# STA261 - Module 4

## Intervals and Model Checking

Rob Zimmerman

University of Toronto

July 23-25, 2024

# Uncertainty in Point Estimates

- In Module 2, we learned how to produce the "best" point estimates of $\theta$ possible using statistics of our data

- The "best" unbiased estimator $\hat{\theta}(\mathbf{X})$ is the one that has the lowest possible variance among all unbiased estimators of $\theta$

- But even so, suppose we observe $\mathbf{X} = \mathbf{x}$ and calculate $\hat{\theta}(\mathbf{x})$; how do we know this is close to the true $\theta$? We don't!

- We can't know for sure

- But we can use the data to get a range of *plausible* values of $\theta$

Eg: UofT heights $\sim N(\mu, 1)$, $\mu \in \mathbb{R}$. Suppose we calculate $\hat{\mu}_{MLE}(\vec{x}) = \bar{X}_n = 5'6''$

It's probably more plausible that the true $\mu$ (in feet) is in $(5, 6)$ than $(2, 4)$

# Random Sets

- Suppose for now that $\Theta \subseteq \mathbb{R}$

a set which is a function of the random sample $\vec{X}$ (eg, $(\bar{X}_n - 1, \bar{X}_n + 1)$)

- If $\hat{\theta}(\mathbf{X})$ is a continuous random variable, then $\mathbb{P}_\theta\left(\theta = \hat{\theta}(\mathbf{X})\right) = 0$ Useless...

- But we can try to find a random set $C(\mathbf{X}) \subseteq \mathbb{R}$ based on $\mathbf{X}$ such that $\mathbb{P}_\theta\left(\theta \in C(\mathbf{X})\right) = 0.95$, for example

- Example 4.1: Let $X \sim \mathcal{N}(\mu, 1)$ where $\mu \in \mathbb{R}$. Show that the region $C(X) = (X + z_{0.975}, X + z_{0.025})$ satisfies $\mathbb{P}_\mu(\mu \in C(X)) = 0.95$ $= 1-\alpha$ ($\alpha = 0.05$)

$$\mathbb{P}_\mu(\mu \in C(\vec{X}))$$
$$= \mathbb{P}_\mu(X + z_{1-\alpha/2} < \mu < X + z_{\alpha/2})$$
$$= \mathbb{P}_\mu(z_{1-\alpha/2} < \mu - X < z_{\alpha/2})$$
$$= \mathbb{P}(z_{1-\alpha/2} < Z < z_{\alpha/2}) \text{ where } Z \sim N(0,1)$$

$$= \Phi(z_{\alpha/2}) - \Phi(z_{1-\alpha/2})$$
$$= 1 - \alpha/2 - \alpha/2$$
$$= 1 - \alpha$$

# Interval Estimators and Confidence Intervals

- Definition 4.1: An **interval estimate** of a parameter $\theta \in \Theta \subseteq \mathbb{R}$ is any pair of statistics $L, U : \mathcal{X}^n \to \mathbb{R}$ such that $L(\mathbf{x}) \leq U(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^n$. The random interval $\underbrace{(L(\mathbf{X}), U(\mathbf{X}))}$ is called an **interval estimator**.

  $\to$ an interval with random endpoints!

- Example 4.2: $N(\mu,1)$: $(X_{(1)}, X_{(n)}+5)$ Bernoulli$(p)$: $(-\bar{X}_n - 4, \bar{X}_n + 5)$

  Are these good? Bad? Depends on your tolerance!

- Definition 4.2: Suppose $\alpha \in [0,1]$. An interval estimator $(L(\mathbf{X}), U(\mathbf{X}))$ is a $\mathbf{(1-\alpha)}$**-confidence interval** for $\theta$ if $\mathbb{P}_\theta\left(L(\mathbf{X}) < \theta < U(\mathbf{X})\right) \geq 1-\alpha$ for all $\theta \in \Theta$. We refer to $1-\alpha$ as the **confidence level** of the interval.

  More generally, we can have a <u>$(1-\alpha)$-confidence region</u> $C(\vec{x}) \subseteq \Theta$, which satisfies $\mathbb{P}_\theta(\theta \in C(\vec{X})) \geq 1-\alpha \ \forall \theta \in \Theta$.

- Example 4.3:

  $X \sim N(\mu, 1) \Rightarrow$ We just showed in Ex. 4.1 that

  $(X + z_{1-\alpha/2}, X + z_{\alpha/2})$ is a $(1-\alpha)$-confidence interval for $\mu$

# One-Sided Intervals

- Definition 4.3: A **lower one-sided** confidence interval is a confidence interval of the form $(L(\mathbf{X}), \infty)$. An **upper one-sided** confidence interval is a confidence interval of the form $(-\infty, U(\mathbf{X}))$.

- Example 4.4: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$. Find a lower one-sided 0.5-confidence interval for $\mu$.

$0.5 = \mathbb{P}(Z < 0)$ where $Z \sim N(0,1)$

$= \mathbb{P}_\mu\left( \frac{\bar{X}_n - \mu}{\sqrt{1/n}} < 0 \right)$

$= \mathbb{P}_\mu(\bar{X}_n < \mu)$

$= \mathbb{P}_\mu(\mu \in (\bar{X}_n, \infty))$

So $(\bar{X}_n, \infty)$ is a lower one-sided 0.5-CI for $\mu$.

"confidence interval"

But $(X_i, \infty)$ is another one!

So $(1-\alpha)$-CIs are not unique!

# Confidence Intervals: Warmups

- The reason for the "$\geq 1-\alpha$" in the definition is that $\mathbb{P}_\theta\left(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})\right)$ may not be free of $\theta$, depending on the choices of $L(\mathbf{X})$ and $U(\mathbf{X})$

- The lower bound means we want $1-\alpha$ confidence even in the "worst case"; equivalently,

$$\inf_{\theta \in \Theta} \underbrace{\mathbb{P}_\theta\left(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})\right)}_{\text{"Coverage probability"}} \geq 1-\alpha$$

- Example 4.5: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$, where $\theta > 0$. Find $a \in \mathbb{R}$ such that $(aX_{(n)}, 2aX_{(n)})$ is a $\underbrace{95\%}_{1-\alpha}$ confidence interval for $\theta$.

$$\begin{aligned}
1-\alpha &= \mathbb{P}_\theta\left(\theta \in (aX_{(n)}, 2aX_{(n)})\right) \\
&= \mathbb{P}_\theta\left(aX_{(n)} < \theta < 2aX_{(n)}\right) \\
&= \mathbb{P}_\theta\left(\frac{\theta}{2a} < X_{(n)} < \frac{\theta}{a}\right) \\
&= F_{X_{(n)}}\left(\frac{\theta}{a}\right) - F_{X_{(n)}}\left(\frac{\theta}{2a}\right) \\
&= \left(\frac{\theta/a}{\theta}\right)^n - \left(\frac{\theta/2a}{\theta}\right)^n \rightarrow \frac{1}{a^n} - \frac{1}{(2a)^n}
\end{aligned}$$

$$\Rightarrow \text{Choose } a = \left(\frac{1-2^{-n}}{1-\alpha}\right)^{\frac{1}{n}}$$

check!

# Poll Time!

On Quercus: Module 4 - Poll 1

# Some Confidence Intervals Are Better Than Others

- A confidence interval is only useful when it tells us something we didn't know before collecting the data

- Example 4.6: Suppose $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim}$ Bernoulli $(\theta)$, where $\theta \in (0, 1)$. Find a 100% confidence interval for $\theta$.

$(0,1)$ ... not helpful at all!

$(X_1 - 1, X_2 + 1)$ .. also not helpful (because $(X_1 - 1, X_2 + 1) \subseteq (0,1)$)

$(X_1 - 200, \infty)$ ... extremely not helpful!

A 100%-CI contains $\Theta$, and therefore tells us nothing! We already know that $\theta \in \Theta$!

- A good confidence interval shouldn't be any longer than necessary

- We interpret the length of the interval as a measure of how accurately the data allow us to know the true value of $\theta$

# Bringing Back Hypothesis Tests

- In Module 3, we learned about test statistics and rejection regions for hypothesis tests

- Pick some arbitrary $\theta_0 \in \Theta$, and suppose we want a level-$\alpha$ test of $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$ using a test statistic $T(\mathbf{X})$

- This means finding a rejection region $R_{\theta_0}$ such that

$$\mathbb{P}_{\theta_0}(T(\mathbf{X}) \in R_{\theta_0}) \leq \alpha$$

- This is equivalent to finding an *acceptance region* $A_{\theta_0} = R^c_{\theta_0}$ such that

$$\mathbb{P}_{\theta_0}(T(\mathbf{X}) \in A_{\theta_0}) \geq 1 - \alpha$$

If this holds $\forall \theta_0 \in \Theta$, then $\{\vec{X} : T(\vec{X}) \in A_{\theta_0}\}$ is a $(1-\alpha)$-Confidence region!

# Confidence Intervals Via Test Statistics

- If the statement $T(\mathbf{X}) \in A_{\theta_0}$ can be manipulated into an equivalent statement of the form $L(\mathbf{X}) < \theta_0 < U(\mathbf{X})$, then

$$\mathbb{P}_{\theta_0}(L(\mathbf{X}) < \theta_0 < U(\mathbf{X})) \geq 1 - \alpha$$

- But $\theta_0 \in \Theta$ was arbitrary!

- So if we did this right, we must have

$$\mathbb{P}_{\theta}\left(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})\right) \geq 1 - \alpha \quad \text{for all } \theta \in \Theta$$

- This method of finding confidence intervals is called *inverting a hypothesis test*

- We can also go the other way! i.e., start with a $(1-\alpha)\cdot$CI $(L(\vec{X}), U(\vec{X}))$ and "invert" it to form a level-$\alpha$ test of $H_0: \theta = \theta_0$ vs $H_A: \theta \neq \theta_0$. (Assignment 4).

