MONASH University

MONASH BUSINESS SCHOOL

Statistical Thinking (ETC2420/ETC5242)

Associate Professor Catherine Forbes

Week 7: Updating discrete probabilities

Learning Goals for Week 7

- Discuss model assessment tools for distributions fitted using MLE
- Transition to Bayesian Statistical Thinking
- Apply Bayes theorem in discrete cases

Assigned reading for Week 7:

- Chapter 2 in *Doing Bayesian Data Analysis*, by J. K. Kruschke

Assessing model fit

Both CLT-based confidence intervals **and** Bootstrap-based confidence intervals

- Constructed from the output of an ML procedure
- Implicitly assume the selected "model" for ML is "correct" for the data

If the model doesn't match the data well

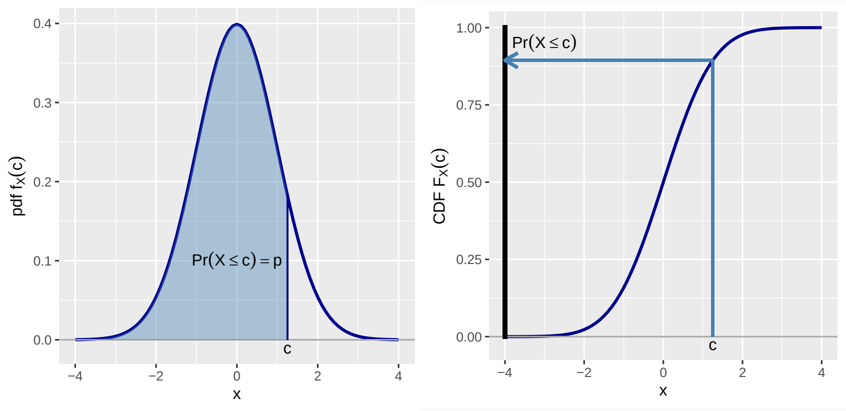
- $\Rightarrow$ parameter estimate and confidence interval(s) will not be very useful!

We need a way to assess the MODEL itself

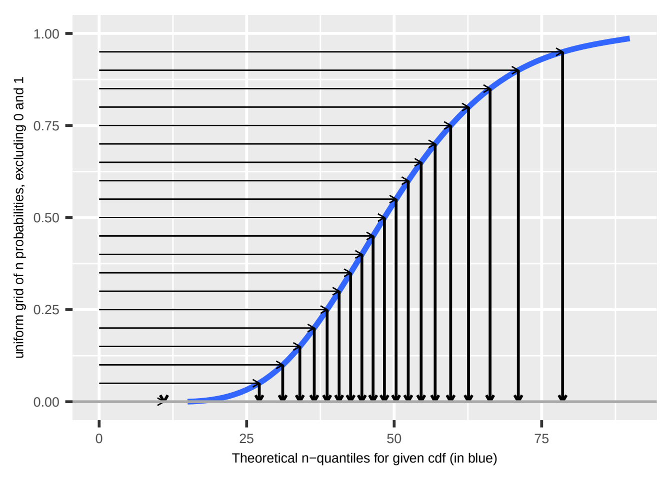
- Is the fitted model suitable for the data?
- Use QQplots, which are based on pairs that match:
  - **theoretical *n*-quantiles** (obtained by inverting the model's cdf) with
  - **empirical *n*-quantiles** (i.e. the sorted sample data values)
- If these pairs "match" then the model is a good fit to the data!

Relationship between quantiles (percentiles), the pdf and the cdf

- 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 $(-\infty, c]$



Inversion of a cdf



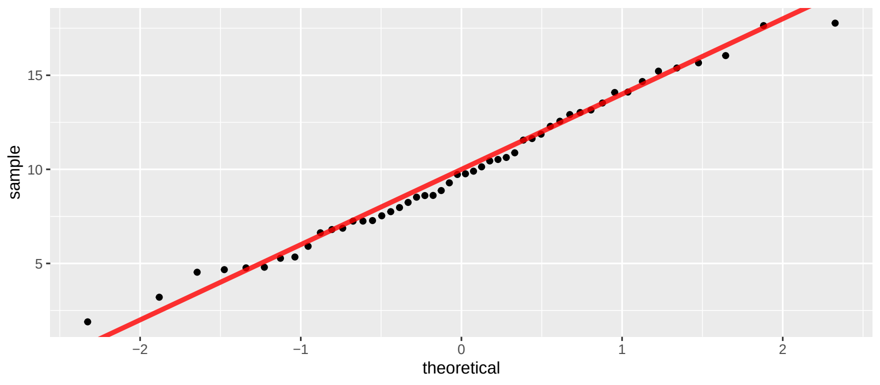
- Avoid potential inversion of cdf at 0 or 1 if range of distribution reaches $-\infty$ or ∞
  - e.g. set (n+1)-quantiles for $p_i = \frac{i}{n} - \frac{1}{2n},\ i = 1, 2, \ldots, n$

