ETC5242 Week 7

Probability & MLE

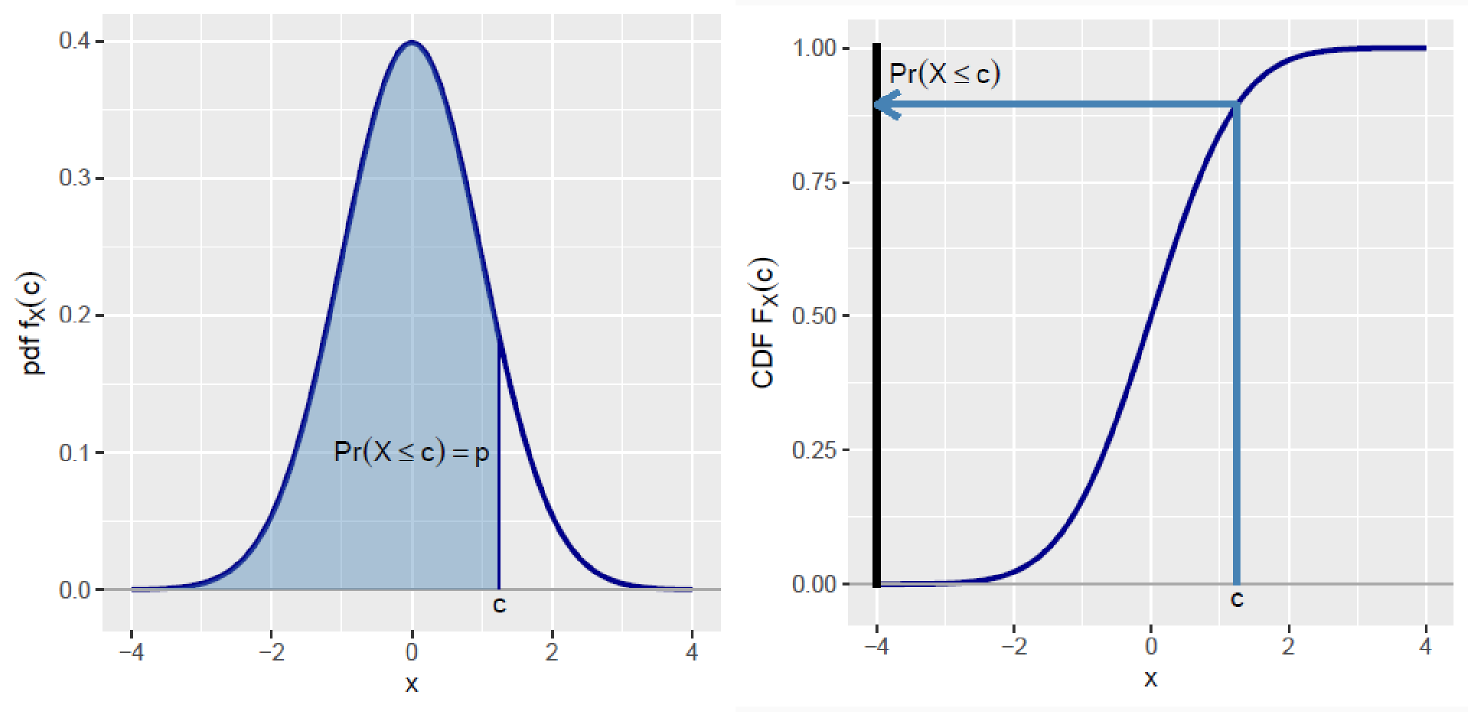
授课老师：Joe

ETC5242平时班 – Week 7

- **Week 7**
  - **Assessing model fitness**
  - **Bayes theorem**

ETC5242平时班 – Week 7

- The cdf of X, denoted $F_X(c)$, returns a value $p \in [0,1]$
- This is equal to the area under the pdf of X, denoted $f_X(c)$, between $(-\infty, c]$



ETC5242平时班 – Week 7

We have the following p.d.f.

