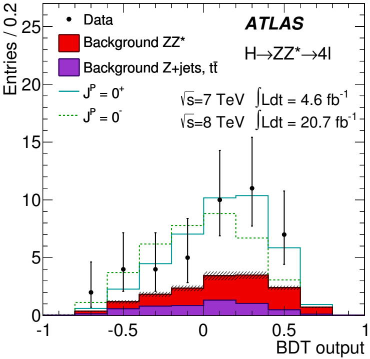
Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

there are seven bins: after normalisation, possible datasets fill a 6D space

> any test statistic corresponds to an *ordering principle* that assigns a number q to every point in the 6D space, collapsing it onto a line

- *goodness-of-fit* puts likely fluctuations up one end ("good"), and unlikely fluctuations at the other ("bad")
- g.o.f.(0^-) is done w/o reference to 0^+
- g.o.f.(0^+) likewise is blind to 0^-



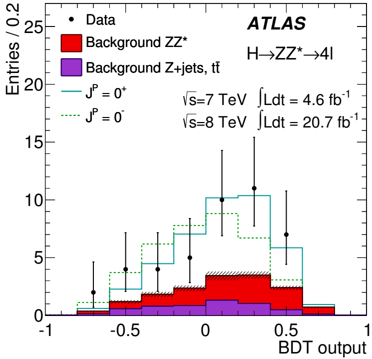
Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

there are seven bins: after normalisation, possible datasets fill a 6D space

> any test statistic corresponds to an *ordering principle* that assigns a number q to every point in the 6D space, collapsing it onto a line

- *goodness-of-fit* puts likely fluctuations up one end ("good"), and unlikely fluctuations at the other ("bad")
- g.o.f.(0^-) is done w/o reference to 0^+
- g.o.f.(0^+) likewise is blind to 0^-
- *hypothesis testing* instead assigns events 0^+-likeness $\longleftrightarrow$ 0^--likeness



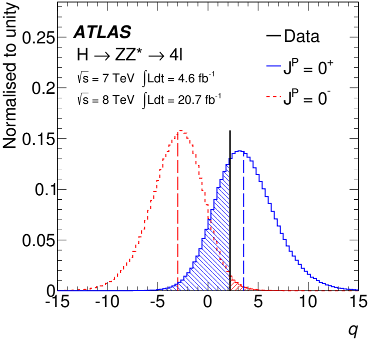
Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

there are seven bins: after normalisation, possible datasets fill a 6D space

> any test statistic corresponds to an *ordering principle* that assigns a number q to every point in the 6D space, collapsing it onto a line

- *goodness-of-fit* puts likely fluctuations up one end ("good"), and unlikely fluctuations at the other ("bad")
- g.o.f.(0^-) is done w/o reference to 0^+
- g.o.f.(0^+) likewise is blind to 0^-
- *hypothesis testing* instead assigns events 0^+-likeness $\longleftrightarrow$ 0^--likeness
- $\prod_i \mathcal{L}(0^+)/\mathcal{L}(0^-)$: 0^+-like-and-0^--unlike versus 0^--like-and-0^+-unlike



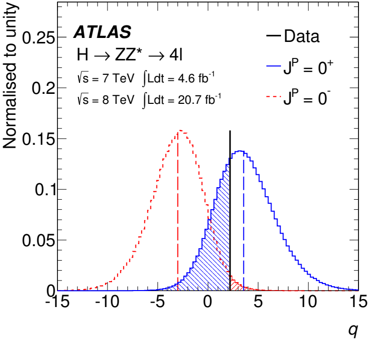
Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

there are seven bins: after normalisation, possible datasets fill a 6D space

> any test statistic corresponds to an *ordering principle* that assigns a number q to every point in the 6D space, collapsing it onto a line

- *goodness-of-fit* puts likely fluctuations up one end ("good"), and unlikely fluctuations at the other ("bad")
- g.o.f.(0^-) is done w/o reference to 0^+
- g.o.f.(0^+) likewise is blind to 0^-
- *hypothesis testing* instead assigns events 0^+-likeness $\longleftrightarrow$ 0^--likeness
- $\prod_i \mathcal{L}(0^+)/\mathcal{L}(0^-)$: 0^+-like-and-0^--unlike versus 0^--like-and-0^+-unlike



Neyman-Pearson theorem: this gives the *most powerful possible test*, i.e. the best direction in 6D (after some nonlinear transform^n)

Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

The example is simple enough that one can check the plot by eye, and calculate the result by hand (ignoring systematics)

Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

The example is simple enough that one can check the plot by eye, and calculate the result by hand (ignoring systematics)

bin	0^-	0^+	data	$\ln(\mathcal{L}(0^+)/\mathcal{L}(0^-))$
-0.7	1.0	0.6	2	
-0.5	3.8	2.25	4	
-0.3	6.2	4.5	4	
-0.1	7.8	7.0	5	
$+0.1$	8.9	10.2	10	
$+0.3$	6.6	10.4	11	
$+0.5$	3.1	5.9	7	

Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

The example is simple enough that one can check the plot by eye, and calculate the result by hand (ignoring systematics)

bin	0^-	0^+	data	$\ln(\mathcal{L}(0^+)/\mathcal{L}(0^-))$
-0.7	1.0	0.6	2	-0.622
-0.5	3.8	2.25	4	-0.546
-0.3	6.2	4.5	4	$+0.418$
-0.1	7.8	7.0	5	$+0.259$
$+0.1$	8.9	10.2	10	$+0.063$
$+0.3$	6.6	10.4	11	$+1.202$
$+0.5$	3.1	5.9	7	$+1.705$
				$+2.479$

Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

The example is simple enough that one can check the plot by eye, and calculate the result by hand (ignoring systematics)

bin	0^-	0^+	data	$\ln(\mathcal{L}(0^+)/\mathcal{L}(0^-))$
-0.7	1.0	0.6	2	-0.622
-0.5	3.8	2.25	4	-0.546
-0.3	6.2	4.5	4	$+0.418$
-0.1	7.8	7.0	5	$+0.259$
$+0.1$	8.9	10.2	10	$+0.063$
$+0.3$	6.6	10.4	11	$+1.202$
$+0.5$	3.1	5.9	7	$+1.705$
				$+2.479$

Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

The example is simple enough that one can check the plot by eye, and calculate the result by hand (ignoring systematics)

bin	0^-	0^+	data	$\ln(\mathcal{L}(0^+)/\mathcal{L}(0^-))$
-0.7	1.0	0.6	2	-0.622
-0.5	3.8	2.25	4	-0.546
-0.3	6.2	4.5	4	$+0.418$
-0.1	7.8	7.0	5	$+0.259$
$+0.1$	8.9	10.2	10	$+0.063$
$+0.3$	6.6	10.4	11	$+1.202$
$+0.5$	3.1	5.9	7	$+1.705$
				$+2.479$

- *cf.* ~ 2.1 in the ATLAS result, with systematics

Hypothesis testing versus goodness-of-fit (2)

A statistics "question arising" from ATLAS' arXiv:1307.1432

The example is simple enough that one can check the plot by eye, and calculate the result by hand (ignoring systematics)

bin	0^-	0^+	data	$\ln(\mathcal{L}(0^+)/\mathcal{L}(0^-))$
-0.7	1.0	0.6	2	-0.622
-0.5	3.8	2.25	4	-0.546
-0.3	6.2	4.5	4	$+0.418$
-0.1	7.8	7.0	5	$+0.259$
$+0.1$	8.9	10.2	10	$+0.063$
$+0.3$	6.6	10.4	11	$+1.202$
$+0.5$	3.1	5.9	7	$+1.705$
				$+2.479$

- *cf.* ~ 2.1 in the ATLAS result, with systematics
- note *luck in the actual dataset obtained is a factor*:
  are fluctuations along the privileged axis? in the right direction?

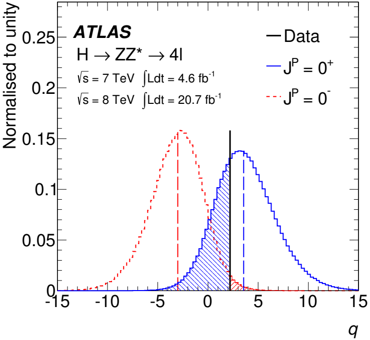
Hypothesis testing versus goodness-of-fit (recap)

A statistics "question arising" from ATLAS' arXiv:1307.1432

there are seven bins: after normalisation, possible datasets fill a 6D space

> any test statistic corresponds to an *ordering principle* that assigns a number q to every point in the 6D space, collapsing it onto a line

- *goodness-of-fit* puts likely fluctuations up one end ("good"), and unlikely fluctuations at the other ("bad")
- g.o.f.(0^-) is done w/o reference to 0^+
- g.o.f.(0^+) likewise is blind to 0^-
- *hypothesis testing* instead assigns events 0^+-likeness $\longleftrightarrow$ 0^--likeness
- $\prod_i \mathcal{L}(0^+)/\mathcal{L}(0^-)$: 0^+-like-and-0^--unlike versus 0^--like-and-0^+-unlike



