Parametric hypothesis testing

Hypothesis testing: quick reminder

Hypothesis testing

sample: $X^n = (X_1, \dots, X_n), \mathrm{X} \sim F_X \in \Omega$

null hypothesis: $H_0: F_X \in \omega, \omega \in \Omega$

alternative hypothesis: $H_1: F_X \notin \omega$

statistic: $T(X^n) \sim F(x)$ when $F_X \in \omega$

$\nsim F(x)$ when $F_X \notin \omega$

Hypothesis testing

sample: $X^n = (X_1, \dots, X_n), \mathrm{X} \sim F_X \in \Omega$

null hypothesis: $H_0: F_X \in \omega, \omega \in \Omega$

alternative hypothesis: $H_1: F_X \notin \omega$

statistic: $T(X^n) \sim F(x)$ when $F_X \in \omega$
$\nsim F(x)$ when $F_X \notin \omega$

observed sample: $x^n = (x_1, \dots, x_n)$

observed statistic: $t = T(x^n)$

p-value: $p(t)$ – probability of getting $T = t$ or more extreme under H_0

Null hypothesis is rejected when $p \leq \alpha$,
α – significance level.

Type I and II errors

	H_0 true	H_0 false
H_0 accepted	👍	Type II error
H_0 rejected	Type I error	👍

Tests are constructed to bound the probability of type I error by significance level α.

Correct test: $P(p \leq \alpha | H_0) \leq \alpha \ \ \forall F_X$

Type I and II errors

	H_0 true	H_0 false
H_0 accepted	👍	Type II error
H_0 rejected	Type I error	👍

Tests are constructed to bound the probability of type I error by significance level α.

Correct test: $P(p \leq \alpha | H_0) \leq \alpha \ \ \forall F_X$

Power: $pow = P(p \leq \alpha | H_1)$
Consistent test: $pow \to 1$ with $n \to \infty$ for every false H_0

Interpretation

Low p – evidence against null hypothesis in favour of the alternative.

If p is not low enough, there is no evidence against null hypothesis in favour of the alternative.

Absence of evidence $\not\Rightarrow$ evidence of absence.

Using statistical tests, it is impossible to confirm null hypothesis!

Statistical and practical significance

The probability of rejecting wrong H_0 depends not only on how far it is from the truth, but on the sample size as well.

With large n, our statistic detects subtle departures from H_0 that might be just not interesting.

When testing hypothesis, one should always estimate an **effect size** – the magnitude of departure from null – and evaluate its practical significance.

Statistical and practical significance

- After 3 years women who exercised >1 hour per day on average gained significantly less weight than women who exercised <20 minutes per day ($p < 0.001$). The difference was 150 g. Lee et al, 2010

Statistical and practical significance

- After 3 years women who exercised >1 hour per day on average gained significantly less weight than women who exercised <20 minutes per day ($p < 0.001$). The difference was 150 g. Lee et al, 2010
- In 2002 a clinical trial of Premarin was terminated. It was discovered that its usage leads to significant increase of risks of breast cancer, stroke (0.08% up each), and heart attack (0.07% up). These effects are small, but practically significant. Ellis, 2010

Statistical and practical significance

- After 3 years women who exercised >1 hour per day on average gained significantly less weight than women who exercised <20 minutes per day ($p < 0.001$). The difference was 150 g. Lee et al, 2010
- In 2002 a clinical trial of Premarin was terminated. It was discovered that its usage leads to significant increase of risks of breast cancer, stroke (0.08% up each), and heart attack (0.07% up). These effects are small, but practically significant. Ellis, 2010
- If during a trial of a drug that slows down progression of the Alzheimer's disease the effect size – difference in average IQ loss between treatment and control groups – is found to be 13 points, but it is not statistically significant, the trial should be continued. Kirk, 1996

Hypothesis tests and confidence intervals

Point hypotheses H_0: $\theta = \theta_0$ could be tested with confidence intervals for θ:

- if θ_0 is outside an interval with confidence $1 - \alpha$, H_0 is rejected at level α
- p-value – maximal α for which θ_0 is inside the corresponding confidence interval

Hypothesis tests and confidence intervals

Point hypotheses H_0: $\theta = \theta_0$ could be tested with confidence intervals for θ:

- if θ_0 is outside an interval with confidence $1 - \alpha$, H_0 is rejected at level α
- p-value – maximal α for which θ_0 is inside the corresponding confidence interval

And the other way around:

- $100(1 - \alpha)\%$ confidence interval for θ consists of all θ^* that a test of H_0: $\theta = \theta^*$ against H_1: $\theta \neq \theta^*$ does not reject at level α

Testing hypotheses with likelihood

Likelihood

$$X^n = (X_1, \dots, X_n), \mathrm{X} \sim \mathrm{f}(x, \theta)$$

MLE for θ:

$$\ln L(X^n, \theta) = \sum_{i=1}^{n} \ln f(X_i, \theta)$$

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \ln L(X^n, \theta)$$

$$I(\theta) = -\frac{\partial^2}{\partial \theta^2} \ln L(X^n, \theta)$$

$$\mathbb{D}\hat{\theta}_{MLE} \approx I^{-1}\left(\hat{\theta}_{MLE}\right)$$

$$S(\theta) = \frac{\partial}{\partial \theta} \ln L(X^n, \theta)$$

$\hat{\theta}_{MLE}$ and $S\left(\hat{\theta}_{MLE}\right)$ are asymptotically normal

Wald test

sample: $X^n = (X_1, \dots, X_n), \mathrm{X} \sim F_X(x, \theta)$

null hypothesis: $H_0: \theta = \theta_0$

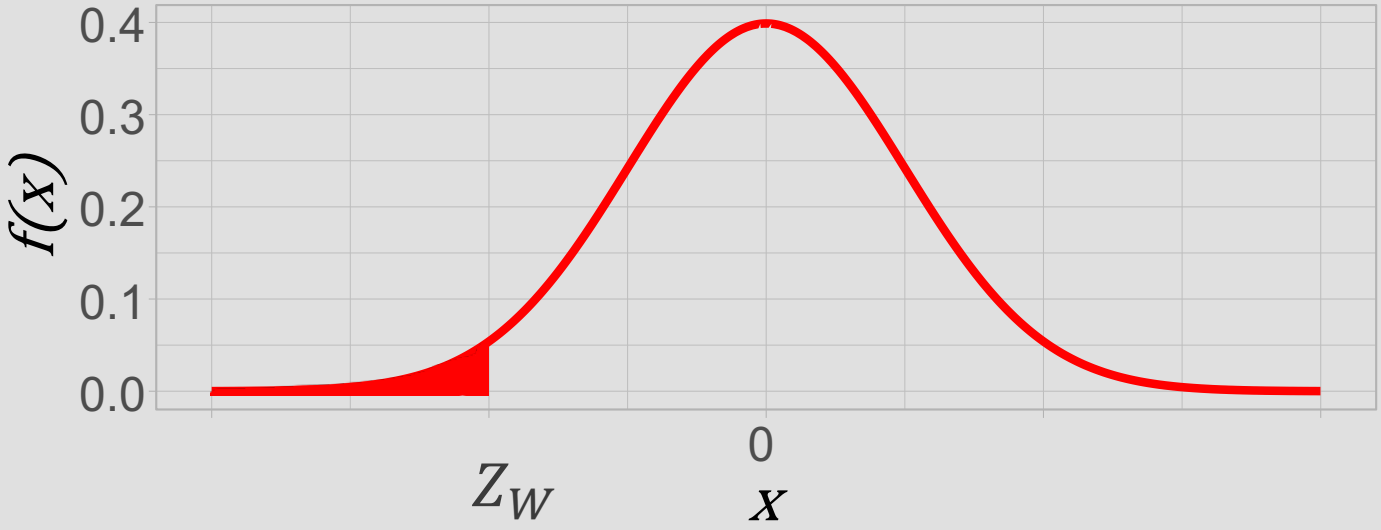
alternative hypothesis: $H_1: \theta <\neq> \theta_0$

statistic: $Z_W = \dfrac{\hat{\theta}_{MLE} - \theta_0}{\sqrt{\mathbb{D}\hat{\theta}_{MLE}}}$

null distribution: $N(0,1)$

P-values and alternatives

When $H_1: \theta < \theta_0$:

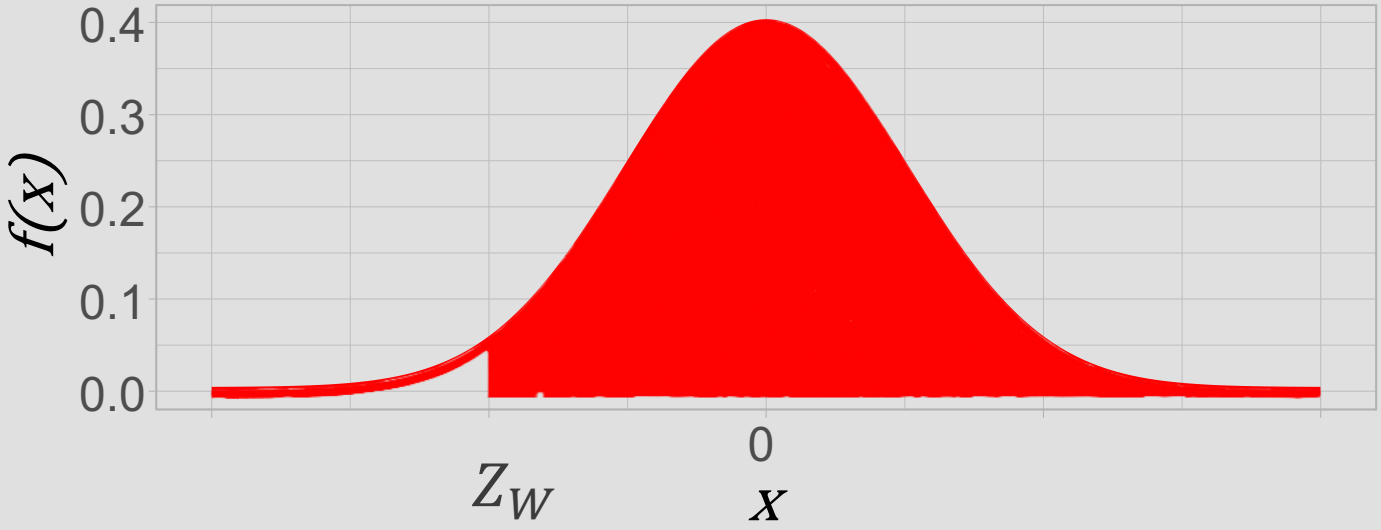


$$\frac{\hat{\theta}_{MLE} - \theta_0}{\sqrt{\mathbb{D}\hat{\theta}_{MLE}}}$$

$$p = \Phi(Z_W)$$

P-values and alternatives

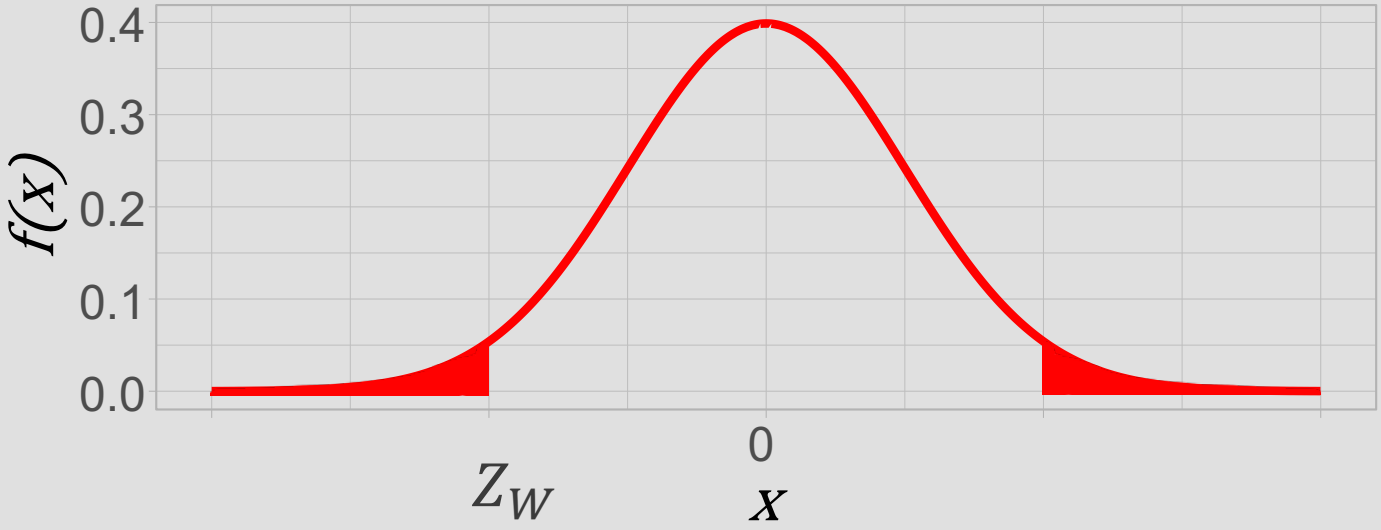
When $H_1: \theta > \theta_0$:



$$p = 1 - \Phi(Z_W)$$

P-values and alternatives

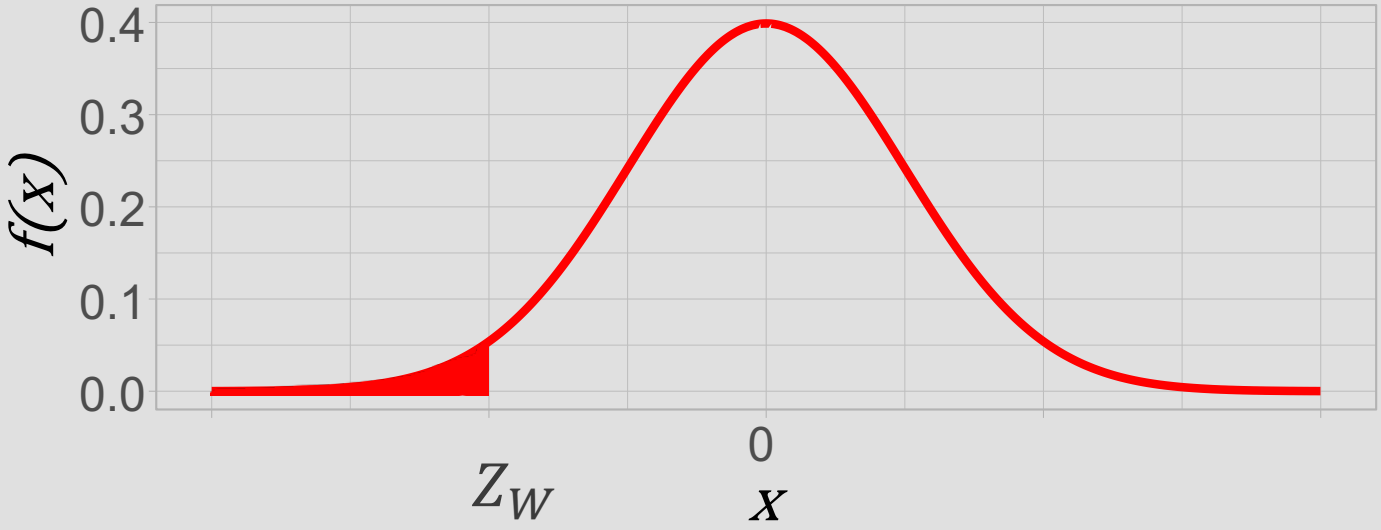
When $H_1: \theta \neq \theta_0$:



$$p = 2\Phi(-|Z_W|)$$

One-sided null hypothesis

When $H_1: \theta < \theta_0$, we could use the same procedure to test $H_0: \theta \geq \theta_0$ instead of $H_0: \theta = \theta_0$ without any changes.



$$p = \Phi(Z_W)$$

Score test

sample: $X^n = (X_1, \dots, X_n), \mathrm{X} \sim F_X(x, \theta)$

null hypothesis: $H_0: \theta = \theta_0$

alternative hypothesis: $H_1: \theta <\neq> \theta_0$

statistic: $Z_S = \dfrac{S(\theta_0)}{\sqrt{I(\theta_0)}}$

null distribution: $N(0,1)$

p-value: $p = \begin{cases} 1 - \Phi(Z_S), H_1: \theta > \theta_0 \\ \Phi(Z_S), \quad H_1: \theta < \theta_0 \\ 2\Phi(-|Z_S|), H_1: \theta \neq \theta_0 \end{cases}$

Likelihood ratio test

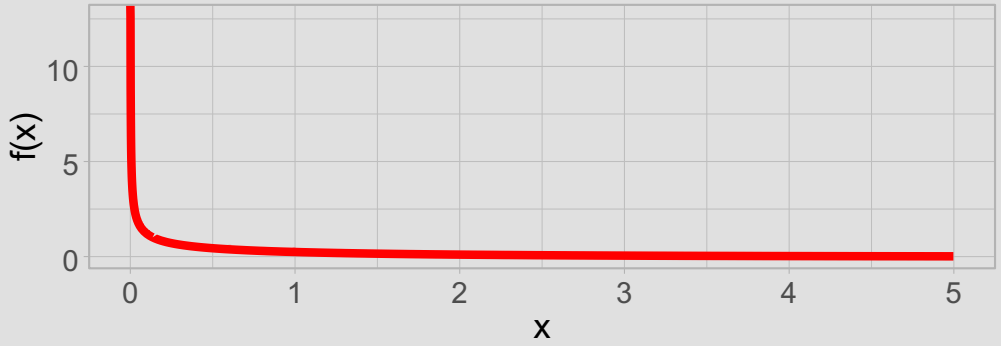
sample: $X^n = (X_1, \ldots, X_n), \mathrm{X} \sim F_X(x, \theta)$

null hypothesis: $H_0: \theta = \theta_0$

alternative hypothesis: $H_1: \theta \neq \theta_0$

statistic: $LR = -2\ln \dfrac{L(X^n, \theta_0)}{L\left(X^n, \hat{\theta}_{MLE}\right)}$