Quantile-Quantile Plot (QQplot)

- 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)

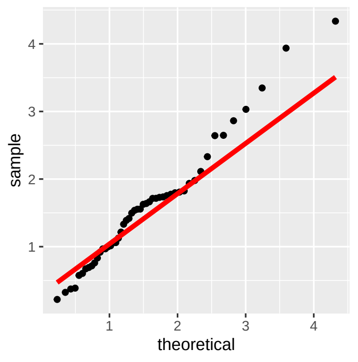
Example 1: N(μ, σ^2) against N(0,1) quantiles

```
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
```



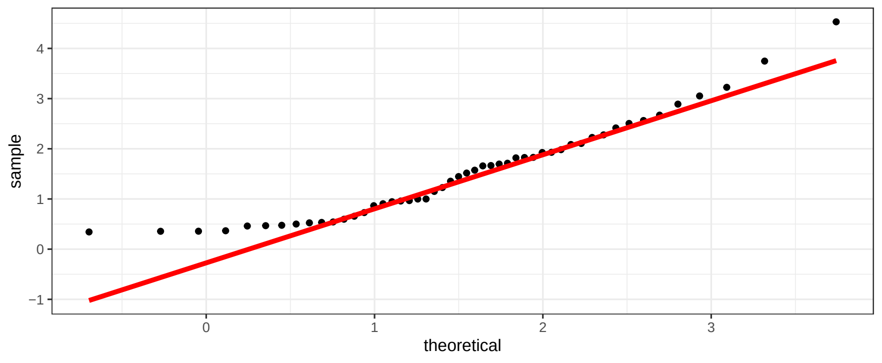
Example 2: stat_qq() for different distributions

```
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)
```



Example 3

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



About QQplots

- 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

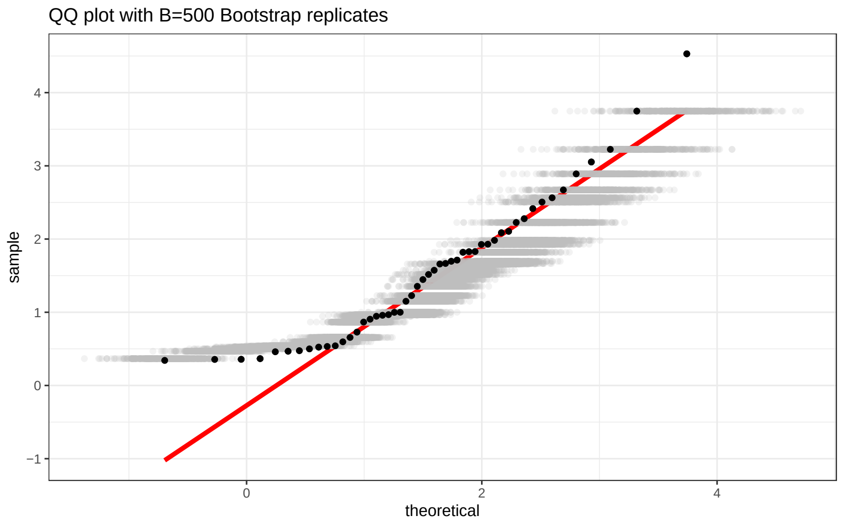
Various approaches available for informal test:

- 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

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
```

Bootstrap MLE QQ plot



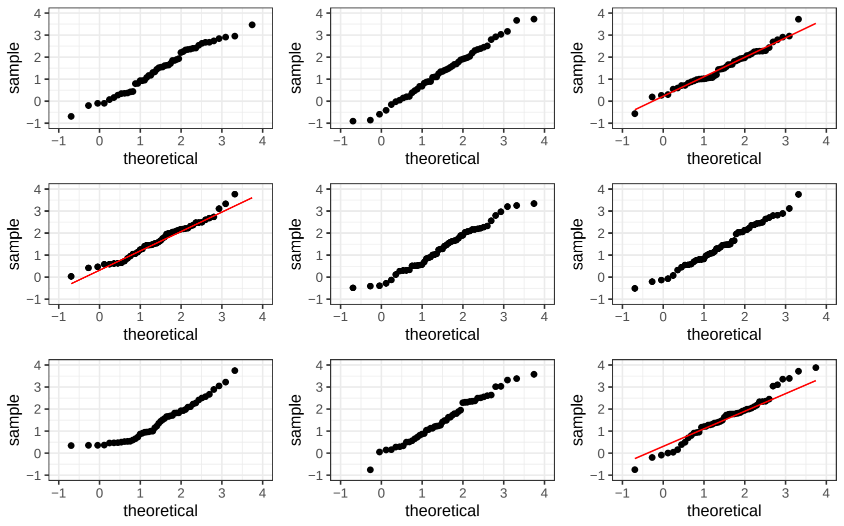
Visual test

- 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 $\alpha = 1/(K)$

Visual test example



What do we do with the MLE?

- What do we do with a good model estimated using MLE?
- Use it to characterise features of the population, e.g. mean, median, IQR, event probabilities:

$$\hat{E}[X \mid \theta] = E[X \mid \hat{\theta}_{MLE}]$$

$$\hat{Median}\{F_X(x \mid \theta)\} = Median\{F_X(x \mid \hat{\theta}_{MLE})\}$$

- Use it to predict future outcomes or construct prediction intervals (assuming i.i.d)

$$\hat{E}[X_{n+1} \mid \theta] = E[X_{n+1} \mid \hat{\theta}_{MLE}]$$

$$\hat{\Pr}(q_{0.025} \leq X_{n+1} \leq q_{0.975} \mid \theta) = \Pr(q_{0.025} \leq X_{n+1} \leq q_{0.975} \mid \hat{\theta}_{MLE})$$

The invariance property of the MLE

- If $\hat{\theta}$ is the MLE of θ, then the MLE of a function $\tau(\theta)$ is $\hat{\tau}(\theta) = \tau(\hat{\theta})$
  - Bootstrap-based confidence intervals can be constructed
  - If $\tau(\theta)$ is a smooth function, then CLT-based confidence intervals can be obtained

Transition to Bayesian Thinking (Wasserman, 2004)

Frequentist inference (everything we have discussed so far. . .)

- Probability refers to **limiting relative frequencies**. Probabilities are objective properties of the real world.
- Parameters are fixed, unknown constants. Because they are not fluctuating, no useful probability statements can be made about parameters.
- Statistical procedures should be designed to have well-defined long run frequency properties. For example, a 95% confidence interval should trap the true value of the parameter with limiting frequency at least 95%.

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.

Introduction to Bayesian statistics

- You all know about Bayes' theorem for calculating conditional probabilities
- Bayesian statistics applied Bayes' rule to modelling data
- e.g.
  - Fit models to data
  - Estimate unknown parameters
  - Characterising uncertainty in parameter estimates
  - Choosing between competing models
  - Predicting future events
  - Ensemble models
  - . . . (and more)
- What make an approach "Bayesian"?
  - Using probabilities to characterise "belief"
  - Treat unknown parameters as "random", rather than being "fixed"
  - Condition on observed data

An example

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**: $X = 1$

A model for the faulty shirt

Let p_i denote the probability that a shirt from M_i is faulty, for $i = 1, 2, 3$

Consider a **randomly selected shirt** could be either faulty ($X = 1$) or not ($X = 0$)

- For $M_i \Rightarrow X \sim Bernoulli(p_i)\}$ random variable:
- The **"success" probability** (when $X = 1$) **depends on the manufacturer**

We **have**

- $\Pr(X = 1 \mid M_i) = p_i$ for $i = 1, 2, 3$

We **want**

- $\Pr(M_i \mid X = 1)$ for $i = 1, 2, 3$
- $\Rightarrow X \mid M_i \sim Bernoulli(p_i)$, where $p_1 = 0.10$, $p_2 = 0.05$ and $p_3 = 0.15$

Frequentist approach: Maximum likelihood estimation

We have a **model** for this one observation $\Rightarrow$ a **likelihood function**:

$$\mathcal{L}(p) = p^X(1-p)^{1-X}, \quad \text{for } p \in \{p_1, p_2, p_3\}$$

This function $\mathcal{L}(p)$ can be maximised!

- view it as a function of p