F(x)	=	1/15	for -3<=x <3
	=	3/40x	for 3<=x <5
	=	0	otherwise

Find the c.d.f. of the above function

ETC5242平时班 – Week 7

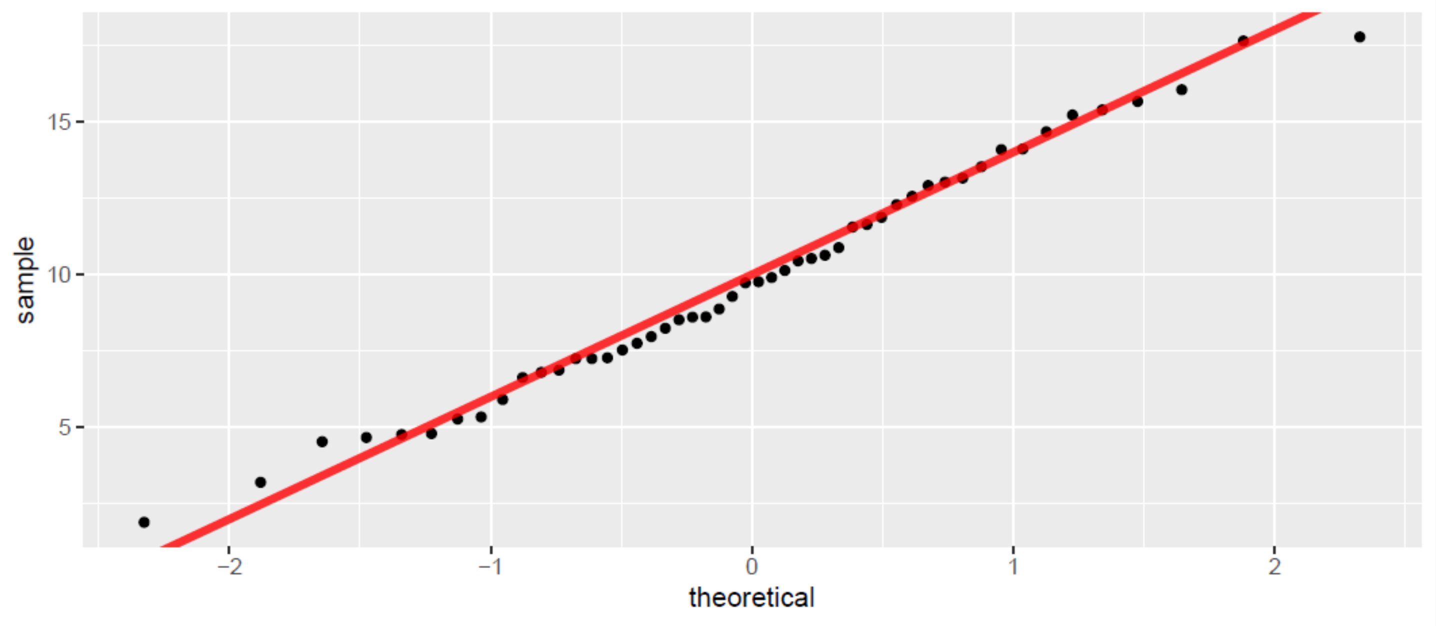
QQ plot

- A graphical tool (subjective visual check) to help assess if plausible that data came from specified distribution
  - e.g. a distribution from MLE fit
- Create scatterplot
  - ordered data (y-axis) against theoretical quantiles (x-axis), or
  - ordered sample data against ordered simulated data
- If both sets of quantiles from same distribution $\Rightarrow$ points should lie on a straight line
  - if not straight, may get an idea of where data doesn't fit
- Often useful to add a line to QQplot
  - 45° line (perfect alignment)
  - line connecting specified quantiles (e.g. 25th- and 75th-%iles)

ETC5242平时班 – Week 7

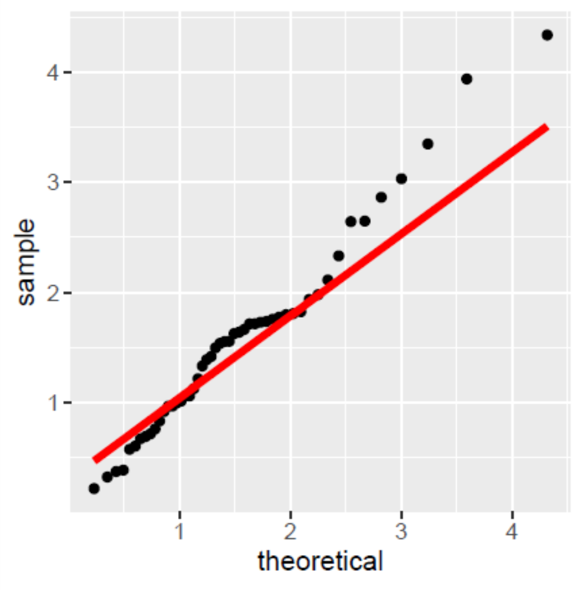
Example 1: N(μ, σ^2) against N(0,1) quantiles

```
n <- 50
df <-tibble(x = rnorm(n, 10, 4))
p <- df %>% ggplot() + geom_qq(aes(sample=x))
p <- p + geom_abline(intercept = 10, slope = 4, color = "red",
              size = 1.5, alpha = 0.8)
p
```