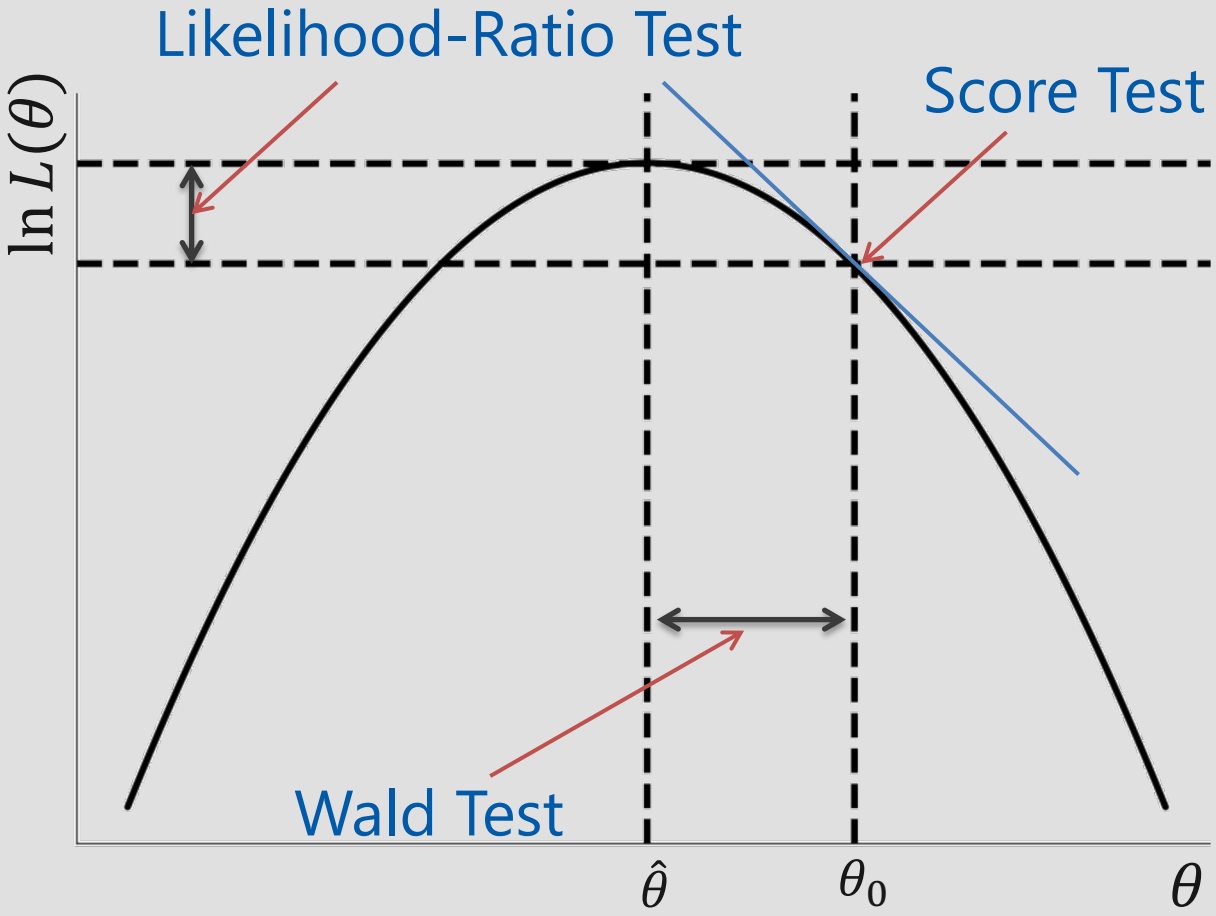
null distribution: χ^2_1



p-value: $p = 1 - F_{\chi^2_1}(LR)$

If θ has dimension k, null distribution has k degrees of freedom.

Three tests



- Wald test uses likelihood only at $\hat{\theta}_{MLE}$, score test – only at θ_0, likelihood ratio test – at both points.
- All three tests are asymptotic, but Wald's has the worst finite sample performance Agresti, A. Categorical Data Analysis

One proportion

Example: exercise

YouGov weekly poll of UK population:
How many times in the past week, if any, have you done 30 minutes or more of physical exercise?

First week of September 2020: of 2007 respondents, 522 said ‘None’.

Would it be accurate to say that a quarter of UK population didn’t exercise?

https://yougov.co.uk

Likelihood

$$X^n = (X_1, \dots, X_n), \mathrm{X} \sim Ber(p), \qquad k = \sum_{i=1}^{n} X_i$$

MLE for p:

$$L(X^n, p) = p^k (1-p)^{n-k}$$

$$\ln L(X^n, p) = k \ln p + (n-k) \ln(1-p)$$

$$\hat{p}_{MLE} = \frac{k}{n} \equiv \hat{p}$$

$$I(p) = \frac{n}{p(1-p)}$$

$$\mathbb{D}\hat{p} \approx \frac{\hat{p}(1-\hat{p})}{n}$$

$$S(p) = \frac{k}{p} - \frac{n-T}{1-p}$$

Tree tests

For testing $H_0: p = p_0$:

$$Z_W = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

$$Z_S = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$$LR = -2 \ln \frac{L(p_0)}{L(\hat{p})}$$

Exercise survey

$$\hat{p} = \frac{522}{2007} \approx 0.26$$

For testing $H_0: p = 0.25$ against $H_0: p \neq 0.25$:

$$Z_W = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx 1.03, p = 0.3028$$

$$Z_S = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx 1.04, p = 0.2965$$

$$LR = -2 \ln \frac{L(p_0)}{L(\hat{p})} \approx 1.08, p = 0.2987$$

$100(1-\alpha)\%$ confidence intervals

Wald:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example: vaccine adverse effects

543 healthy adults were vaccinated with ChAdOx1 nCoV-19 vaccine; none had serious adverse events in 4 weeks after.

What could be the risk of serious adverse events in the population?

Folegatti et al, 2020

Example: vaccine adverse effects

543 healthy adults were vaccinated with ChAdOx1 nCoV-19 vaccine; none had serious adverse events in 4 weeks after.

What could be the risk of serious adverse events in the population?

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

95% Wald interval: $\{0\}$
Does it mean the vaccine is guaranteed to be safe?

Folegatti et al, 2020

Example: vaccine adverse effects

543 healthy adults were vaccinated with ChAdOx1 nCoV-19 vaccine; assume 1 had serious adverse events in 4 weeks after.

What could be the risk of serious adverse events in the population?

Example: vaccine adverse effects

543 healthy adults were vaccinated with ChAdOx1 nCoV-19 vaccine; assume 1 had serious adverse events in 4 weeks after.

What could be the risk of serious adverse events in the population?

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

95% Wald interval: $[-0.0018, 0.0054]$
What do negative values even mean?

$100(1-\alpha)\%$ confidence intervals

Wald:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Could span outside $[0,1]$
- Turns into a point with $\hat{p} = 0$ or $\hat{p} = 1$

$100(1-\alpha)\%$ confidence intervals

Wilson (score):

$$\frac{k+\frac{1}{2}z_{1-\frac{\alpha}{2}}}{n+z_{1-\frac{\alpha}{2}}} \pm \frac{z_{1-\frac{\alpha}{2}}\sqrt{n}}{z^2_{1-\frac{\alpha}{2}}+n}\sqrt{\hat{p}(1-\hat{p})+\frac{z^2_{1-\frac{\alpha}{2}}}{4n}}$$

$100(1-\alpha)\%$ confidence intervals

Wilson (score):

$$\frac{k+\frac{1}{2}z_{1-\frac{\alpha}{2}}}{n+z_{1-\frac{\alpha}{2}}} \pm \frac{z_{1-\frac{\alpha}{2}}\sqrt{n}}{z^2_{1-\frac{\alpha}{2}}+n}\sqrt{\hat{p}(1-\hat{p})+\frac{z^2_{1-\frac{\alpha}{2}}}{4n}}$$

- Center is between $\hat{p}$ and $\frac{1}{2}$
- Looks a little bit more intimidating

$100(1-\alpha)\%$ confidence intervals

Wilson (score):

$$\frac{k+\frac{1}{2}z_{1-\frac{\alpha}{2}}}{n+z_{1-\frac{\alpha}{2}}} \pm \frac{z_{1-\frac{\alpha}{2}}\sqrt{n}}{z^2_{1-\frac{\alpha}{2}}+n}\sqrt{\hat{p}(1-\hat{p})+\frac{z^2_{1-\frac{\alpha}{2}}}{4n}}$$

- Center is between $\hat{p}$ and $\frac{1}{2}$
- Looks a little bit more intimidating

- 95% Wilson intervals for vaccine adverse effect risk:
  - for $k=0$: $[0, 0.007]$
  - for $k=1$: $[0.0003, 0.0104]$

Wald and Wilson intervals' coverage

Wald interval is **anticonservative** – has coverage lower than nominal!