# Famous Examples: $Z$-Intervals

- Example 4.7: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\sigma^2$ is known. Find a $(1-\alpha)$-confidence interval for $\mu$ by inverting the two-sided $Z$-test.

Let $\mu_0 \in \mathbb{R}$. We need a level-$\alpha$ test of $H_0: \mu = \mu_0$ vs $H_A: \mu \neq \mu_0$.

From Example 3.15, $R_{\mu_0} = \left\{ \vec{x} \in \mathcal{X}^n : \left| \frac{\bar{x}_n - \mu_0}{\sqrt{\sigma^2/n}} \right| > z_{\alpha/2} \right\}$

$\Rightarrow A_{\mu_0} = \left\{ \vec{x} \in \mathcal{X}^n : \left| \frac{\bar{x}_n - \mu_0}{\sqrt{\sigma^2/n}} \right| < z_{\alpha/2} \right\}$

Therefore, $1 - \alpha = \mathbb{P}_\mu(\vec{X} \in A_\mu)$

$= \mathbb{P}_\mu\left( -z_{\alpha/2} < \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} < z_{\alpha/2} \right)$

$=$

$\vdots$

$= \mathbb{P}_\mu\left( \bar{X}_n - z_{\alpha/2}\sqrt{\sigma^2/n} < \mu < \bar{X}_n + z_{\alpha/2} \cdot \sqrt{\sigma^2/n} \right)$

So a $(1-\alpha)$-CI for $\mu$ is $\left( \bar{X}_n - z_{\alpha/2}\sqrt{\sigma^2/n},\ \bar{X}_n + z_{\alpha/2}\sqrt{\sigma^2/n} \right)$ "z-interval"

# Famous Examples: One-Sided $Z$-Intervals

- Example 4.8: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ where $\mu \in \mathbb{R}$ and $\sigma^2$ is known. Find a lower one-sided $(1-\alpha)$-confidence interval for $\mu$ by inverting an appropriate one-sided $Z$-test.

Ex 3.16: $R_{\mu_0} = \left\{ \vec{x} \in \mathcal{X}^n : \frac{\bar{x}_n - \mu_0}{\sqrt{\sigma^2/n}} > z_\alpha \right\} \Rightarrow A_{\mu_0} = \left\{ \vec{x} \in \mathcal{X}^n : \frac{\bar{x}_n - \mu_0}{\sqrt{\sigma^2/n}} < z_\alpha \right\}.$

$$\text{So } 1-\alpha = \mathbb{P}_\mu\left( \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} < z_\alpha \right)$$

$$= \mathbb{P}_\mu\left( -\mu < z_\alpha \cdot \sqrt{\frac{\sigma^2}{n}} - \bar{X}_n \right)$$

$$= \mathbb{P}_\mu\left( \mu > \bar{X}_n - z_\alpha \cdot \sqrt{\frac{\sigma^2}{n}} \right)$$

$$\Rightarrow \text{Choose } \left( \bar{X}_n - z_\alpha \cdot \sqrt{\frac{\sigma^2}{n}}, \infty \right).$$

EXERCISE: find an upper $(1-\alpha)$-CI for $\mu$ using this technique!

# Famous Examples: $t$-Intervals

- Example 4.9: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Find a $(1-\alpha)$-confidence interval for $\mu$ by inverting the two-sided $t$-test.

Ex 3.17: $R_{\mu_0} = \left\{ \vec{x} \in \mathcal{X}^n : \left| \frac{\bar{x}_n - \mu_0}{\sqrt{S_n^2/n}} \right| > t_{n-1,\alpha/2} \right\}$

$\Rightarrow A_{\mu_0} = R_{\mu_0}^c = \left\{ \vec{x} \in \mathcal{X}^n : \left| \frac{\bar{x}_n - \mu_0}{\sqrt{S_n^2/n}} \right| < t_{n-1,\alpha/2} \right\}$

So $1-\alpha = \mathbb{P}_\mu\left( -t_{n-1,\alpha} < \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} < t_{n-1,\alpha/2} \right)$

$= \mathbb{P}_\mu\left( \bar{X}_n - t_{n-1,\alpha/2}\sqrt{\frac{S_n^2}{n}} < \mu < \bar{X}_n + t_{n-1,\alpha/2}\sqrt{\frac{S_n^2}{n}} \right)$

$\Rightarrow$ Choose $\left( \bar{X}_n - t_{n-1,\alpha/2}\sqrt{\frac{S_n^2}{n}}, \bar{X}_n + t_{n-1,\alpha/2}\sqrt{\frac{S_n^2}{n}} \right)$ "t-interval"

# Famous Examples: One-Sided $t$-Intervals

- Example 4.10: Let $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Find an upper one-sided $(1-\alpha)$-confidence interval for $\mu$ by inverting an appropriate one-sided $t$-test.

EXERCISE!

Should the corresponding $H_0$ be

$H_0: \mu \leq \mu_0$ or $H_0: \mu \geq \mu_0$?

Figure it out!

# An LRT-Based Interval

$F_\theta(x) = (1 - e^{-(x-\theta)}) \cdot \mathbb{1}_{x \geq \theta}$

- Example 4.11: Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f_\theta(x) = e^{-(x-\theta)} \cdot \mathbb{1}_{x \geq \theta}$, where $\theta \in \mathbb{R}$. Find a $(1-\alpha)$-confidence interval for $\theta$ by inverting an LRT.

From Ex. 3.21, the LRT of $H_0: \theta \leq \theta_0$ vs $H_A: \theta > \theta_0$ had a rejection region of the form

$$R_{\theta_0} = \{\vec{x} \in \mathcal{X}^n : x_{(1)} > \theta_0 - \frac{\log(c)}{n} \text{ OR } x_{(1)} < \theta_0\}$$

$$\rightarrow A_{\theta_0} = \{\vec{x} \in \mathcal{X}^n : x_{(1)} < \theta_0 - \frac{\log(c)}{n} \text{ AND } x_{(1)} > \theta_0\} = \{\vec{x} \in \mathcal{X}^n : x_{(1)} + \frac{\log(c)}{n} < \theta_0 < x_{(1)}\}$$

So if we choose $c$ to make that $R_{\theta_0}$ a size-$\alpha$ test, then $\left(X_{(1)} + \frac{\log(c)}{n}, X_{(1)}\right)$ will be a $(1-\alpha)$-CI for $\theta$. How?

$$\begin{aligned}
1-\alpha &= \mathbb{P}_\theta\Big(X_{(1)} \leq \theta - \frac{\log(c)}{n} \wedge \underbrace{X_{(1)} \geq \theta}_{\text{always true!}}\Big) \\
&= \mathbb{P}_\theta\Big(X_{(1)} \leq \theta - \frac{\log(c)}{n}\Big) \\
&= 1 - \Big(1 - F_\theta\big(\theta - \frac{\log(c)}{n}\big)\Big)^n \\
&= 1 - \Big(1 - 1 + \exp\big(-(\theta - \frac{\log(c)}{n} - \theta)\big)\Big)^n \\
&= 1 - c
\end{aligned}$$

$\Rightarrow$ Choose $c = \alpha$ $\Rightarrow \left(X_{(1)} + \frac{\log(\alpha)}{n}, X_{(1)}\right)$ is a $(1-\alpha)$-CI for $\theta$.

# Functions of the Data *and* the Parameter

- In constructing our confidence intervals, we've often encountered statements that look like
$$\mathbb{P}_\theta\left(a < Q(\mathbf{X}, \theta) < b\right) \geq 1 - \alpha,$$
where $Q : \mathcal{X}^n \times \Theta \to \mathbb{R}$ is a function of the data $\mathbf{X}$ *and* the parameter $\theta$, and $a, b$ are constants

- We were able to choose those constants $a$ and $b$ because we knew exactly what the distribution of $Q(\mathbf{X}, \theta)$ was

- We could then "invert" the statement $a < Q(\mathbf{X}, \theta) < b$ to produce a confidence interval for $\theta$

$Q(\vec{X}, \mu) \sim N(0,1)$

- Example 4.12: $N(\mu, \sigma^2)$, $\sigma^2$ known: $\mathbb{P}_\mu\left(-z_\alpha < \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} < z_\alpha\right) = 1-\alpha$

- Example 4.13: $\text{Unif}(0,\theta)$: $\mathbb{P}_\theta\left(\frac{1}{2a} \leq \frac{X_{(n)}}{\theta} \leq \frac{1}{a}\right) = 1-\alpha$, where $a$ was chosen as before

$Q(\vec{X}, \theta)$, distribution was free of $\theta$

# Pivotal Quantities

- The key in these examples was that the *distribution* of $Q(\mathbf{X}, \theta)$ is free of $\theta$
- Definition 4.4: A random variable $Q(\mathbf{X}, \theta)$ is a **pivotal quantity** (or **pivot**) for $\theta$ if its distribution is free of $\theta$.
- So if $\mathbf{X} \sim f_{\theta_1}$ and $\mathbf{Y} \sim f_{\theta_2}$, then $Q(\mathbf{X}, \theta_1) \overset{d}{=} Q(\mathbf{Y}, \theta_2)$
- Every ancillary statistic is a pivotal quantity
- Example 4.14: $N(\mu, \sigma^2)$, $\sigma^2$ known: $\mathbb{P}_\mu\left(-z_\alpha < \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} < z_\alpha\right) = 1-\alpha$

  $\sim N(0,1)$ which is free of $\mu$ $\Rightarrow$ pivotal!
- Example 4.15: $\text{Exp}(\lambda)$: $Q(\vec{X}, \lambda) = \frac{X_1}{\lambda} \sim \text{Exp}(1) \leftarrow$ free of $\lambda \Rightarrow \frac{X_1}{\lambda}$ is pivotal for $\lambda$

# Poll Time!

We can calculate $Q(x, \theta')$ for any $x \in \mathcal{X}$ and $\theta' \in \Theta$.