Example 2: stat_qq() for different distributions

```
df <- tibble(mydata=rgamma(n=50, shape=3, rate=2))
fit <- fitdistr(df$mydata, "gamma")
params <- fit$estimate
ggplot(df, aes(sample = mydata)) +
  stat_qq(distribution = qgamma, dparams = params) +
  stat_qq_line(distribution = qgamma,
               dparams = params, color = "red", size=1.5) +
  theme(aspect.ratio = 1)
```



ETC5242平时班 – Week 7

- Can we test?
- H_0: data comes from the specified model vs. H_1 data does not come from the specified model
- In most cases, fit will not be perfect

Various approaches available for informal test:

- Use a ‘thick-marker’ judgment approach
- Use a bootstrap technique to obtain “confidence set”
- Embed QQplot from among many QQplots from data simulated from the model

ETC5242平时班 – Week 7

Bootstrap MLE QQ plot

```
MLE.x <- fit$estimate # point estimate
boot.seq <- seq(1,n,1)/n-1/(2*n)
B <- 500
MLE.x_boot <- matrix(rep(NA,2*B), nrow=B, ncol=2)
for(i in 1:B){
  temp <- sample(df$mydata, size=n, replace=TRUE)
  df <- df %>% mutate(temp=temp)
  MLE.x_boot[i,] <- fitdistr(temp, "normal")$estimate
  params_boot <- MLE.x_boot[i,]
  p <- p + stat_qq(aes(sample=temp), distribution = qnorm,
                   dparams = params_boot, colour="grey",
                   alpha=0.2)
}
p <- p + stat_qq(aes(sample=mydata), distribution = qnorm,
                 dparams = params) +
  ggtitle("QQ plot with B=500 Bootstrap replicates")
p
```