Wald and Wilson intervals' coverage

Wald interval is **anticonservative** – has coverage lower than nominal!

When repeated many times on independent samples, $100(1-\alpha)\%$ of intervals would contain true parameter value

Wald and Wilson intervals' coverage

Wald interval is **anticonservative** – has coverage lower than nominal!

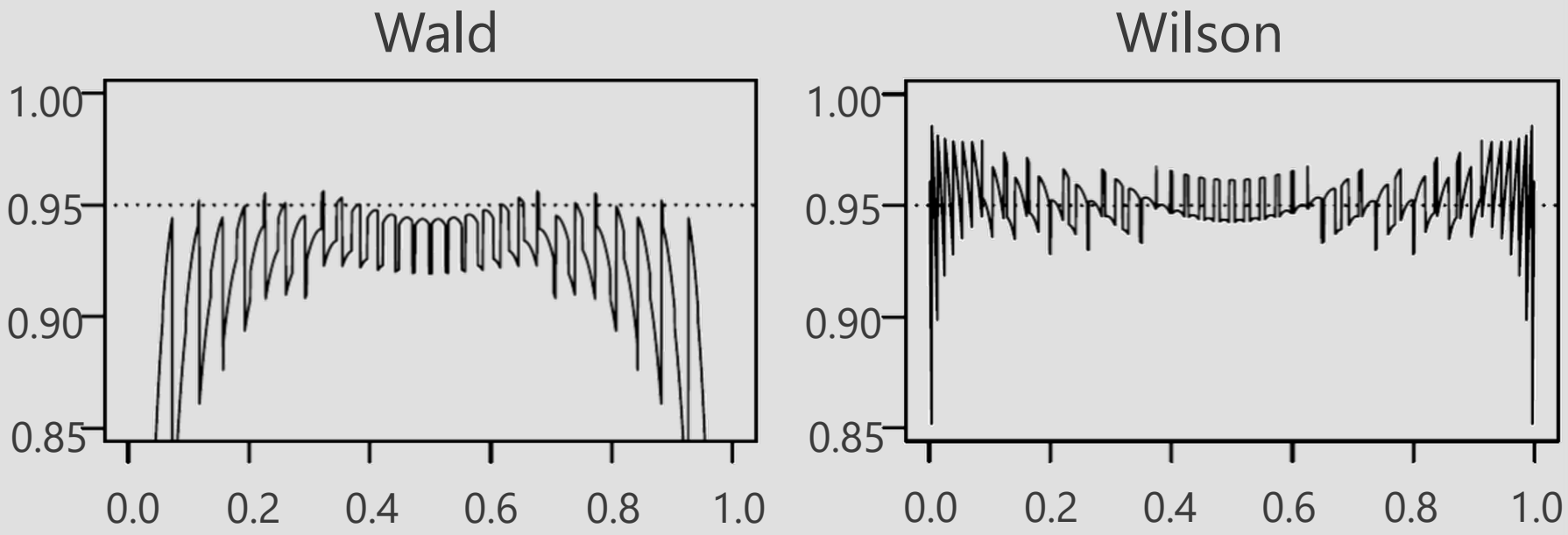
When repeated many times on independent samples, $100(1-\alpha)\%$ of intervals would contain true parameter value

Low coverage means that $100(1-\alpha)\%$ interval has confidence lower than $1-\alpha$

Wald and Wilson intervals' coverage

Wald interval is **anticonservative** – has coverage lower than nominal!

Simulation results for $n = 40$ and $\alpha = 0.05$:



Bilder, C.R., Loughin, T.M. Analysis of categorical data with R

What about LR confidence interval?

There's no analytical formula, but it could be constructed numerically – by finding smallest and largest p_0 such that $H_0: p = p_0$ is not rejected against $H_1: p \neq p_0$.

Two proportions: independent samples

Example: exercise

YouGov weekly poll of UK population:
How many times in the past week, if any, have you done 30 minutes or more of physical exercise?

First week of September 2020: of 2007 respondents, 522 said 'None'.
271 of them were from London, and 79 of them said 'None'.

Does London have the same proportion of people who did not exercise as the rest of the UK?

https://yougov.co.uk

Z test for a difference in proportions

samples: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), \mathrm{X}_1 \sim Ber(p_1)$

$X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), \mathrm{X}_1 \sim Ber(p_2)$

samples are independent

hypotheses: $H_0: p_1 = p_2,\ H_1: p_1 <\neq> p_2$

statistic: $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{P(1-P)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$

$k_1 = \sum_{i=1}^{n_1} X_{1i},\ \ k_2 = \sum_{i=1}^{n_2} X_{2i},$

$\hat{p}_1 = \frac{k_1}{n_1},\ \ \hat{p}_2 = \frac{k_2}{n_2},\ \ P = \frac{k_1+k_2}{n_1+n_2}$

null distribution: $N(0,1)$

Example: exercise

London:

$$\hat{p}_1 = \frac{79}{271} \approx 0.29$$

Rest of the UK:

$$\hat{p}_2 = \frac{522 - 79}{2007 - 271} = \frac{443}{1736} \approx 0.26$$

$H_0: p_1 = p_2$, $H_1: p_1 \neq p_2$:

$$Z \approx 1.27$$

$$p = 0.2048$$

Wald confidence interval for the difference

Directly equivalent to Z-test:

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\frac{\alpha}{2}} \sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Wald confidence interval for the difference

Directly equivalent to Z-test:

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\frac{\alpha}{2}} \sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- Also anticonservative!

Wilson confidence interval for the difference

$$[C_L, C_U] = [\hat{p}_1 - \hat{p}_2 - \delta, \hat{p}_1 - \hat{p}_2 + \varepsilon],$$

$$\delta = z_{1-\frac{\alpha}{2}}\sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}},$$

$$\varepsilon = z_{1-\frac{\alpha}{2}}\sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{n_2}},$$

l_1, u_1 – roots of the equation $|x - \hat{p}_1| = z_{1-\frac{\alpha}{2}}\sqrt{\frac{x(1-x)}{n_1}}$,

l_2, u_2 – roots of the equation $|x - \hat{p}_2| = z_{1-\frac{\alpha}{2}}\sqrt{\frac{x(1-x)}{n_2}}$

Wilson confidence interval for the difference

$$[C_L, C_U] = [\hat{p}_1 - \hat{p}_2 - \delta, \hat{p}_1 - \hat{p}_2 + \varepsilon],$$

$$\delta = z_{1-\frac{\alpha}{2}} \sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}},$$

$$\varepsilon = z_{1-\frac{\alpha}{2}} \sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{n_2}},$$

l_1, u_1 – roots of the equation $|x - \hat{p}_1| = z_{1-\frac{\alpha}{2}} \sqrt{\frac{x(1-x)}{n_1}}$,

l_2, u_2 – roots of the equation $|x - \hat{p}_2| = z_{1-\frac{\alpha}{2}} \sqrt{\frac{x(1-x)}{n_2}}$

- Has good coverage properties
- Could be numerically inverted to build hypothesis test!

Two proportions: paired samples

Example: prime minister's approval rating

In a poll of 1600 voting-age British citizens, 944 indicated approval of the prime minister's performance in the office. Followed-up 6 months later, of the same people only 880 indicated the approval.

I \ *II*	+	-	Σ
+	794	150	944
-	86	570	656
Σ	880	720	1600

Did approval rating change?

Agresti, A. Categorical data analysis

Z test for a difference in proportions, paired

$$X_1^n = (X_{11}, \dots, X_{1n}), \mathrm{X}_1 \sim Ber(p_1)$$
$$X_2^n = (X_{21}, \dots, X_{2n}), \mathrm{X}_1 \sim Ber(p_2)$$

samples:

X_1^n \ X_2^n	1	0
1	a	b
0	c	d

hypotheses: $H_0: p_1 = p_2,\ H_1: p_1 <\neq> p_2$

statistic: $$Z = \frac{b - c}{\sqrt{b + c - \frac{(b-c)^2}{n}}}$$

null distribution: $N(0,1)$

Example: prime minister's approval rating

I \ II	+	-	Σ
+	794	150	944
-	86	570	656
Σ	880	720	1600

$$\hat{p}_1 = \frac{944}{1600} \approx 0.59, \hat{p}_2 = \frac{880}{1600} \approx 0.55$$

$H_0: p_1 = p_2$, $H_1: p_1 \neq p_2$

Z test for paired samples: $p = 2.8 \times 10^{-5}$

Z test for independent samples: $p = 0.0222$

Wilson confidence interval for the difference

Let's just say it:

- Exists, but would not fit on this slide
- Has good coverage
- Also could be numerically inverted to test hypothesis

Wilson confidence interval for the difference

Let's just say it:

- Exists, but would not fit on this slide
- Has good coverage
- Also could be numerically inverted to test hypothesis

- For prime minister's rating:
  - using version for paired samples: $[-5.9, -2.1]$, corresponding p-value $p = 3.1 \times 10^{-5}$
  - using version for independent samples: $[-10.5, -3.7]$

Takeaways about proportions

- There are many ways to test hypotheses about proportions
- For each test there is a corresponding confidence interval
- The simplest way – based on Wald's method – is often slightly incorrect
- If Wilson's method is available, use it instead, it is never worse!