But if $\vec{X} \sim f_\theta$, we may not know the distribution of $Q(X, \theta')$ if $\theta' \neq \theta$...

On Quercus: Module 4 - Poll 2

# Confidence Intervals from Pivotal Quantities

- Example 4.16: Let $X_1, X_2, \dots, X_n \overset{iid}{\sim}$ Exp $(\lambda)$, $\lambda > 0$. Show that $Q(\mathbf{X}, \lambda) = 2\lambda \sum_{i=1}^{n} X_i$ is a pivotal quantity, and use it to find a $1 - \alpha$ confidence interval for $\lambda$.

Use mgfs! $M_{\Sigma X_i}(t) = \left(\frac{\lambda}{\lambda - t}\right)^n$, $t < \lambda \Rightarrow M_{2\lambda\Sigma X_i}(t) = \left(\frac{\lambda}{\lambda - 2\lambda t}\right)^n = \left(\frac{1}{1-2t}\right)^n$

The mgf is free of $\lambda$, so the distribution of $2\lambda\Sigma X_i$ is too $\Rightarrow 2\lambda\Sigma X_i$ is pivotal!

In fact, the mgf tells us that $2\lambda\Sigma X_i \sim \chi^2_{(2n)}$. (FYI)

So set $1-\alpha = \mathbb{P}_\lambda(a < 2\lambda\Sigma X_i < b)$ for some $a, b \in \mathbb{R}$ with $a < b$.

They must satisfy $1-\alpha = F_{\chi^2_{(2n)}}(b) - F_{\chi^2_{(2n)}}(a)$. Many choices!

For example: if we choose $a = 0$, then $1-\alpha = F_{\chi^2_{(2n)}}(b) \Rightarrow b = F^{-1}_{\chi^2_{(2n)}}(1-\alpha) =: \chi^2_{(2n),\alpha}$

So $1-\alpha = \mathbb{P}_\lambda(0 < 2\lambda\Sigma X_i < \chi^2_{(2n),\alpha})$

$\Rightarrow$ Choose $\left(0, \frac{\chi^2_{(2n),\alpha}}{2\Sigma X_i}\right)$

# Finding Pivotal Quantities

- There's no all-purpose strategy to finding pivotal quantities, but there's a neat trick that sometimes lets us pull one out of the pdf of a statistic $T(\mathbf{X})$
- Theorem 4.1: Suppose that $T(\mathbf{X}) \sim f_\theta$ is univariate and continuous, such that the pdf can be expressed as

$$f_\theta(t) = g(Q(t,\theta)) \cdot \left| \frac{\partial}{\partial t} Q(t,\theta) \right|$$

for some function $g(\cdot)$ which is free of $\theta$ and some function $Q(t,\theta)$ which is continuously differentiable and one-to-one as a function of $t$ (i.e., with $\theta$ fixed). Then $Q(T(\mathbf{X}),\theta)$ is a pivot.

*Proof.*

Keep scrolling...

Fix $\theta \in \Theta$ and let $h_\theta(q)$ be the pdf of $Q(T(\vec{X}), \theta) =: Q_\theta(T(\vec{X}))$.

Let $Q_\theta^{-1}(q)$ be the functional inverse of $Q_\theta(t)$.

Then...

$$h_\theta(q) = f_\theta\left(Q_\theta^{-1}(q)\right) \cdot \left| \frac{d}{dq} Q_\theta^{-1}(q) \right| \quad \text{by the usual transformation of variables formula}$$

$$= f_\theta\left(Q_\theta^{-1}(q)\right) \cdot \left| \frac{d}{dt} Q_\theta(t) \Big|_{t = Q_\theta^{-1}(q)} \right|^{-1}$$

$$= \underbrace{g\left(Q_\theta\left(Q_\theta^{-1}(q)\right)\right) \cdot \left| \frac{d}{dt} Q_\theta(t) \Big|_{t = Q_\theta^{-1}(q)} \right|}_{\text{by assumption}} \cdot \left| \frac{d}{dt} Q_\theta(t) \Big|_{t = Q_\theta^{-1}(q)} \right|^{-1}$$

$$= g(q), \text{ which is free of } \theta.$$

So the distribution of $Q(T(\vec{X}), \theta)$ is free of $\theta$. $\square$

# Finding Pivotal Quantities: Examples

$$f_\theta(t) = g(Q(t,\theta)) \cdot \left|\frac{\partial}{\partial t} Q(t,\theta)\right|$$

- Example 4.17: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0,\theta)$ where $\theta > 0$. Find a pivotal quantity based on $T(\mathbf{X}) = X_{(n)}$, and use it to construct a $1-\alpha$ confidence interval for $\theta$.

The pdf of $T(\vec{X})$ is $\underbrace{n \cdot f_\theta(t) \cdot F_\theta(t)^{n-1}}_{\text{Assignment 0}} = n \cdot \frac{1}{\theta} \cdot \left(\frac{t}{\theta}\right)^{n-1} = \frac{nt^{n-1}}{\theta^n} = 1 \cdot \left|\frac{\partial}{\partial t} \frac{t^n}{\theta^n}\right|$

$g(\cdot)$ $\quad Q(t,\theta)$

By Theorem 4.1, $Q(X_{(n)},\theta) = \frac{X_{(n)}^n}{\theta^n}$ is a pivotal quantity.

What's its distribution? For $x \in (0,1)$,

$$\mathbb{P}_\theta\left(\frac{X_{(n)}^n}{\theta^n} \le x\right)$$
$$= \mathbb{P}_\theta\left(X_{(n)} \le \theta x^{\frac{1}{n}}\right)$$
$$= F_\theta\left(\theta x^{\frac{1}{n}}\right)^n$$
$$= \left(\frac{\theta x^{\frac{1}{n}}}{\theta}\right)^n$$
$$= x \implies Q(X_{(n)},\theta) \sim \text{Unif}(0,1).$$

→ Choose $a<b$ s.t.

$$1-\alpha = \mathbb{P}_\theta\left(a < \frac{X_{(n)}^n}{\theta^n} < b\right) = \mathbb{P}(a < U < b),$$

$U \sim \text{Unif}(0,1)$.

For example, take $a = \alpha/2$ and $b = 1-\alpha/2$.

$$\implies 1-\alpha = \mathbb{P}_\theta\left(\alpha/2 < \frac{X_{(n)}^n}{\theta^n} < 1-\alpha/2\right)$$
$$= \mathbb{P}_\theta\left(\frac{X_{(n)}^n}{1-\alpha/2} < \theta^n < \frac{X_{(n)}^n}{\alpha/2}\right)$$

$\implies$ Choose $\left(\frac{X_{(n)}}{(1-\alpha/2)^{1/n}}, \frac{X_{(n)}}{(\alpha/2)^{1/n}}\right)$

# Finding Pivotal Quantities: Examples

- Example 4.18: Let $X \sim f_\theta(x) = \frac{2(\theta - x)}{\theta^2} \cdot \mathbb{1}_{0 \le x \le \theta}$, where $\theta > 0$. Find a pivotal quantity based on $X$, and use it to construct a $1-\alpha$ confidence interval for $\theta$.

Observe that if $Q(x,\theta) = \frac{\theta - x}{\theta}$, then $f_\theta(x) = \underbrace{2 \cdot Q(x,\theta)}_{= g(Q(x,\theta)),\ \text{where } g(x) = 2x} \cdot \left| \frac{\partial}{\partial x} Q(x,\theta) \right|$.

By Theorem 1.4, $Q(X,\theta) = \frac{\theta - X}{\theta}$ is a pivotal quantity. What's its distribution?

For $x \in (0,1)$,

$$\mathbb{P}_\theta\left(\frac{\theta - X}{\theta} \le x\right)$$
$$= \mathbb{P}_\theta(X \ge (1-x)\cdot\theta)$$
$$= \int_{(1-x)\cdot\theta}^{\theta} \frac{2(\theta - t)}{\theta^2}\, dt$$
$$= x^2$$

Plenty of choices to make $1-\alpha = \mathbb{P}_\theta\left(a < \frac{\theta - X}{\theta} < b\right)$
$= b^2 - a^2$.

For example, if $a = 0$, then $b = \sqrt{1-\alpha}$. Then

$$1-\alpha = \mathbb{P}_\theta\left(0 < \frac{\theta - X}{\theta} < \sqrt{1-\alpha}\right)$$
$$= \mathbb{P}_\theta\left(X < \theta < \frac{X}{1-\sqrt{1-\alpha}}\right)$$

$\Longrightarrow$ Choose $\left(X, \frac{X}{1-\sqrt{1-\alpha}}\right)$.

# Confidence Intervals: Interpretations

- Confidence intervals are almost as widely misinterpreted as $p$-values

- Suppose that in a published scientific study, you see a stated 95% confidence interval such as $(0.932, 1.452)$

  for $\theta$

- How do you interpret this correctly?

  $(0.932, 1.452)$ is $\underline{\text{an}}$ "observed" value of the 95%-CI $(L(\vec{X}), U(\vec{X}))$.

  $(L(\vec{X}), U(\vec{X}))$ is random! $(L(\vec{x}), U(\vec{x}))$ is observed!

  random variables constants

- Should we be surprised if we try and reproduce the study and calculate a 95% confidence interval of $(0.824, 1.734)$?

- What about $(-0.232, 1.440)$?