Neyman-Pearson theorem: this gives the *most powerful possible test*, i.e. the best direction in 6D (after some nonlinear transform^n)

Hypothesis testing versus goodness-of-fit (3)

A statistics "question arising" from ATLAS' arXiv:1307.1432

There is a close analogy which requires no calculation:

Hypothesis testing versus goodness-of-fit (3)

A statistics "question arising" from ATLAS' arXiv:1307.1432

There is a close analogy which requires no calculation:

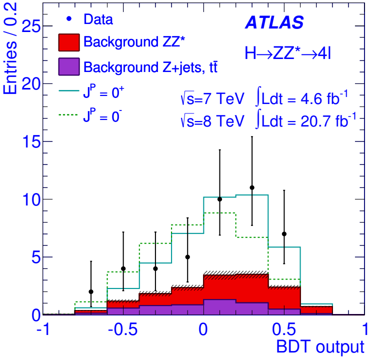
Gaussian measurements of an underlying constant

Hypothesis testing versus goodness-of-fit (3)

A statistics "question arising" from ATLAS' arXiv:1307.1432

There is a close analogy which requires no calculation:
Gaussian measurements of an underlying constant

- consider $N = 50\ \mathcal{G}(\mu, 1)$ meas^{ts}

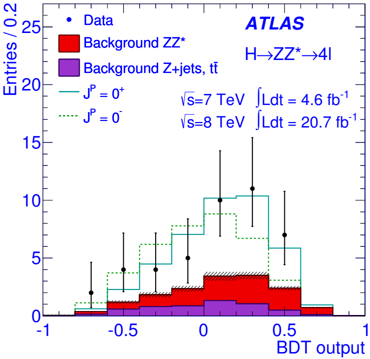


Hypothesis testing versus goodness-of-fit (3)

A statistics "question arising" from ATLAS' arXiv:1307.1432

There is a close analogy which requires no calculation:
Gaussian measurements of an underlying constant

- consider $N = 50\ \mathcal{G}(\mu, 1)$ meas^{ts}
- H_+: $\mu_+ = +0.2$

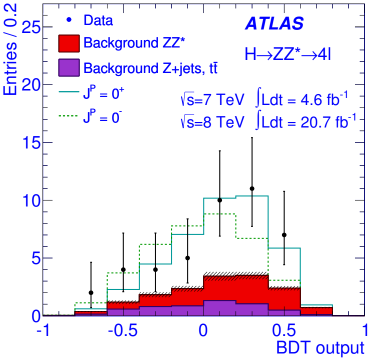


Hypothesis testing versus goodness-of-fit (3)

A statistics "question arising" from ATLAS' arXiv:1307.1432

There is a close analogy which requires no calculation:
Gaussian measurements of an underlying constant

- consider $N = 50\ \mathcal{G}(\mu, 1)$ meas^{ts}
- H_{+}: $\mu_{+} = +0.2$
- H_{-}: $\mu_{-} = -0.2$

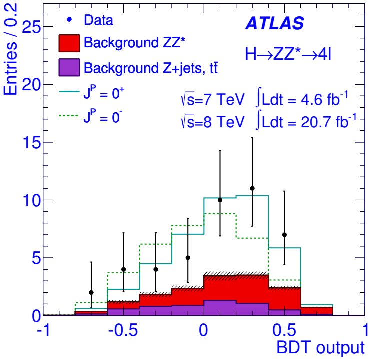


Hypothesis testing versus goodness-of-fit (3)

A statistics "question arising" from ATLAS' arXiv:1307.1432

There is a close analogy which requires no calculation:
Gaussian measurements of an underlying constant

- consider $N = 50\ \mathcal{G}(\mu, 1)$ meas^{ts}
- H_+: $\mu_+ = +0.2$
- H_-: $\mu_- = -0.2$
- Poisson errors on each bin (8–10 bins) will span difference in pred^{ns}, *cf.* $\rightarrow$

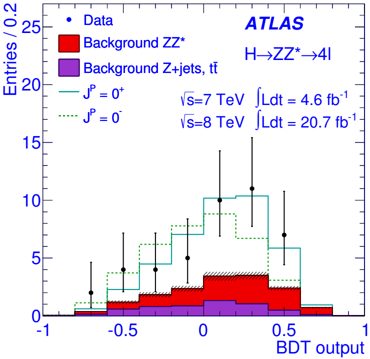


Hypothesis testing versus goodness-of-fit (3)