Normal means, one sample

Example: birth weight

Average birth weight in the US – 3300 g; for women living in poverty – 2800 g.

25 women living in poverty participated in new prenatal care program; hospital records show that the average birthweight of their babies was 3075 g with a standard deviation of 500 g.

Does the program work?

https://ccnmtl.columbia.edu/

Z test for a mean

sample: $X^n = (X_1, \dots, X_n), \mathrm{X} \sim N(\mu, \sigma^2)$, σ is known.

null hypothesis: $H_0: \mu = \mu_0$

alternative hypothesis: $H_1: \mu <\neq> \mu_0$

statistic: $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

null distribution: $N(0,1)$

T test for a mean

sample: $X^n = (X_1, \dots, X_n), \mathrm{X} \sim N(\mu, \sigma^2)$, σ is unknown.

null hypothesis: $H_0: \mu = \mu_0$

alternative hypothesis: $H_1: \mu <\neq> \mu_0$

statistic: $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$

null distribution: $St(n-1)$

T test for a mean

sample: $X^n = (X_1, \dots, X_n), \mathrm{X} \sim N(\mu, \sigma^2)$, σ is unknown.

null hypothesis: $H_0: \mu = \mu_0$

alternative hypothesis: $H_1: \mu <\neq> \mu_0$

statistic: $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$

null distribution: $St(n-1)$

The larger sample size is, the smaller the difference between t and z tests

T-test for a mean

For birthweight example:
the program works ⇔ birthweight for women in the program is higher than it would have been without it.

$H_0: \mu = 2800$, $H_1: \mu \geq 2800$

$$T = \frac{3075 - 2800}{500/\sqrt{25}} \approx 2.75,$$

$p = 0.0056.$

T-test for a mean

For birthweight example:
the program works ⇔ birthweight for women in the program is higher than it would have been without it.

$H_0: \mu = 2800$, $H_1: \mu \geq 2800$

$$T = \frac{3075 - 2800}{500/\sqrt{25}} \approx 2.75,$$

$p = 0.0056.$

Average birthweight significantly increases by 275 g ($p = 0.0056$, t-test with one-sided alternative of the increase, 95% confidence interval – $[234, 316]$ g).

T-test for a mean

For birthweight example:
the program works ⇔ birthweight for women in the program is higher than it would have been without it.

$H_0: \mu = 2800$, $H_1: \mu \geq 2800$

$$T = \frac{3075 - 2800}{500/\sqrt{25}} \approx 2.75,$$

$p = 0.0056.$

Average birthweight significantly increases by 275 g ($p = 0.0056$, t-test with one-sided alternative of the increase, 95% confidence interval – $[234, 316]$ g).

The standard way of reporting hypothesis testing results – include:

- test, alternative
- p-value
- effect size estimate and confidence interval for it

One sided alternative

Should only be used when the direction of change could be hypothesized in advance.

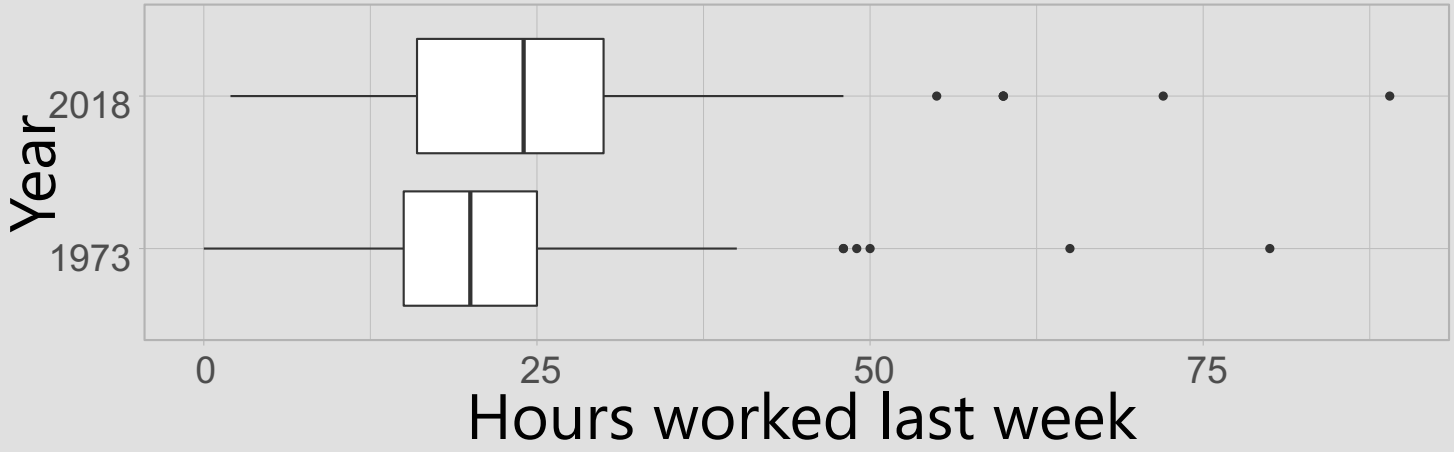
It's not OK to pick the side of the alternative after you've seen the data!

Normal means, two samples

Example: part time working hours

USA's General Social Survey is a sociological survey created and regularly collected since 1972 by the National Opinion Research Center at the University of Chicago.

In 1973, 134 of the respondents were working part-time, in 2018 – 255. For each of them we know the number of working hours in a week before the survey:



Did the average number of working hours change?

https://gssdataexplorer.norc.org/

T test for means of 2 independent samples

samples: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma_1^2)$
$X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_1 \sim N(\mu_2, \sigma_2^2)$
samples are independent
σ_1, σ_2 are unknown

hypotheses: $H_0: \mu_1 = \mu_2, H_1: \mu_1 <\neq> \mu_2,$

statistic: $$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

null distribution: $\approx St(\nu)$

T test for means of 2 independent samples

samples: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma_1^2)$

$X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_1 \sim N(\mu_2, \sigma_2^2)$

samples are independent

σ_1, σ_2 are unknown

hypotheses: $H_0: \mu_1 = \mu_2, H_1: \mu_1 <\neq> \mu_2,$

statistic: $$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Also known as Welch t-test

null distribution: $\approx St(\nu)$

T test for means of 2 independent samples

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1 - 1)} + \frac{S_2^4}{n_2^2(n_2 - 1)}}$$

T test for means of 2 independent samples

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1 - 1)} + \frac{S_2^4}{n_2^2(n_2 - 1)}}$$

- Null distribution is approximate, not exact! There is no exact solution (Behrens–Fisher problem)
- The approximation is very accurate when $n_1 = n_2$ or $[n_1 > n_2] = [\sigma_1 > \sigma_2]$

T test for means of 2 independent samples

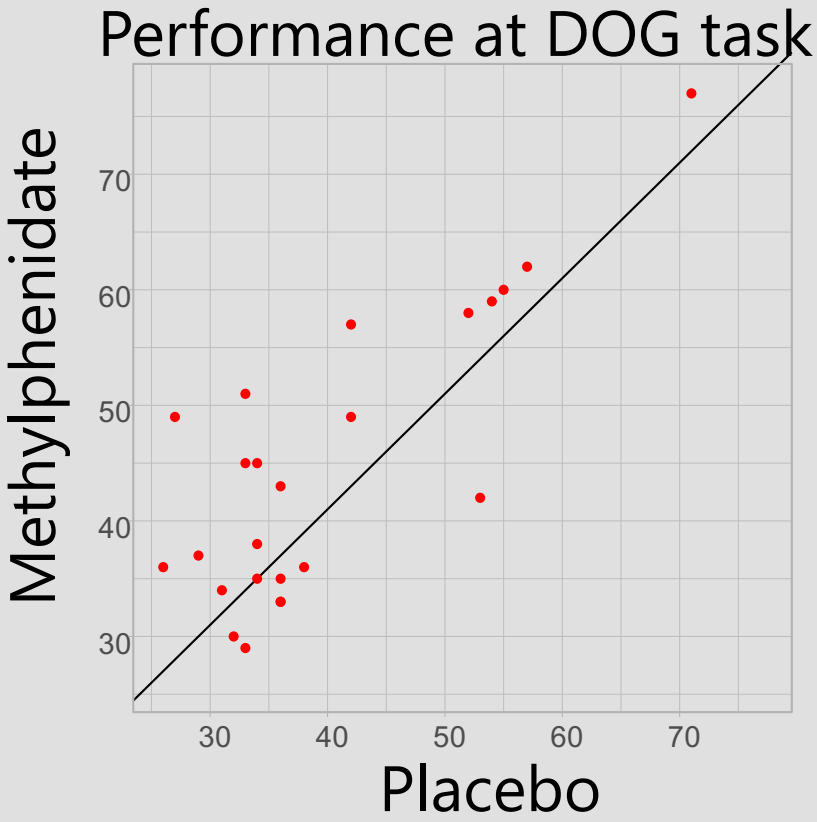
$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1 - 1)} + \frac{S_2^4}{n_2^2(n_2 - 1)}}$$

- Null distribution is approximate, not exact! There is no exact solution (Behrens–Fisher problem)
- The approximation is very accurate when $n_1 = n_2$ or $[n_1 > n_2] = [\sigma_1 > \sigma_2]$

- For GSS example: the average number of working hours for part-time workers increased significantly by 2.6 hours ($p = 0.0276$ for two-sided alternative, 95% confidence interval for the increase – $[0.29, 4.95]$ hours).