Poll Time! If $X_1, \dots, X_{100} \overset{iid}{\sim} f_\theta$, By definition, $0.95 \leq \mathbb{P}(L(X_i) < \theta < U(X_i))$

$$\mathbb{E}[\#\text{ of } \theta \text{ coverages}]$$

$$= \mathbb{E}\left[\sum_{i=1}^{100} \mathbb{1}_{L(X_i) < \theta < U(X_i)}\right]$$

$$= \sum_{i=1}^{100} \mathbb{P}(L(X_i) < \theta < U(X_i))$$

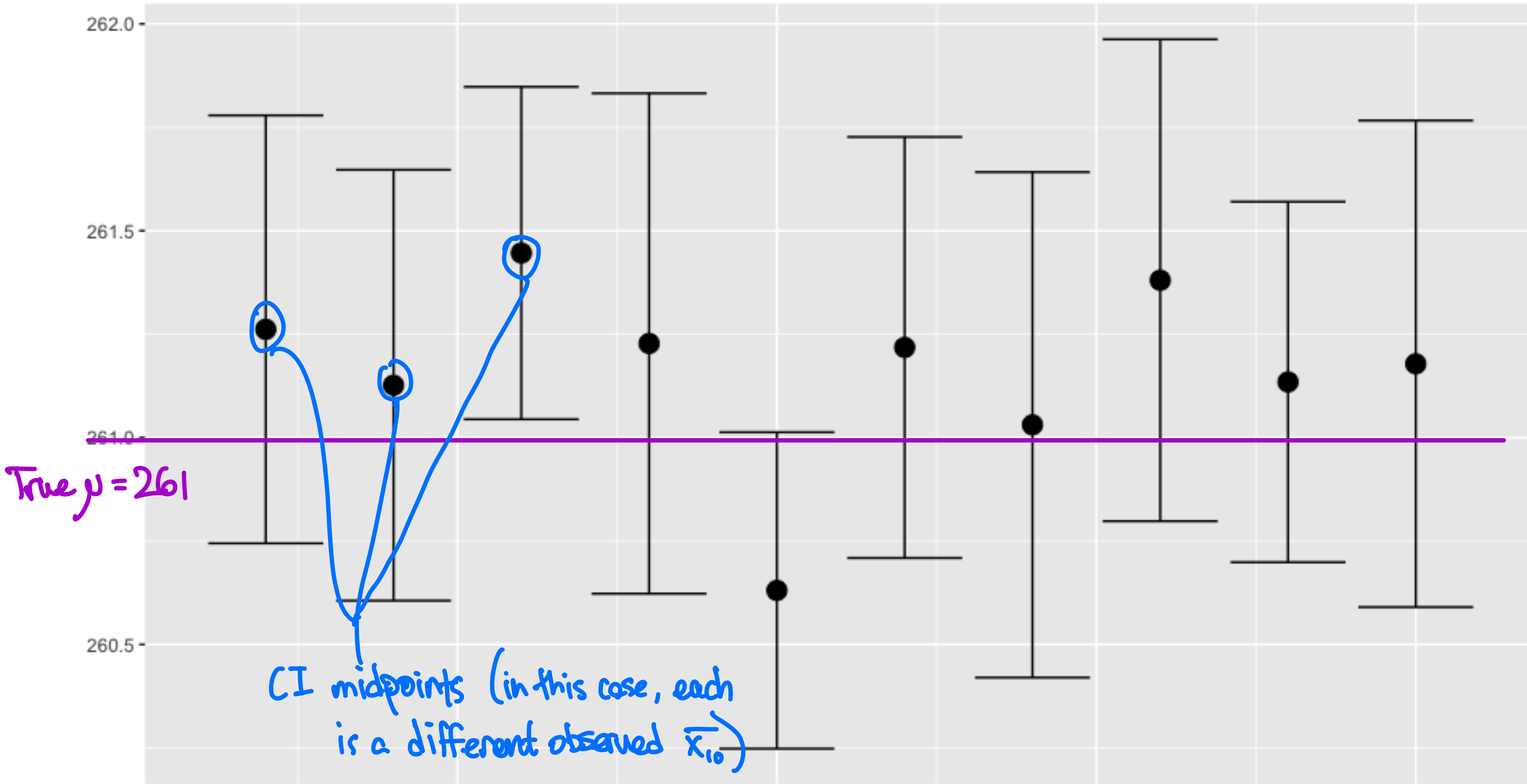
$$\overset{?}{=} \sum_{i=1}^{100} 0.95$$

On Quercus: Module 4 - Poll 3

$$= 95$$

# Confidence Intervals: Interpretations

- Here are ten observed 95% $Z$-intervals for $\mu$ calculated from ten random samples of size $n = 15$ from a $\mathcal{N}(\mu, 1)$ distribution:



# Questioning Our Assumptions...

- All of the theory we've done up to this point has depended on the assumption of an underlying statistical model

- When we say "Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$...", we're assuming the data follows one of the distributions in the parametric family $\{f_\theta : \theta \in \Theta\}$ and only the parameter $\theta$ is unknown

- If we get the statistical model wrong, then any inferences we make about $\theta$ are likely to be completely invalid

- So it's extremely important to be able to check that statistical model assumption

# Nothing Is Certain

- Of course, we can't *know* for sure that a model is correct
- Unless we generate the data ourselves... but then there would be no point in doing inference!
- But we can perform checks that give us confidence in our assumptions
- This is called *model checking*
- We will study two kinds of model checks: visual diagnostics and goodness-of-fit tests

# Histograms: Preliminaries

- Suppose we have iid data $X_1, X_2, \ldots, X_n$, which we hypothesize are distributed according to a cdf $F_\theta$

- Let's group the range of the data into bins $[h_1, h_2], (h_2, h_3], \ldots, (h_{m-1}, h_m]$

- By the (weak) law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \in (h_j, h_{j+1}]} \xrightarrow{p} \overbrace{\mathbb{P}_\theta(X_1 \in (h_j, h_{j+1}])}^{\mathbb{E}_\theta[\mathbb{1}_{X_1 \in (h_j, h_{j+1}]}]}$$

$$= \mathbb{P}_\theta(h_j < X_1 \le h_{j+1}) = F_\theta(h_{j+1}) - F_\theta(h_j)$$

- So if $n$ is large and we're correct about $F_\theta$, then

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \in (h_j, h_{j+1}]} \approx F_\theta(h_{j+1}) - F_\theta(h_j)$$

# The Histogram Density Function

$= F'_\theta(\cdot)$

- If, in addition, we believe the $X_i$'s are continuous with pdf $f_\theta$, then there exists $h^* \in (h_j, h_{j+1})$ such that

$$\frac{1}{n(h_{j+1} - h_j)} \sum_{i=1}^{n} \mathbb{1}_{X_i \in (h_j, h_{j+1}]} \approx \frac{F_\theta(h_{j+1}) - F_\theta(h_j)}{h_{j+1} - h_j} = f_\theta(h^*)$$

by the mean value theorem!

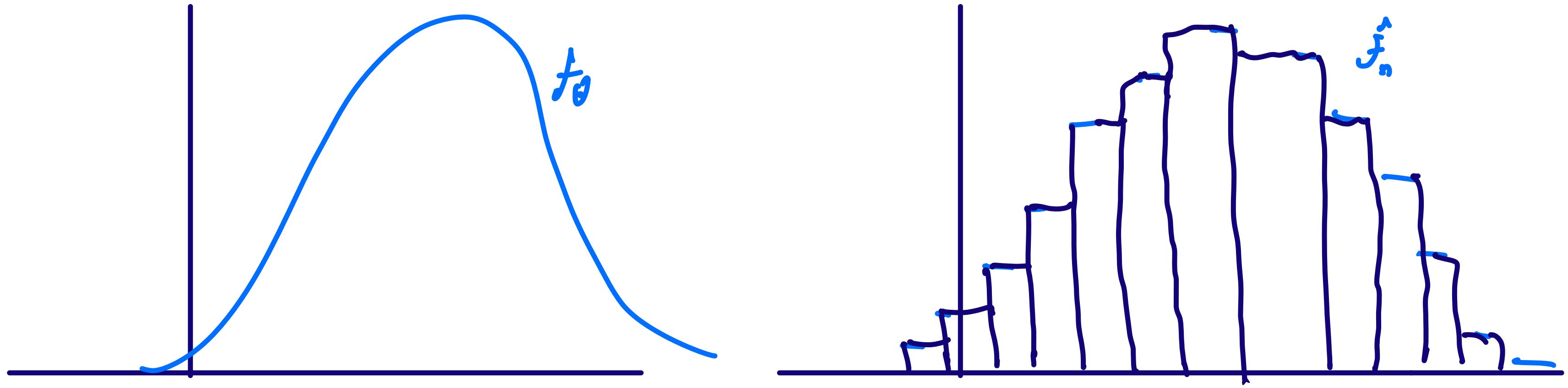
- Definition 4.5: Given data $X_1, \ldots, X_n$ and a partition $h_1 < h_2 < \cdots < h_m$, the **density histogram function** is defined as

$$\hat{f}_n(t) = \begin{cases} \frac{1}{n(h_{j+1}-h_j)} \sum_{i=1}^{n} \mathbb{1}_{X_i \in (h_j, h_{j+1}]}, & t \in (h_j, h_{j+1}] \\ 0, & \text{otherwise} \end{cases}$$

A random function (since it's implicitly a function of the r.v.'s $X_1, \ldots, X_n$)