- with X **{**fixed} at the observed $X = 1$

Manufacturer M_i	Value of p p_i	Likelihood $p_i = p_i^1(1-p_i)^0$
M_1	0.10	0.09
M_2	0.05	0.0475
M_3	0.15	0.1275

$\Rightarrow$ M_3 **appears to be MOST LIKELY** (not surprising!)

- Note we cannot assess uncertainty around this guess

Prior information

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**:

  Posterior $\propto$ **Likelihood** $\times$ **Prior**

Bayes' theorem: inverting probabilities

Review Probability from Week 6

$$\Pr(A \mid B\ ) = \frac{\Pr(B \cap A)}{\Pr(B)} = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B \mid A)\Pr(A) + \Pr(B \mid A^c)\Pr(A^c)}$$

- $\Pr(A)$ and $\Pr(A^c)$ are **marginal probabilities** (**"prior"**)
- $Pr(A \mid B)$ and $Pr(A^c \mid B)$ are **conditional probabilities** (**"posterior"**, after update)

More possibilities:

$$\Pr(A_1 \mid B\ ) = \frac{\Pr(B \cap A_1)}{\Pr(B)} = \frac{\Pr(B \mid A_1)\Pr(A_1)}{\Pr(B \mid A_1)\Pr(A_1) + \cdots + \Pr(B \mid A_k)\Pr(A_k)}$$

- $\Pr(A_k)$'s are **marginal probabilities** (**"prior"**)
- e.g. $Pr(A_1 \mid B)$'s is a **conditional probability** (**"posterior"**, after observing B)

Note this is for **discrete set** of possibilities $A_1, A_2, \ldots, A_k$

Bayesian Solution to the retail problem

- 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
- $\Pr(M_1 \mid X = 1) = 67\%$
- $\Pr(M_2 \mid X = 1) = 17\%$
- $\Pr(M_3 \mid X = 1) = 17\%$

What if you didn’t believe a specific coin was fair?

Bayesians could put prior probabilities over a collection $p \in \{p_1, p_2, \ldots, p_k\}$

Or could assume belief for p over continuum $p \in (0, 1)$ (e.g. $p \sim Uniform(0, 1)$)

In either case

- $\Rightarrow$ Can work out updated probabilites after viewing tosses of **specific** coin (data) using **Bayes’ theorem**
- Then we update **subjective prior belief**, given data

Let data $X =$ number of heads (“successes”) in n coin tosses

- This is a **model** for the random variable X, given parameter p
- $\Rightarrow X \sim Binomial(n, p)$

$$P(X = x \mid n, p) = \binom{n}{x} p^x (1 - p)^{n-x} \quad \text{for } x \in \{0, 1, 2, \ldots, n\}$$

- When viewed as a function of p, with $X = x$ fixed $\Rightarrow$ **likelihood function** $\mathcal{L}(p)$

Bayes theorem for Binomial observation, with a discrete prior

- After observing data $X = x$, we update our beliefs and calculate the posterior distribution
- Assign **prior probabilities** $\{\pi_1, \pi_2, \ldots, \pi_K\}$ over a discrete set of points $\{p_1, p_2, \ldots, p_K\}$,
  - i.e. $\Pr(p = p_i) = \pi_i$, for $i = 1, 2, \ldots, K$
  - p_i is like A_i, and X is like B in Bayes' theorem
- We use Bayes theorem similar to in the "shirt problem"
  - $\Pr(p_i \mid X = x) \propto \textit{Prior} \times \textit{Likelihood}$, subject to $\sum_{i=1}^{K} \Pr(p_i \mid X = x) = 1$

p_i	Prior $\Pr(p = p_i)$	Likelihood $\Pr(X = x \mid p_i)$	Prior $\times$ Likelihood $\Pr(p = p_i)\Pr(X = x \mid p_i)$	Posterior $\Pr(p_i \mid X = x)$
p_1	π_1	$p_1^x\,(1 - p_1)^{n-x}$	$\pi_1\, p_1^x (1 - p_1)^{n-x}$	$\pi_1\, p_1^x (1 - p_1)^{n-x} / m(x)$
p_2	π_2	$p_2^x\,(1 - p_2)^{n-x}$	$\pi_2\, p_2^x (1 - p_2)^{n-x}$	$\pi_2\, p_2^x (1 - p_2)^{n-x} / m(x)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_K	π_K	$p_K^x (1 - p_K)^{n-x}$	$\pi_K\, p_K^x\,(1 - p_K)^{n-x}$	$\pi_K\, p_K^x (1 - p_K)^{n-x} / m(x)$
Column Total	1.0	–	$m(x) = \sum_{k=1}^{K} \pi_k\, p_k^x\,(1 - p_k)^{n-x}$ **(denominator for Bayes' theorem)**	1.0