Example: ADHD treatment

24 children with ADHD had their performance measured on delay of gratification (DOG) task 60 minutes after taking Methylphenidate, and, on a different week, after taking placebo.



Does Methylphenidate work?

Pearson et al, 2003

T test for means of 2 paired samples

samples: $X_1^n = (X_{11}, \dots, X_{1n}), \mathrm{X}_1 \sim N(\mu_1, \sigma_1^2)$
$X_2^n = (X_{21}, \dots, X_{2n}), \mathrm{X}_1 \sim N(\mu_2, \sigma_2^2)$
samples are paired

hypotheses: $H_0: \mu_1 = \mu_2, H_1: \mu_1 <\neq> \mu_2,$

statistic: $T = \frac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{n}},$
$S = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2, D_i = X_{1i}\text{-}X_{1i}$

null distribution: $St(n-1)$

T test for means of 2 paired samples

samples: $X_1^n = (X_{11}, \dots, X_{1n}), X_1 \sim N(\mu_1, \sigma_1^2)$
$X_2^n = (X_{21}, \dots, X_{2n}), X_1 \sim N(\mu_2, \sigma_2^2)$
samples are paired

hypotheses: $H_0: \mu_1 = \mu_2, H_1: \mu_1 <\neq> \mu_2,$

statistic: $T = \frac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{n}},$
$S = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2, D_i = X_{1i} - X_{1i}$

null distribution: $St(n-1)$

Same as one sample t test on pairwise differences

T test for means of 2 paired samples

- For ADHD example: the average score on delay of gratification test after taking Methylphenidate is significantly higher by 4.95 points ($p = 0.0038$ for two-sided alternative, 95% confidence interval for the increase – $[1.78, 8.14]$ points).

Takeaways about normal means

- T tests for every configuration
- Comparing means for paired samples is the same as comparing mean of pairwise differences to 0
- If you have additional information about variances, you could build a slightly more powerful test
- Samples do not have to be from normal distributions! If sample sizes are large, and population distributions are fairly symmetric, T test works by CLT

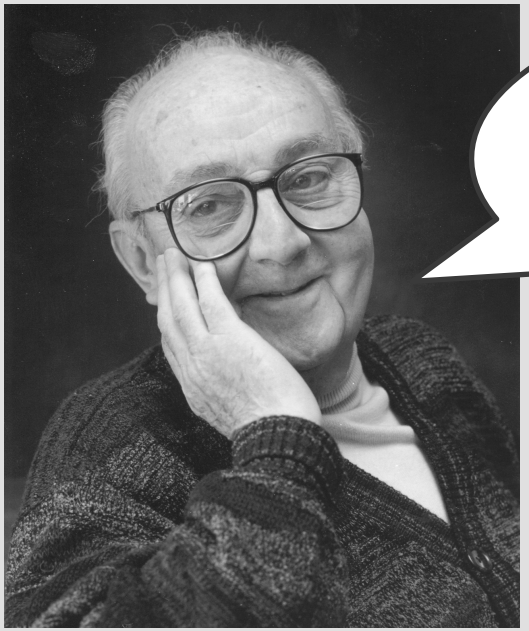
Goodness-of-fit tests

Is $X \sim F$?

- Does my sample really come from $X \sim F$?
- Probably not, very few data generating processes follow a distribution exactly

Is $X \sim F$?

- Does my sample really come from $X \sim F$?
- Probably not, very few data generating processes follow a distribution exactly

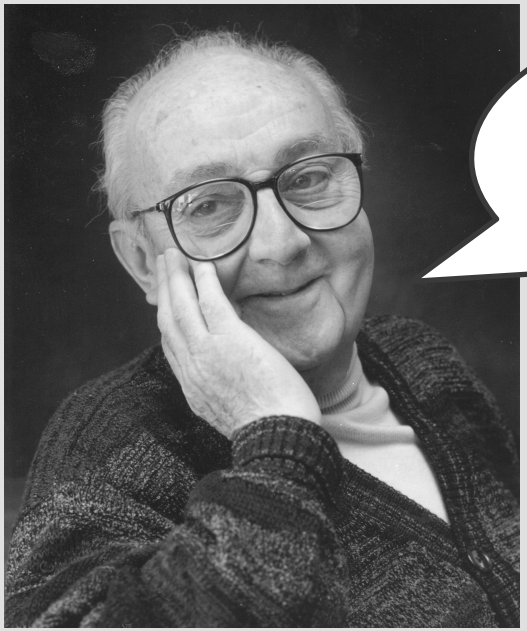


«All models are wrong, but some are useful»

George Box

Is $X \sim F$?

- Does my sample really come from $X \sim F$?
- Probably not, very few data generating processes follow a distribution exactly

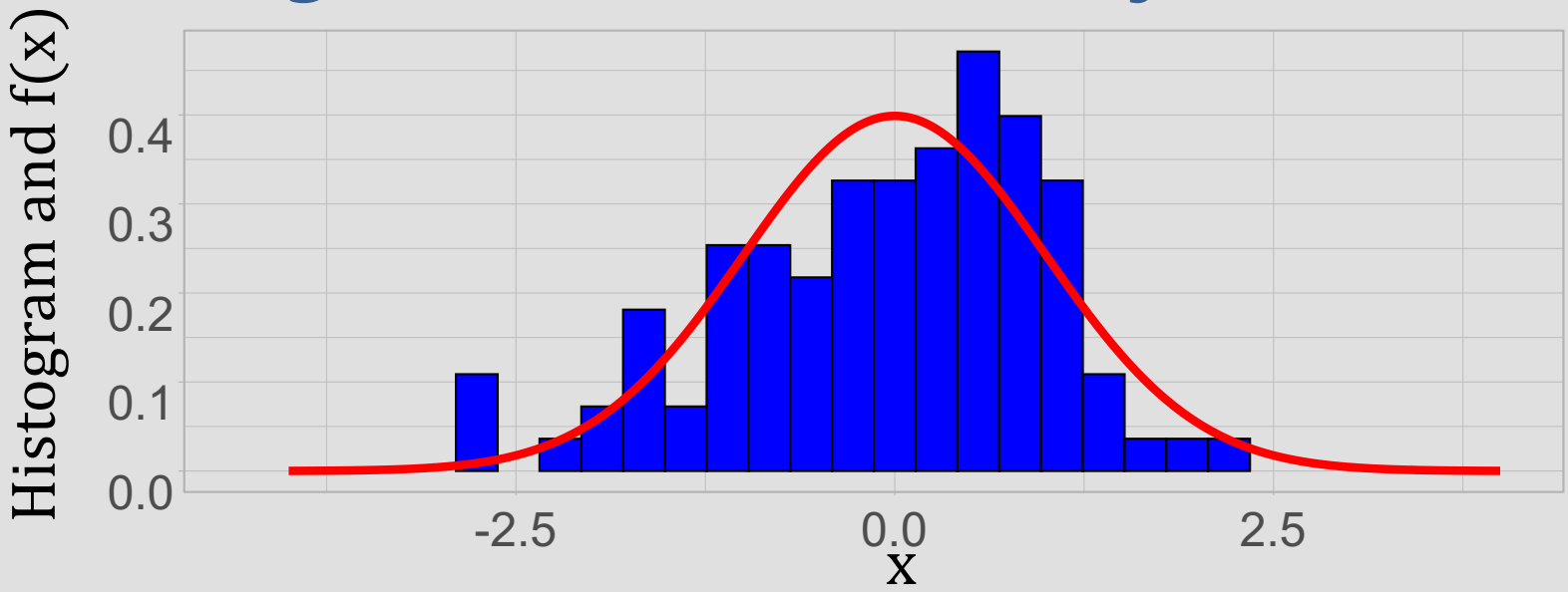


«All models are wrong, but some are useful»