A statistics "question arising" from ATLAS' arXiv:1307.1432

There is a close analogy which requires no calculation:
Gaussian measurements of an underlying constant

- consider $N = 50\ \mathcal{G}(\mu, 1)$ meas^{ts}
- H_+: $\mu_+ = +0.2$
- H_-: $\mu_- = -0.2$
- Poisson errors on each bin (8–10 bins) will span difference in pred^{ns}, *cf.* $\rightarrow$
- $\sqrt{(V(\hat{\mu})} \sim 1/\sqrt{50} = 0.14$

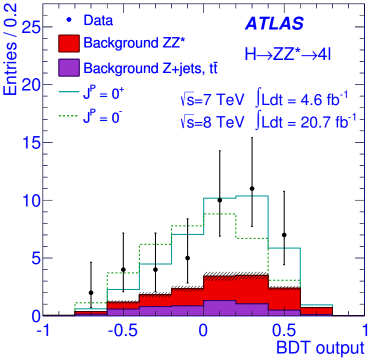


Hypothesis testing versus goodness-of-fit (3)

A statistics "question arising" from ATLAS' arXiv:1307.1432

There is a close analogy which requires no calculation:
Gaussian measurements of an underlying constant

- consider $N = 50\ \mathcal{G}(\mu, 1)$ meas^{ts}
- H_+: $\mu_+ = +0.2$
- H_-: $\mu_- = -0.2$
- Poisson errors on each bin (8–10 bins) will span difference in pred^{ns}, *cf.* $\rightarrow$
- $\sqrt{(V(\hat{\mu})} \sim 1/\sqrt{50} = 0.14$
- 50% of H_+ cases will fluctuate up: $\hat{\mu} > 0.2$; $z > 0.40/0.14 = 2.86$!

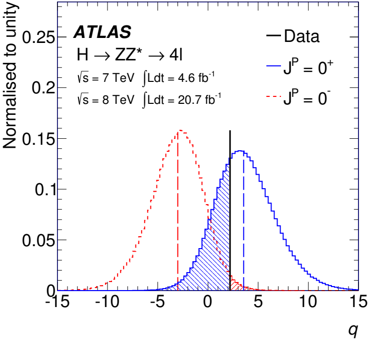


Hypothesis testing versus goodness-of-fit (3)

A statistics "question arising" from ATLAS' arXiv:1307.1432

There is a close analogy which requires no calculation:
Gaussian measurements of an underlying constant

- consider $N = 50\ \mathcal{G}(\mu, 1)$ meas^{ts}
- H_+: $\mu_+ = +0.2$
- H_-: $\mu_- = -0.2$
- Poisson errors on each bin (8–10 bins) will span difference in pred^{ns}, *cf.* $\rightarrow$
- $\sqrt{(V(\hat{\mu})} \sim 1/\sqrt{50} = 0.14$
- 50% of H_+ cases will fluctuate up: $\hat{\mu} > 0.2$; $z > 0.40/0.14 = 2.86$!



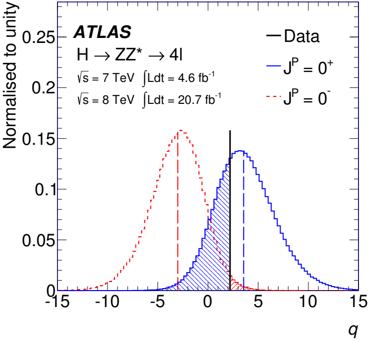
- likewise $\lesssim 10\%$ of H_+ cases will fluctuate to a H_--like result

Hypothesis testing versus goodness-of-fit (3)

A statistics "question arising" from ATLAS' arXiv:1307.1432

There is a close analogy which requires no calculation:
Gaussian measurements of an underlying constant

- consider $N = 50\ \mathcal{G}(\mu, 1)$ meas^{ts}
- H_+: $\mu_+ = +0.2$
- H_-: $\mu_- = -0.2$
- Poisson errors on each bin (8–10 bins) will span difference in pred^{ns}, *cf.* $\rightarrow$
- $\sqrt{(V(\hat{\mu})} \sim 1/\sqrt{50} = 0.14$
- 50% of H_+ cases will fluctuate up: $\hat{\mu} > 0.2$; $z > 0.40/0.14 = 2.86$!



- likewise $\lesssim$ 10% of H_+ cases will fluctuate to a H_--like result
- *i.e.* it may be essential to be correct, but it's also important to be lucky

Exercise: Show that the $\mathcal{L}/\mathcal{L}$ technique gives an equivalent answer.