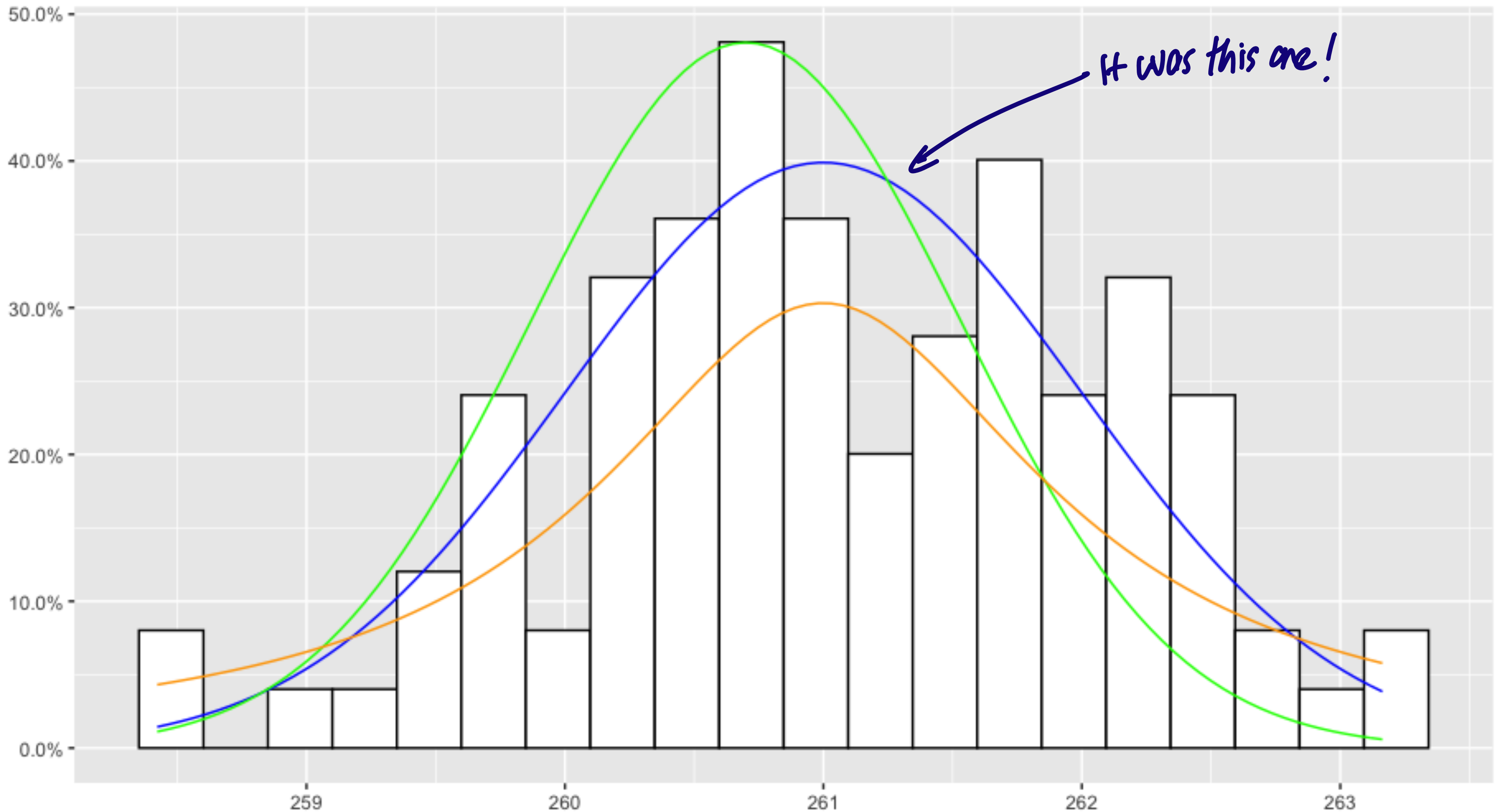
# Histograms

- If we believe that our observed data $x_1, \ldots, x_n$ are realizations of $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim} f_\theta$, then the observed $\hat{f}_n(t)$ should look like a "discretized" version of $f_\theta(t)$
- ...and the resemblance should improve as $n$ gets larger and each bin size $h_{j+1} - h_j$ gets smaller
- Definition 4.6: A plot of a density histogram function $\hat{f}_n(t)$ with vertical lines drawn at each $h_j$ is called a **histogram**. A histogram where each bin width $h_{j+1} - h_j = 1$ is called a **relative frequency plot**.



# Histograms: An Example

- Here's a histogram $(n = 100)$ overlaid with three hypothesized pdfs; which is more likely to have generated the data?



# Poll Time!

On Quercus: Module 4 - Poll 4

# Empirical CDFs

- We might prefer to deal with the cdf $F_\theta$ instead
- If we fix any $t \in \mathbb{R}$, then the law of large numbers says that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \le t} \xrightarrow{p} \mathbb{P}_\theta\left(X \le t\right)$$

$X \sim F_\theta$

- So if $n$ is large and we're correct about $F_\theta$, then

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \le t} \approx F_\theta(t)$$

This <u>will</u> approximate the true data-generating cdf (regardless of whether or not that's the hypothesised $F_\theta$)

- Definition 4.7: Given a random variables $X_1, \ldots, X_n$, the **empirical distribution function (ecdf)** is defined as

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \le t}$$

# Empirical CDFs Are Nice

- If we believe that our observed data $x_1, \ldots, x_n$ are realizations of $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$, then $\hat{F}_n(t)$ should look like $F_\theta(t)$
- In fact, a famous result called the **Glivenko-Cantelli theorem** says that if $F_\theta$ really is the true cdf, then $\hat{F}_n(t) \longrightarrow F_\theta(t)$ as $n \to \infty$ in a *much* stronger sense than convergence in probability "uniform almost sure convergence": (FYI)
  $\mathbb{P}_\theta\left(\{\omega \in \Omega : \lim_{n\to\infty} \sup_{t \in \mathbb{R}} |F_n(t)(\omega) - F_\theta(t)| > \epsilon\}\right) = 0.$
- Theorem 4.2: For any fixed $t \in \mathbb{R}$, the ecdf $\hat{F}_n(t)$ is an unbiased estimator of $F_\theta(t)$, and it has a lower variance than $\mathbb{1}_{X_i \le t}$.

*Proof.* $\mathbb{1}_{X_i \le t} \sim \text{Bernoulli}\left(\mathbb{P}_\theta(\mathbb{1}_{X_i \le t} = 1)\right)$
$= \text{Bernoulli}\left(\mathbb{P}_\theta(X_i \le t)\right)$
$= \text{Bernoulli}\left(F_\theta(t)\right)$

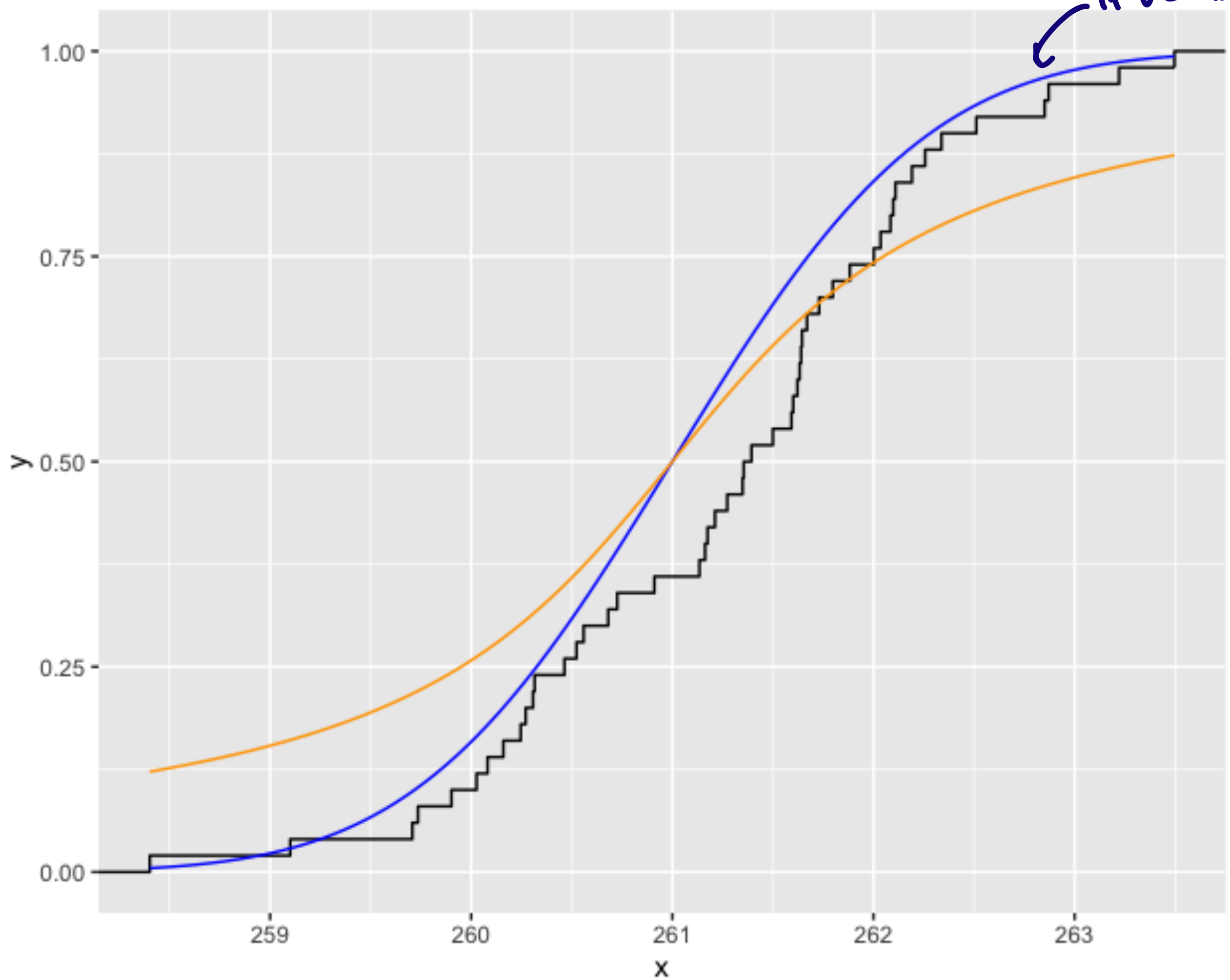
Therefore, $\mathbb{E}_\theta[\hat{F}_n(t)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{1}_{X_i \le t}] = F_\theta(t)$.