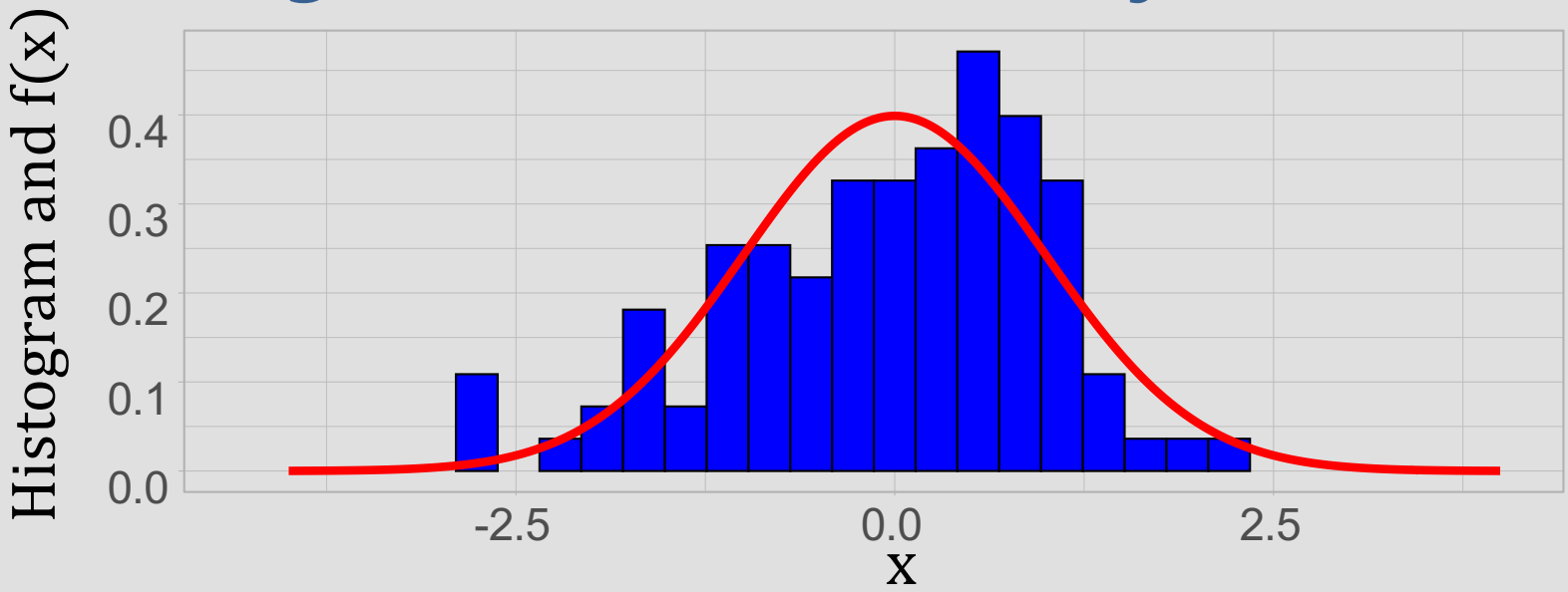
George Box

- Is it sensible to approximate my population with F_X?

Does histogram look like density?



Does histogram look like density?



$[a_{i-1}, a_i], i = 1, \dots, K$ – histogram bins

O_i – number of observations in the ith bin

$p_i = F(a_i) - F(a_{i-1})$ – theoretical probability of getting an observation in the ith interval assuming $X \sim F(x)$