ETC5242平时班 – Week 7

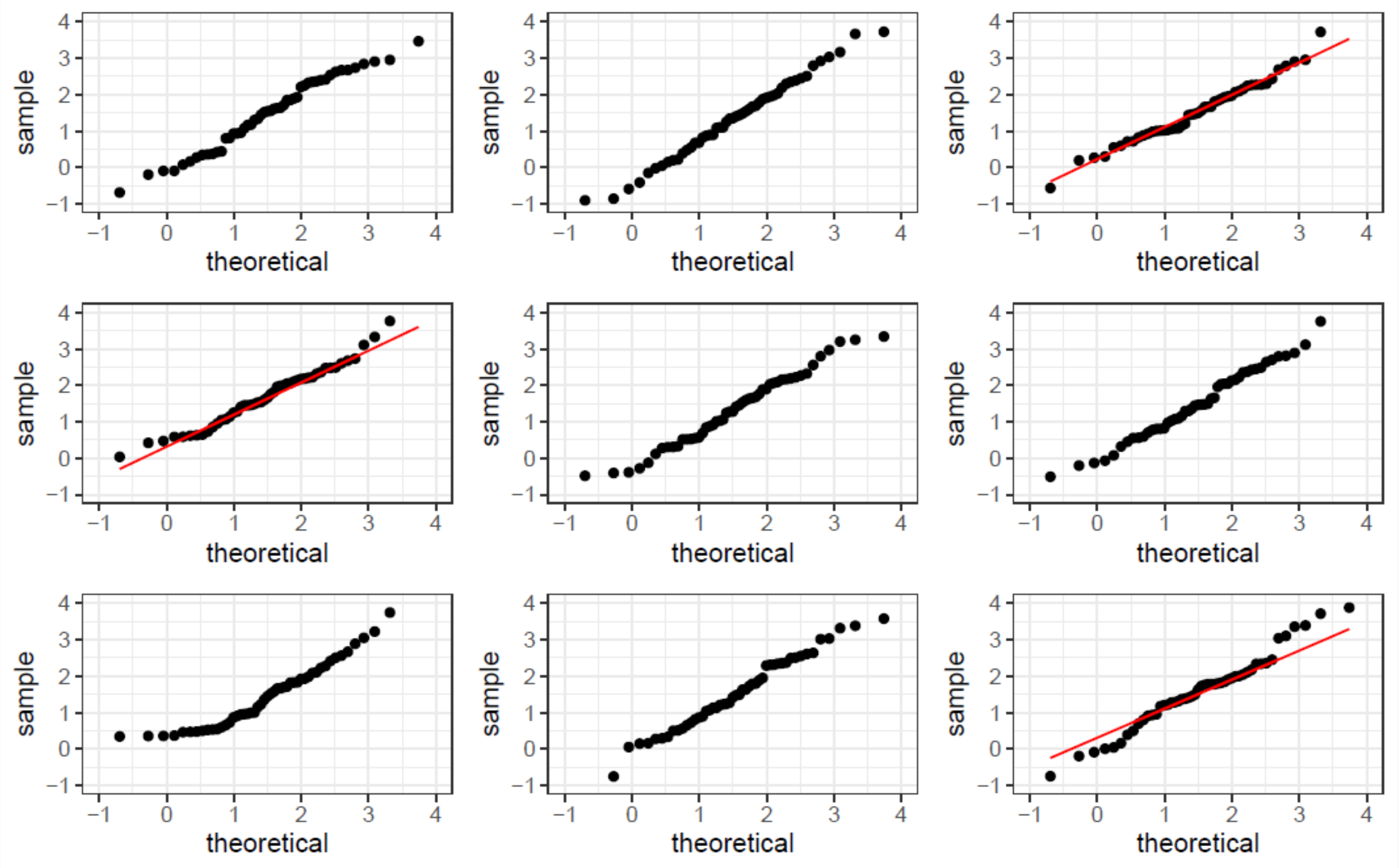
Visual test

- Simulate $K-1$ samples of the same size from the fitted distribution
- Make additional QQ-plots for these simulated samples
- Randomly place QQ-plot of actual data among the $K-1$ comparator QQ-plots
- Try to spot the "odd-one-out" where data least compatible with the 45° line

Null and alternative hypotheses for visual test

- H_0: actual data is a random sample from the fitted distribution, vs.
- H_1: actual data is not a from the fitted distribution
- $\Rightarrow$ Reject H_0 if you can detect the QQplot constructed from the actual data
- Under H_0, the chance of incorrectly rejecting H_0 is $\alpha = 1/(K)$

ETC5242平时班 – Week 7



ETC5242平时班 – Week 7

Bayesian inference

- Probability describes **degree of belief**, and are inherently subjective. Prior belief can be updated, using data and a model for its behaviour.
- Probability statements can be made about parameters, even if parameters are conceived as being fixed, because our knowledge about them need not be fixed.
- We make inferences about a parameter, θ, by producing a probability distribution for θ. Inferences, such as point estimates and interval estimates, may then be extracted from this distribution.

- What make an approach “Bayesian”?
  - Using probabilities to characterise “belief”
  - Treat unknown parameters as “random”, rather than being “fixed”
  - Condition on observed data

ETC5242平时班 – Week 7

You are the manager of a retail clothing store

- A customer returns a **shirt** purchased from the store that **is faulty**
- There are **only 3 manufacturers** who supply this particular shirt

Suppose **it is known** that

- **10%** of the clothing from M_1 (manufacturer 1) faulty
- **5%** from M_2 faulty
- **15%** from M_3 faulty

Which **manufacturer** produced the faulty shirt?

- Can statistics tell us anything about this?
- Note have only a **single data point**: $X = 1$

ETC5242平时班 – Week 7

Suppose we had some **additional ("prior") information**:

- 60% of the stock comes from M_1
- 30% from M_2
- 10% from M_3

Would knowing this prior information change your guess?

- After all, there are relatively few shirts from M_3

- **Bayesian statistics** helps us to answer questions like these
  - *And more...*

- For this we need to use **Bayes' theorem**:

$$\Pr(p_i \mid X = 1) = \frac{\Pr(X = 1 \mid p_i) \Pr(p_i)}{\sum_{j=1}^{3} \Pr(X = 1 \mid p_j) \Pr(p_j)}, \text{ for } i = 1, 2, 3$$

- Notice the general form of **Bayes' theorem**:

$$\text{Posterior} \propto \textbf{Likelihood} \times \textbf{Prior}$$

ETC5242平时班 – Week 7

- Here we have **prior probabilities** for each $p_i, i = 1, 2, 3$
- Bayes theorem calculation:

M_i	Prior $\Pr(M_i)$	Likelihood $\Pr(X = 1 \mid M_i) = p_i$	Prior $\times$ Likelihood $\Pr(M_i)\Pr(X = 1 \mid M_i)$	Posterior $\Pr(M_i \mid X = 1)$
1	0.60	0.10	0.060	0.67
2	0.30	0.05	0.015	0.17
3	0.10	0.15	0.015	0.17
Column Total	1.0	–	0.09 (denominator for Bayes' theorem)	1.0

- Now $\Rightarrow$ M_1 **appears to be MOST PROBABLE**, with
- $\Pr(M_1 \mid X = 1) = 67\%$
- $\Pr(M_2 \mid X = 1) = 17\%$
- $\Pr(M_3 \mid X = 1) = 17\%$

ETC5242平时班 – Week 7

You work at the treasury and know for a fact that only 1 coin in 1 million is biased. If it is biased, it is biased so that Heads comes up $2/3$ instead of $1/2$ the time. You pick a new coin at random from a bin, and toss it 10 times in a row. It comes up heads 10 times in a row. Which is more probable, and why?