Also, $\text{Var}_\theta(\hat{F}_n(t)) = \frac{1}{n}\text{Var}_\theta(\mathbb{1}_{X_i \le t}) = \frac{1}{n} \cdot F_\theta(t) \cdot (1 - F_\theta(t)) \le F_\theta(t) \cdot (1 - F_\theta(t)) = \text{Var}_\theta(\mathbb{1}_{X_i \le t})$. $\square$

# Empirical CDFs: An Example

- Here's an ecdf ($n = 50$) overlaid with two hypothesized cdfs; which is more likely to have generated the data?

It was this one



# Poll Time!

$$X_1, \ldots, X_n \overset{iid}{\sim} N(0,1)$$

$$\mathbb{E}[\hat{F}_n(0)] = \Phi(0) = 1/2.$$

On Quercus: Module 4 - Poll 5

# Bringing Back Ancillarity and Sufficiency

- We know from Module 1 that if $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim} f_\theta$, the distribution of an ancillary statistic $S(\mathbf{X})$ is free of $\theta$

- But if we've gotten the model $\{f_\theta : \theta \in \Theta\}$ wrong, $S(\mathbf{X})$ could very well depend on $\theta$!
  (or some other unknown parameter in the "true" model)

- So some ancillary statistics provide a model check: if our realization $S(\mathbf{x})$ is "surprising", we have evidence against the model being true

- Similarly, if $T(\mathbf{X})$ is sufficient for $\theta$, then $\mathbf{X} \mid T(\mathbf{X}) = t$ shouldn't depend on $\theta$

- This leads to the idea of **residual analysis**

- Loosely speaking, residuals are based on the information in the data that is left over after we have fit the model

  (there's no formal definition of "residual")

# Residual Plots

- Example 4.19: Let $X_1, \dots, X_n$ be a random sample from a suspected $\mathcal{N}(\mu, \sigma^2)$ distribution, where $\mu \in \mathbb{R}$ and $\sigma^2$ is known. If we're correct, then $R(\mathbf{X}) = (X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)$ is ancillary for $\mu$, because

$$X_i - \bar{X}_n \sim \mathcal{N}\left(0, \frac{n-1}{n}\sigma^2\right), \quad i = 1, \dots, n$$

and therefore **standardized residuals**

$$R_i^*(\mathbf{X}) := \frac{X_i - \bar{X}_n}{\sqrt{\frac{n-1}{n}\sigma^2}} \sim \mathcal{N}(0,1).$$

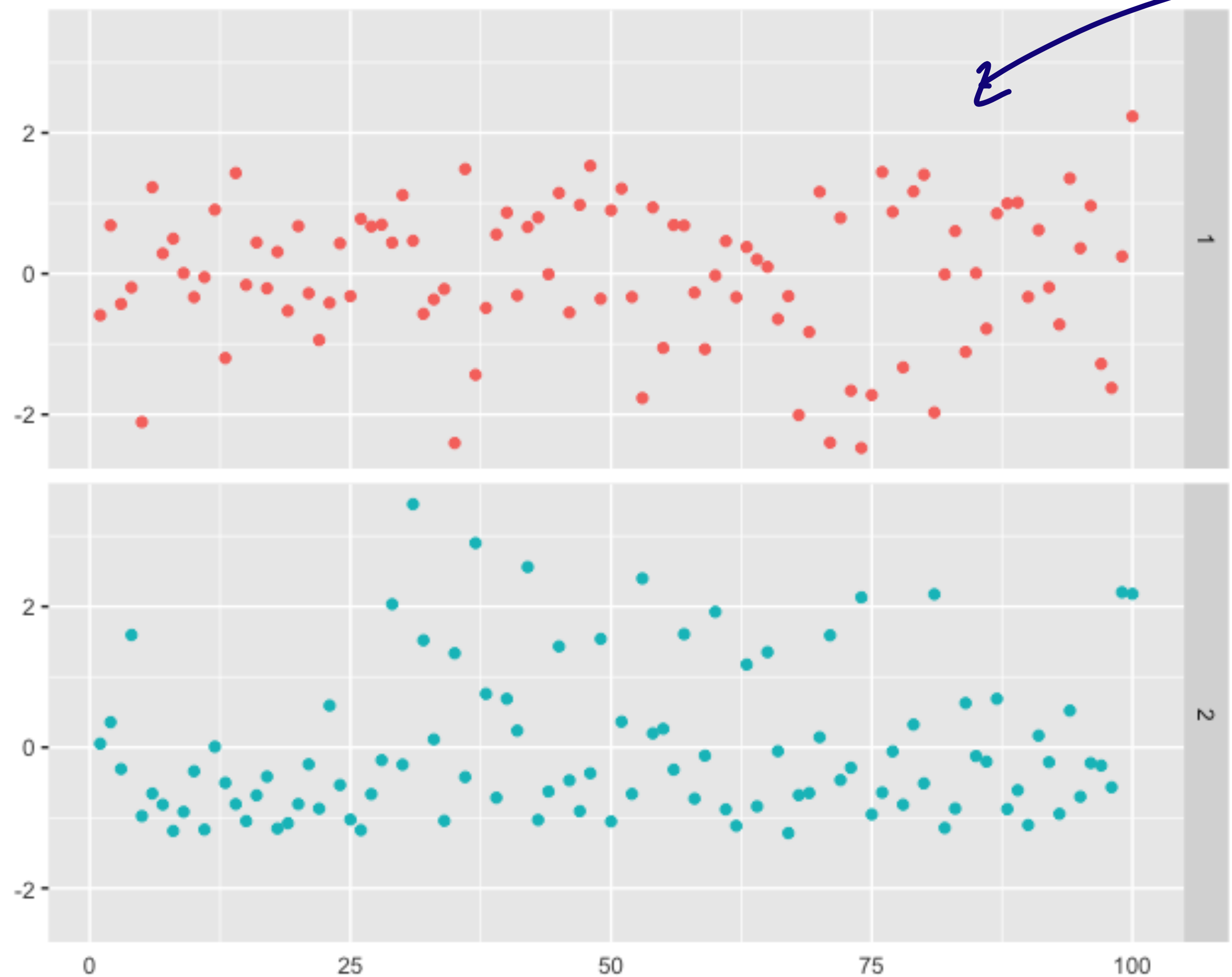
$\longleftarrow$ If $\sigma^2$ is unknown, we can just replace $\sigma^2$ by $S_n^2$, whence $R_i^* \sim t_{(n-1)}$

So if we're right about $\mathcal{N}(\mu, \sigma^2)$, then a scatterplot of the residuals shouldn't exhibit any discernable pattern, and should mostly stay within $(-3, 3)$

# Residual Plots

- Example 4.20: Here are two standardized residual plots constructed from two samples ($n = 100$) with equal variances $\sigma^2$; which looks more like it came from a $\mathcal{N}(\mu, \sigma^2)$ distribution?

It was this one!



# Probability Plots

- Probability plots extend this idea
- We need a fundamental result of probability theory first
- Theorem 4.3 (**Probability integral transform**): Let $X$ be a continuous random variable with cdf $F_\theta(x)$, and let $U = F_\theta(X)$. Then $U \sim \text{Unif}\,(0,1)$.

Proof: EXERCISE!

- The order statistics of $U_1, \ldots, U_n \overset{iid}{\sim} \text{Unif}\,(0,1)$ follow a Beta distribution: $U_{(j)} \sim \text{Beta}\,(j, n-j+1)$, and so $\mathbb{E}\left[U_{(j)}\right] = \frac{j}{n+1}$ (Assignment 0)
- This suggests a recipe: if we hypothesize $X_1, \ldots, X_n \overset{iid}{\sim} F_\theta$, then we can plot

$$\Big(\underbrace{F_\theta(x_{(j)})}, \frac{j}{n+1}\Big), \; j = 1, \ldots, n.$$

If it doesn't look like the points lie along a straight line, we should question the assumption of $F_\theta$.

i.e., $y = \frac{x}{n+1}$

$F_\theta(X_{(j)}) = [F_\theta(X)]_{(j)}$ because cdfs are increasing $\overset{d}{=} U_{(j)}$ if we're correct about $F_\theta$