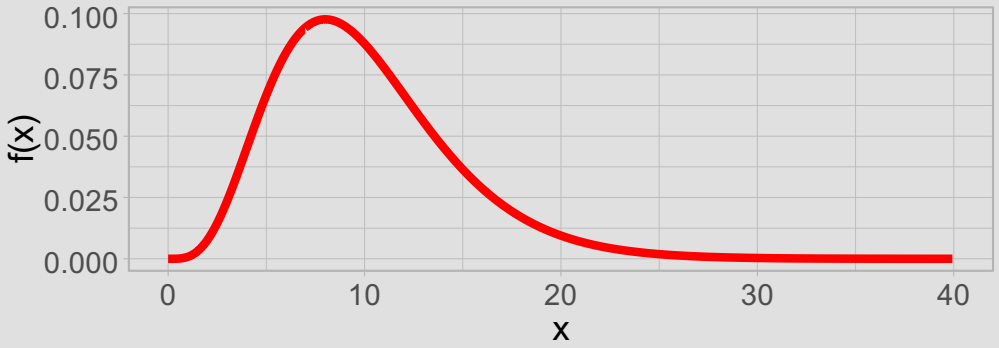
Pearson's chi-squared test

sample: $X^n = (X_1, \dots, X_n)$

hypotheses: $H_0: X \sim F(x), H_1: X \nsim F(x)$

statistic: $\chi^2 = \sum_{i=1}^{K} \frac{(O_i - np_i)^2}{np_i}$

null distribution: χ^2_{K-p}, p – number of fitted parameters



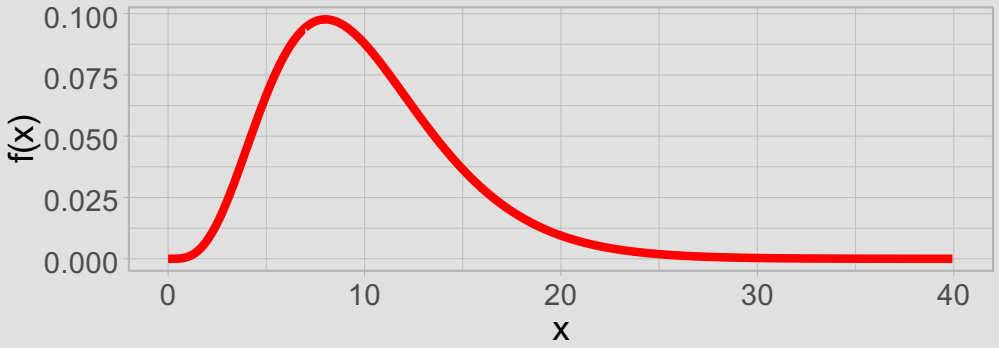
Pearson's chi-squared test

sample: $X^n = (X_1, \dots, X_n)$

hypotheses: $H_0: X \sim F(x), H_1: X \nsim F(x)$

statistic: $\chi^2 = \sum_{i=1}^{K} \frac{(O_i - np_i)^2}{np_i}$

null distribution: χ^2_{K-p}, p – number of fitted parameters

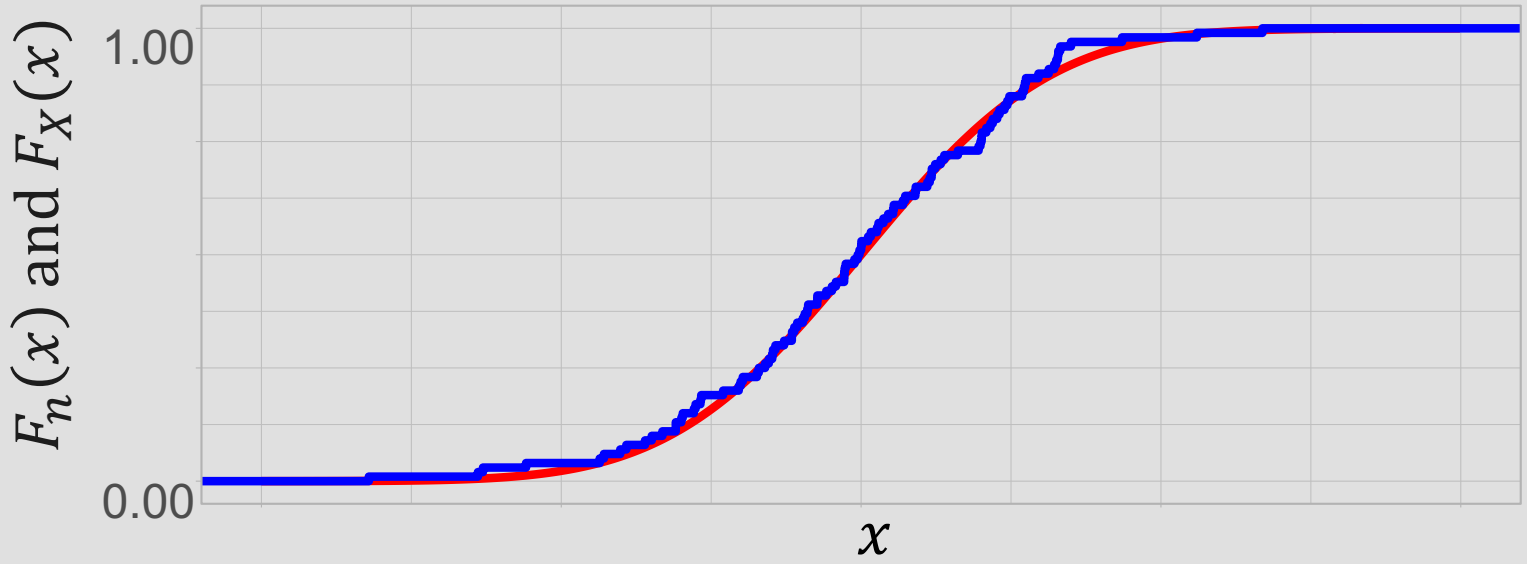


If you test for normality, but don't know μ and σ, then $p = 2$

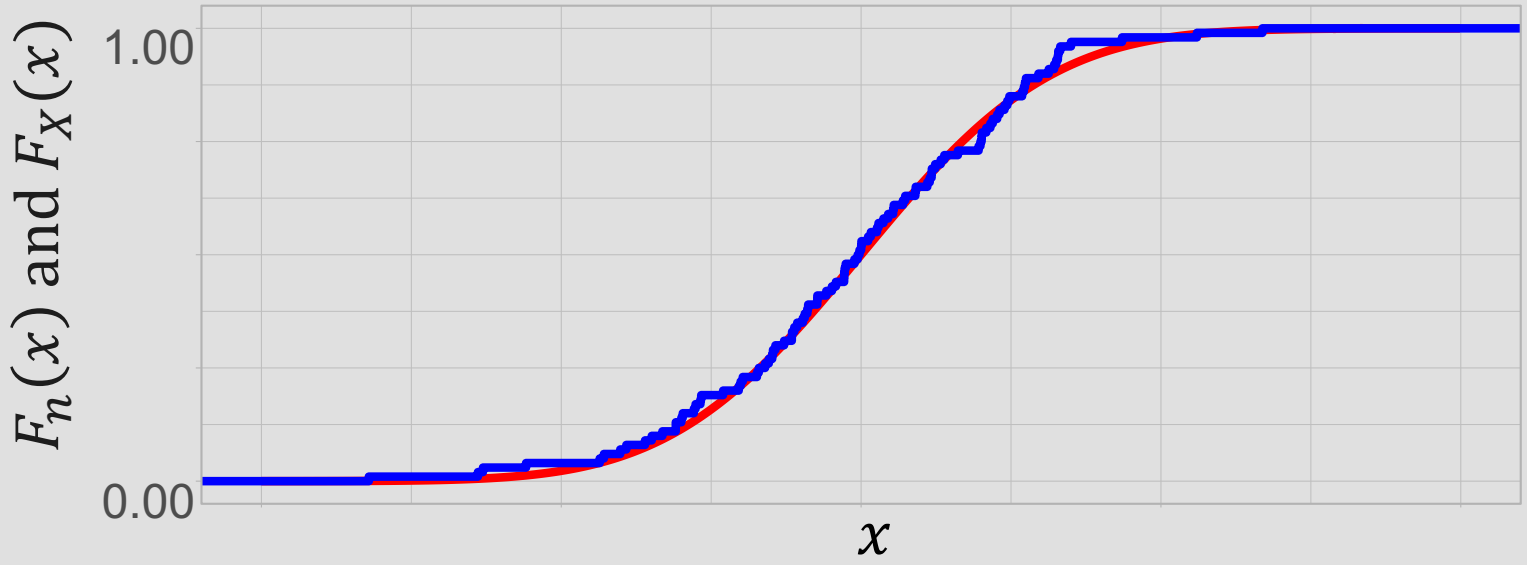
Pearson's chi-squared test

- Binning is arbitrary
- Requires $np_i > 5$ in 80% of bins, and no bins with $p_i = 0$

Does ecdf look like cdf?



Does ecdf look like cdf?



- Kolmogorov-Smirnov test:

$$\sup_{-\infty<x<\infty} |F_n(x) - F(x)|$$

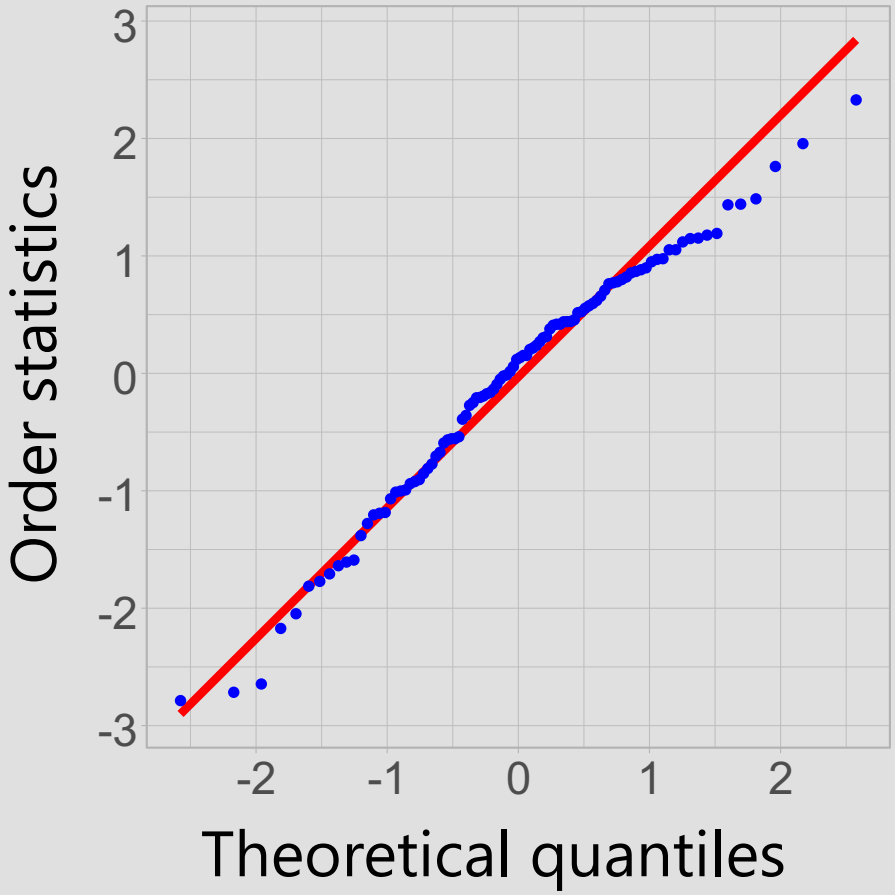
Low power!

- Anderson-Darling test:

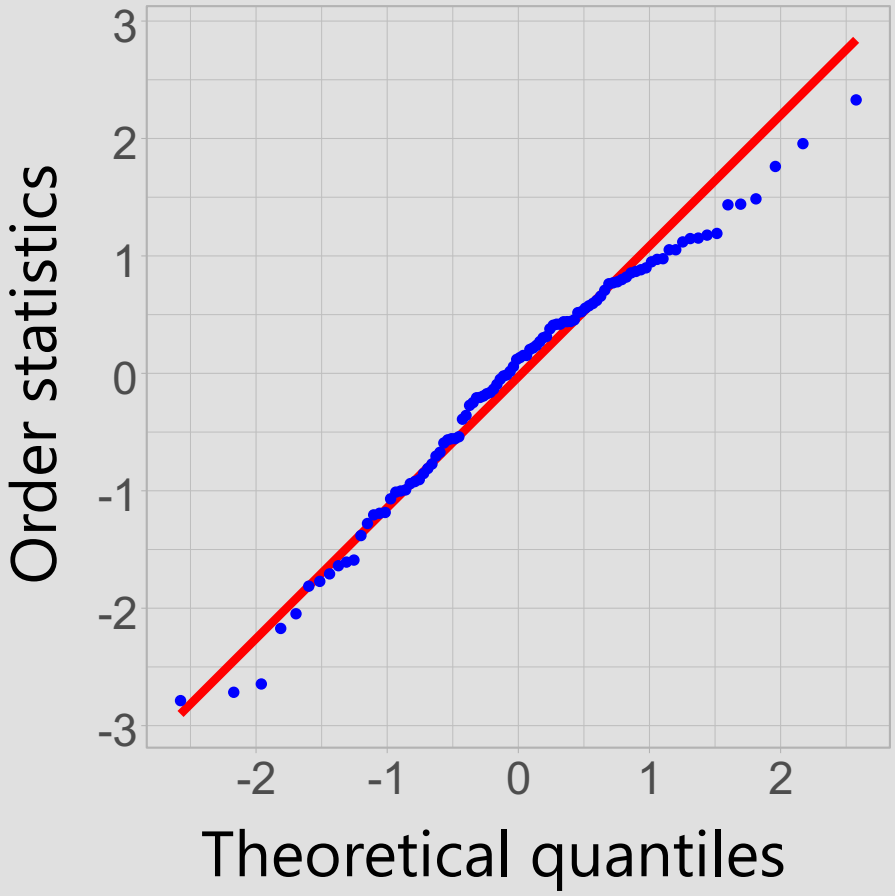
$$n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x)$$

Null distributions are tabulated

Does q-q plot look linear?



Does q-q plot look linear?



For $H_0: X \sim N$, linearity is formally tested with Shapiro-Wilk test

Takeaways about GoF tests

- Many ways to test whether $X \sim F$
- But it's very important to keep in mind why are you doing it
- Some tests are easy to generalize to two samples – testing

$$H_0\colon F_{X_1} = F_{X_2}, \qquad H_1\colon F_{X_1} \neq F_{X_2}$$

E.g., Kolmogorov-Smirnov, Anderson-Darling tests switch to measuring distance between two ecdfs