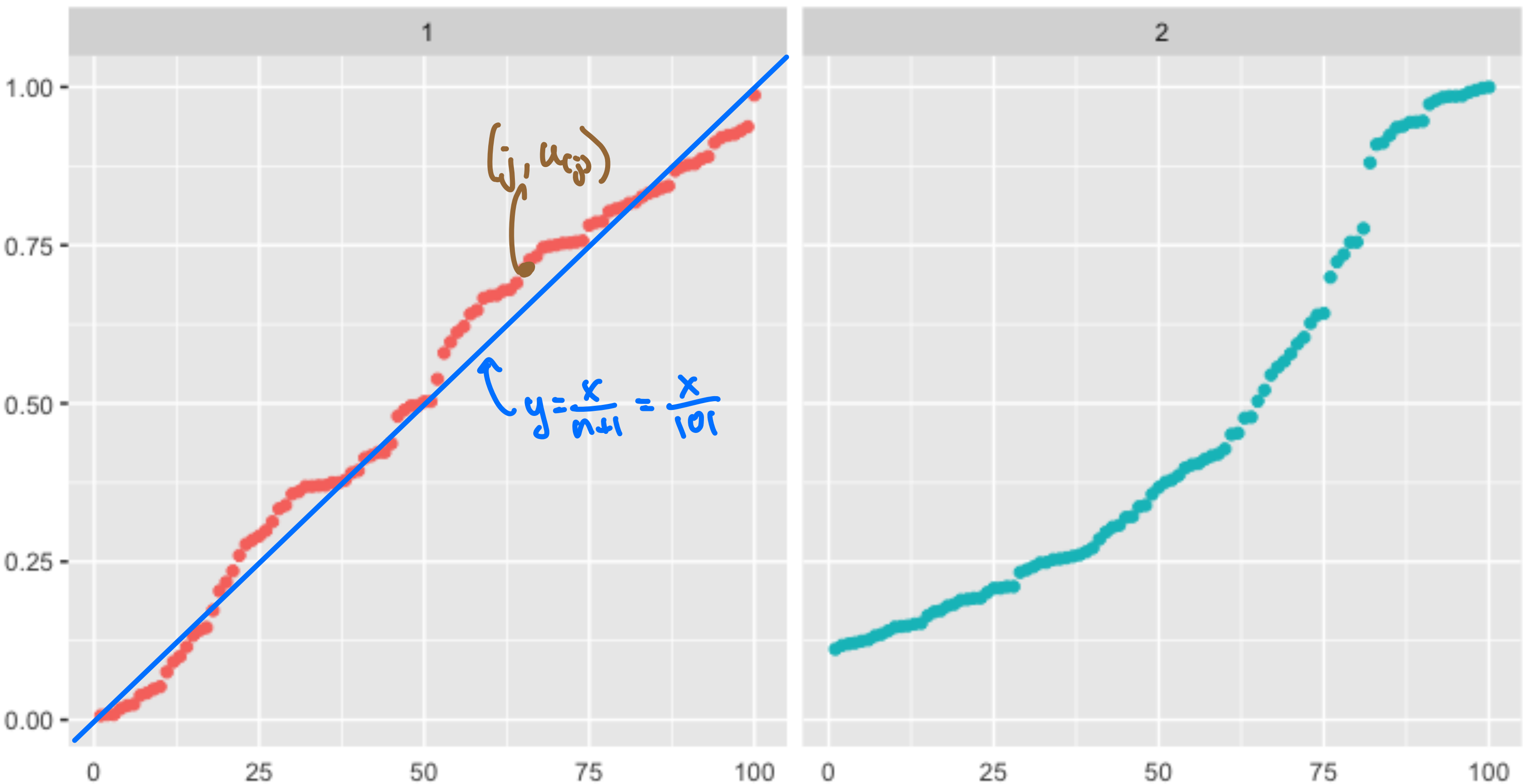
OR: $j = (n+1) \cdot \mathbb{E}_\theta(F_\theta(X_{(j)}))$ and compare to $y = x$

# Probability Plots

$n=100$

- Example 4.21: Here are two probability plots constructed from the standardized residuals as before, using $F_\theta(x) = \Phi(x)$. Which looks more like it came from a $\mathcal{N}(\mu, \sigma^2)$ distribution?

$u_{(j)}$




$j$

# Q-Q Plots

"Quantile-Quantile"

- We could also go in the other direction by looking at the quantiles
- Definition 4.8: Let $X$ be a random variable with cdf $F_\theta$. The **inverse cdf** (or the **quantile function**) is defined by $F_\theta^{-1}(t) = \inf\{x : F_\theta(x) \geq t\}$.

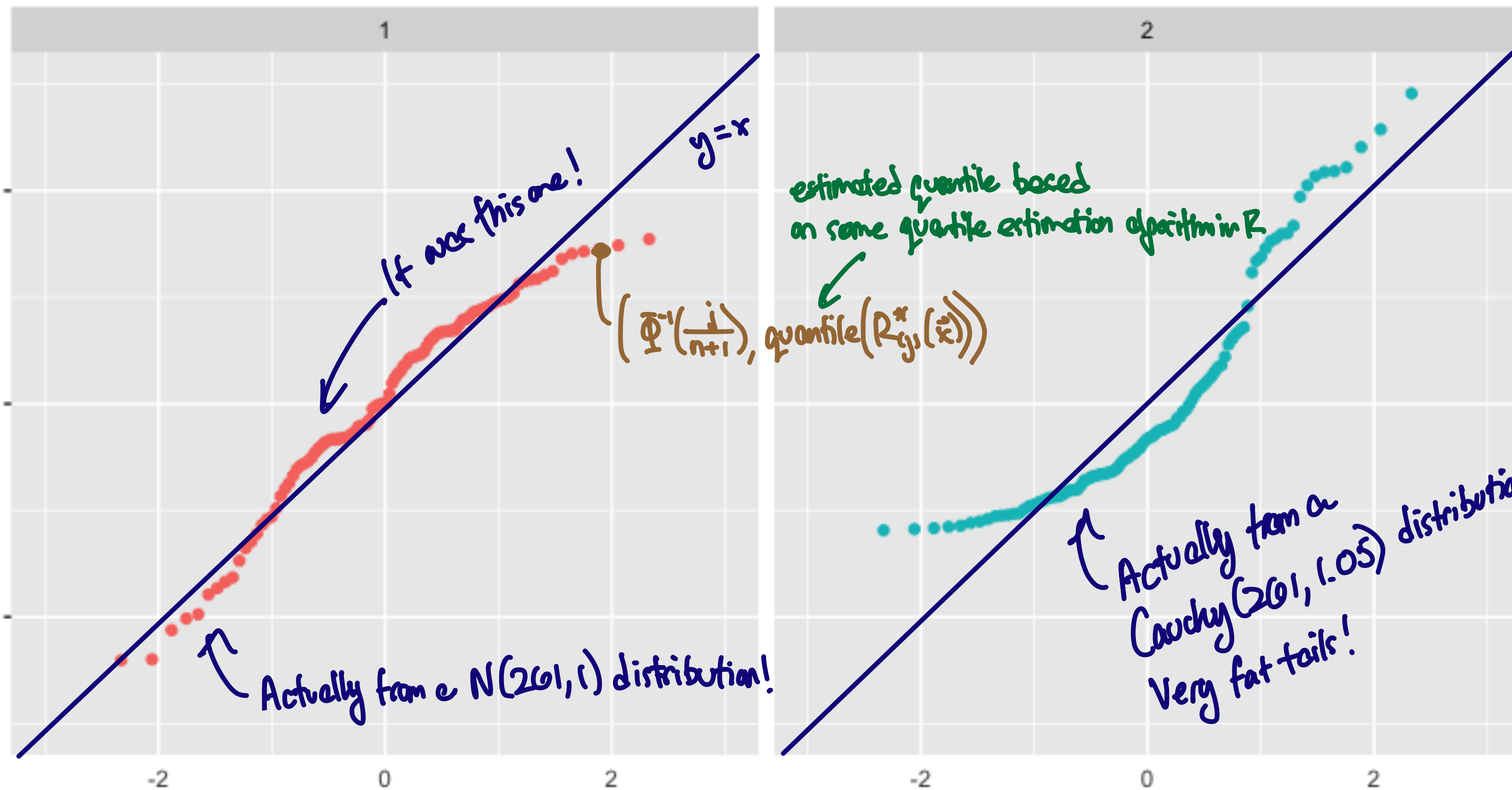
  "generalized inverse of $F_\theta$"
- When $X$ is continuous, the inverse cdf is simply the functional inverse of $F_\theta$
- There are plenty of software algorithms that can estimate the quantiles from a sample $x_1, \ldots, x_n$
- If we hypothesize $X_1, \ldots, X_n \sim F_\theta$ and we can compute $F_\theta^{-1}$, then we have another recipe for model checking:

Plot the observed quantiles versus the theoretical ones! If it doesn't look (roughly) like they lie on the line $y=x$, we should question the assumption of $F_\theta$.

# Q-Q Plots

By far, the most common use is when $F_\theta = \Phi$.
We use this when we want to check if the N(0,1) distribution does a good job of capturing the EXTREME observations (i.e., in the tails)

- Example 4.22: Here are two Q-Q plots constructed from the standardized residuals as before, using $F_\theta^{-1}(x) = \Phi^{-1}(x)$. Which looks more like it came from a $\mathcal{N}(\mu, \sigma^2)$ distribution?



# Q-Q Plots

- Q-Q plots are most frequently used as a test for normality
- But technically there's no reason why we can't use them to compare *any* two distributions, observed or hypothesized
- ...provided we can actually compute (or estimate) their quantiles, of course
- Q-Q plots are particularly useful when we want to see how the "outliers" in our data compare to the extreme values predicted by the tails of a hypothesized distribution

Check out "Chernoff faces" in the optional readings!

# Goodness of Fit Tests

- Instead of using visual diagnostics, we can use hypothesis tests as model checks

- Definition 4.9: A **goodness of fit test** for a statistical model $\{f_\theta : \theta \in \Theta\}$ is a hypothesis test that determines how well the model suits a given set of observations $x_1, \ldots, x_n$.

- This time, the null hypothesis $H_0$ is that the model $\{f_\theta : \theta \in \Theta\}$ for our data is "correct"

  $H_0$: "the data are normally distributed"

  $H_0$: "the two samples are independant"

  $H_0$: "the observations themselves are all independant"

- As usual, we have a test statistic $T(\mathbf{X})$ that follows some known distribution under $H_0$

- An observed value $T(\mathbf{x})$ which is very unlikely under $H_0$ (as quantified by a $p$-value, for example) provides evidence that the model is wrong

# Towards a Foundational Test

- Suppose we observe iid random variables $W_1, W_2, \ldots, W_n$ taking values in sample space $\mathcal{X} = \{1, 2, \ldots, k\}$, which we think of as *labels* or *categories*

- We want to test whether the $W_i$'s are distributed according to some hypothesized probability measure $\mathbb{P}_0$ on $\mathcal{X}$

- Let $X_j = \sum_{i=1}^{n} \mathbb{1}_{W_i = j}$ and let $p_j = \mathbb{P}_0(\{j\})$ so that under $H_0$,

$$(X_1, X_2, \ldots, X_k) \sim \text{Multinomial}(n; p_1, \ldots, p_k)$$

$X_j \overset{d}{=} \sum_{i=1}^{n} Y_i \sim \text{Bin}(n, p_j)$

where $Y_1, \ldots, Y_n \sim \text{Bernoulli}(p_j)$

- Now define

$$R_j = \frac{X_j - \mathbb{E}[X_j]}{\sqrt{\text{Var}(X_j)}} \overset{H_0}{=} \frac{X_j - np_j}{\sqrt{np_j(1 - p_j)}}$$

by the central limit theorem

- Since $R_j \xrightarrow{d} \mathcal{N}(0, 1)$ under $H_0$, it's tempting to think $\sum_{j=1}^{k} R_j^2 \xrightarrow{d} \chi^2_{(k)}$, but that's not true because the $X_j$'s (and thus the $R_j$'s) aren't independent!

If $\vec{X} \sim \text{Multinomial}(n, p_1, \ldots, p_k)$, then $\sum_{i=1}^{k} X_i = n$.

# Pearson's Chi-Squared Test

- Instead, we have the following result
- Theorem 4.4: If $(X_1, X_2, \dots, X_k) \sim$ Multinomial$(n; p_1, \dots, p_k)$, then

$$\sum_{j=1}^{k} (1 - p_j) R_j^2 \overset{\text{check!}}{=} \sum_{j=1}^{k} \frac{(X_j - np_j)^2}{np_j} \xrightarrow{d} \underbrace{\chi^2_{(k-1)}}_{\text{the "asymptotic distribution" (Module 5 for more)}}.$$

- The statistic $\chi^2(\mathbf{X}) = \sum_{j=1}^{k} \frac{(X_j - np_j)^2}{np_j}$ is called a **chi-square statistic**, and it's almost always written as

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} \longrightarrow \chi^2_{(k-1)}$$

$O_i$ = # of "observed" i's

$E_i$ = # of "expected" i's

- The chi-squared test is an *approximate test*, because the test statistic only has the $\chi^2_{(k-1)}$ distribution in the limit (more on this in Module 5)

# A Famous Example: Fisher and Mendel's Pea Data

- Mendelian laws of inheritance establish relative frequencies of dominant and recessive phenotypes across new generations

- Gregor Mendel was known for his pioneering experiments with pea plants in the mid-1800s

- If you cross smooth, yellow male peas with wrinkled, green female peas, Mendelian inheritance predicts these relative frequencies of traits in the progeny:

| | Yellow | Green |
|---|---|---|
| Smooth | $\frac{9}{16}$ | $\frac{3}{16}$ |
| Wrinkled | $\frac{3}{16}$ | $\frac{1}{16}$ |

Relabel:

$1 \Leftrightarrow$ Yellow + Smooth

$2 \Leftrightarrow$ Yellow + Wrinkled

$3 \Leftrightarrow$ Green + Smooth

$4 \Leftrightarrow$ Green + Wrinkled

"$\mathbb{P}_0$": $\mathbb{P}_0(\{1\}) = 9/16$ $\quad \mathbb{P}_0(\{3\}) = 3/16$

$\mathbb{P}_0(\{2\}) = 3/16$ $\quad \mathbb{P}_0(\{4\}) = 1/16$

# A Famous Example: Fisher and Mendel's Pea Data

- Mendel crossed 556 such pairs of peas together and recorded the following counts:

OBSERVED COUNTS

| | Yellow | Green |
|---|---|---|
| Smooth | 315 | 108 |
| Wrinkled | 102 | 31 |

EXPECTED COUNTS

| | Yellow | Green |
|---|---|---|
| Smooth | 312.75 | 104.25 |
| Wrinkled | 104.25 | 34.75 |

- Example 4.23: Do these results support the predicted frequencies?

$$\chi^2(\vec{x}) = \frac{(315-312.75)^2}{312.75} + \frac{(108-104.25)^2}{104.25} + \frac{(102-104.25)^2}{104.25} + \frac{(31-34.75)^2}{34.75} \approx 0.6043$$

Our p-value is $p(\vec{x}) = \mathbb{P}\left(\chi^2_{(3)} \geq \chi^2(\vec{x})\right)$

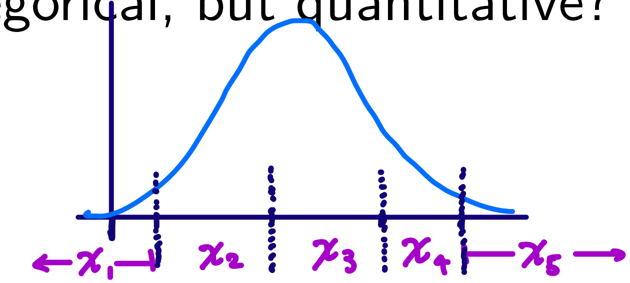
$$= 1 - F_{\chi^2_{(3)}}(0.6043)$$

$\simeq 0.895$. So we (really) fail to reject $H_0$ at the 0.05 significance level.

Check out the "Mendelian paradox"!

# Extending the Chi-Squared Test

- What if our hypothesized distribution is not categorical, but quantitative?




- We can still use a chi-squared test – but how?

- The trick is to partition the sample space $\mathcal{X}$ into $k$ disjoint subsets $\mathcal{X}_1, \ldots, \mathcal{X}_k$, and let $X_j = \sum_{j=1}^{n} \mathbb{1}_{W_j \in \mathcal{X}_j}$ and $p_j = \mathbb{P}_0(\mathcal{X}_j) = \mathbb{P}_0(W_i \in \mathcal{X}_j)$

  Eg: $\mathcal{X} = \mathbb{R}$. Maybe $\mathcal{X}_1 = (-\infty, -3]$, $\mathcal{X}_2 = (-3, 2]$, $\mathcal{X}_3 = (2, 3]$, $\mathcal{X}_4 = (3, \infty)$...

- The finer our partition, the better we can distinguish between distributions

- But of course, we need to increase our sample size accordingly so that each category gets sufficiently "filled" with data

Guideline: each $\mathcal{X}_j$ should contain at least 5 observations before doing this!

If we have 0 observations inside some $\mathcal{X}_j$, then we can't reasonably hypothesize anything except $p_j = 0$

# A Famous Example: Testing for Uniformity

- There are many reasons why we might want to test whether some data $U_1, \ldots, U_n$ arises from a Unif $(0, 1)$ distribution

* Probability plots: we use the probability integral transform to make $F_0(X_1), \ldots, F_0(X_n) \sim \text{Unif}(0,1)$ under $H_0$: "$F_0$ generated the $X_i$'s." The chi-squared test is essentially a quantitative version of the probability plots from before.

* Random number generation: when simulating data from some distribution $F_0$, we typically need to start with $U_1, \ldots, U_n \overset{iid}{\sim} \text{Unif}(0,1)$ random variables, and then transform them (e.g. $F_0^{-1}(U_i) \sim F_0$ check!). We can't generate truly random numbers,* but we can construct a deterministic sequence $u_1, u_2, u_3, \ldots$ that "looks" random enough.

- We can use a chi-squared test for this by binning $[0, 1]$ into $k$ equal-sized sub-intervals of length $1/k$, and letting $X_i = \sum_{j=1}^{n} \mathbb{1}_{U_j \in (\frac{i-1}{k}, \frac{i}{k}]}$ and $p_i = 1/k$

* Exception: numbers generated by radioactive decay ("Hotbits")

# A Famous Example: Testing for Uniformity

- Example 4.24: How can we test whether an iid sequence $U_1, \dots, U_n$ arises from a Unif $(0, 1)$ distribution using $10$ categories?

Partition $(0,1]$ into $(0, \frac{1}{10}], (\frac{1}{10}, \frac{2}{10}], \dots, (\frac{9}{10}, 1]$

and let $X_j = \sum_{i=1}^{n} \mathbb{1}_{U_i \in (\frac{j-1}{10}, \frac{j}{10}]}$ , $j = 1, \dots, 10$.

OR: let $V_i = \lceil 10 \cdot U_i \rceil$ so that

$V_1, \dots, V_n \overset{iid}{\sim} \text{Unif}(\{1, .., 10\})$ under $H_0$

and let $X_j = \sum_{i=1}^{n} \mathbb{1}_{U_i = j}$

Then carry out a chi-squared goodness of fit test by calculating $\chi^2(\vec{x}) := \sum_{i=1}^{10} \frac{(x_i - n/10)^2}{n/10}$,

and compare that to a $\chi^2_{(9)}$ distribution: $p(\vec{x}) = \mathbb{P}(\chi^2_{(9)} > \chi^2(\vec{x}))$

$$= 1 - F_{\chi^2_{(9)}}(\chi^2(\vec{x})).$$

FYI: this is actually a very underpowered test. :(

There are much better randomness tests out there. The "Diehard tests" are standard these days.

# Other Goodness of Fit Tests

- Pearson's chi-squared isn't the only goodness of fit test out there; there are countless others

- Many apply to one particular parametric family specifically

Eg: for testing normality, there are the "Shapiro-Wilk test", the "Anderson-Darling test", the "Jarque-Bera test"...

- Others are completely generic and test for equality between *any* two distributions

The "Kolmogorov-Smirnov test" and the "Cramer-von Mises test" are the most popular

- These latter tests allow us to compare an ecdf $\hat{F}_n$ to a hypothesized cdf $F_\theta$

They're very helpful! They're like quantitative versions of the $\hat{F}_n$-vs-$F_\theta$ visual diagnostic

# Other Goodness of Fit Tests

- In most cases, the distribution of the test statistic under $H_0$ is only known in the limit as $n \to \infty$
- Even then, cutoffs often can't be calculated exactly and require simulations to approximate
- When there's more than one test out there for the same thing, it's always a good idea to read up on the benefits/drawbacks of each one before deciding which to use
- One might have a lower probability of type I error, another might higher power for lower sample sizes, another might be more robust to outliers, and so on

Fairly active area of research!